A Deep Learning Based Prediction of Arabic Manuscripts Handwriting Style

Manal Khayyat[1] and Lamiaa Elrefaei[2]
[1]Computer Science Department, King Abdulaziz University, Saudi Arabia
[2]Electrical Engineering Department, Benha University, Egypt

Abstract: *With the increasing amounts of existing unorganized images on the internet today and the necessity to use them efficiently in various types of applications. There is a critical need to discover rigid models that can classify and predict images successfully and instantaneously. Therefore, this study aims to collect Arabic manuscripts images in a dataset and predict their handwriting styles using the most powerful and trending technologies. There are many types of Arabic handwriting styles, including Al-Reqaa, Al-Nask, Al-Thulth, Al-Kufi, Al-Hur, Al-Diwani, Al-Farsi, Al-Ejaza, Al-Maghrabi, Al-Taqraa, etc. However, the study classified the collected dataset images according to the handwriting styles and focused on only six types of handwriting styles that existed in the collected Arabic manuscripts. To reach our goal, we applied the MobileNet pre-trained deep learning model on our classified dataset images to automatically capture and extract the features from them. Afterward, we evaluated the performance of the developed model by computing its recorded evaluation metrics. We reached that MobileNet convolutional neural network is a promising technology since it reached 0.9583 as the highest recorded accuracy and 0.9633 as the average F-score*.

Keywords: *Deep Learning Model, Convolutional Neural Network, Handwriting Style Prediction, Arabic Manuscript Images.*

Received October 6, 2019; accepted April 6, 2020
https://doi.org/10.34028/iajit/17/5/3

1. Introduction

Images classification and prediction is a technique of Computer vision to classify queried images based on some specifications. We can utilize deep learning technology for training the dataset to predict specific traits of our dataset images successfully. Deep learning is a subfield of machine learning, which is itself a subfield of artificial intelligence [12]. The concept of deep learning technology based on simulating real human perceptions in visualizing images. Hence, it aims to extract higher-level features such as activities or objects presented within images automatically. Al-Ayyoub *et al*. [4] admit that deep learning mimics humans' brains through leveraging multiple complex algorithms to discover the right model for extracting the distinguishing features. Moreover, Tyagi [20] claims that deep neural networks proved their efficiency in resembling humans' brains by developing many non-linear transformations that create hierarchal abstract layers that can intelligently learn complicated features. Thereby, they are able to retrieve accurate results.

According to Zhou and Jia [24], utilizing a deep neural network as a learning method accomplished substantial success compared with other methods that depend on classical computations. Because the traditional approaches require domain experts, time consuming, error-prone, and scalable to new problems. On the other hand, the deep learning approach is a computing model that learns from data, easy to extend, and able to speed up using GPUs [3]. Thus, the deep learning technique proved its rigidness in many domains and generated high evaluation parameters.

There are five main tasks of deep learning as following: detection, classification, segmentation, prediction, and recommendation. In addition, there are many different types of deep neural networks; the most common types are Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Liang *et al*. [14] believe that CNN is a multi-layer network, whereas each layer has its neuron nodes. The nodes utilize weights and biases to be able to learn the relationships between input variables and output variables [17]. According to Das [9], there are many different architectural models for CNN, such as LeNet, AlexNet, MobileNet, ZFNet, GoogLeNet, VGGNet, ResNet, …etc., However, all of them based on a similar architecture consisting of one input layer, one output layer, and many hidden layers in between them responsible for performing the computations and representations.

Regarding the used dataset in the study, which is Arabic manuscripts; they are historical hand-written papers constituting books. They considered historical because they were written either before the existence of the "Hijra" time period, which is also known by the "Hijri" date or they were written after the "Hijri" date but, before the year of 1305. Therefore, there was no

standard format used by the authors of the historical Arabic manuscripts, which made them different and usually hard to read. In addition, the automatic manipulation and understanding of the Arabic language are challenging due to its distinct characteristics [4]. Al-Jawfi [5] claims that the handwritten Arabic language has characteristics that make recognizing them harder than other languages. For instance, the fact that the Arabic words are written from right to left and their dots might appear above or below the letters, also the dots are ranging from 0-3 makes the recognition of the Arabic language more challenging. Moreover, even though the Arabic alphabets are consisting of only 28 letters, however, most of them come in four different forms depending on their position within the word, which increases the alphabet patterns from 28 to around 60.

Despite the fact that the Arabic characters are difficult to process, the ancient Arabic manuscripts that are handwritten, very old, and having low-quality resolution are much more difficult to manipulate. These characteristics made them complex to be visualized and being able to predict their handwriting style correctly. Yahia [22] believes that the historical Arabic manuscripts, in particular, pose additional challenges due to their degraded quality of blotched papers and faded inks.

Throughout the generations, many Arabic handwriting styles used. Discussing the various types of handwriting styles in Arabic manuscripts, the oldest handwriting style and which used in writing the holy Quran during the first five Hijri centuries, is Al-Kufi handwriting style. Another Arabic handwriting style used in witting the holy Quran is Al-Nask. It is easier in writing, and it is considered a branch from Al-Thulth handwriting style. There is also another branch found from Al-Thulth handwriting style, which is named Al-Taqraa.

One of the easiest and most used Arabic handwriting styles is Al-Reqaa. Moreover, the Iranian calligraphers used Al-Farsi handwriting style in writing their poems and manuscripts. In Andalusia and Morocco, Al-Maghrabi handwriting style used. While in Al-Iraq country, Al-Ejaza handwriting style used, which is also known as Al-Tawqe and Al-Rayhani. This handwriting style was developed further by the Othmanian Empire. Another common Arabic handwriting style that was found by the Othmanian Empire was Al-Diwani. It became an official handwriting style used by the Othmanian government in the 857 years of Hijra [18].

Contributions of this study present in; "to the best of our knowledge" being the first research study experimenting the use of deep learning technology to predict the handwriting style of Arabic manuscripts. In addition, we collected large amount of historical Arabic images in a dataset and classify them to be able to predict their handwriting style successfully.

Arabic handwriting styles are significant social and political appliances in the Islamic world. The unique handwriting style used in writing the manuscript transfers its historical background, including the region that the manuscript was written at, the specific genre of the manuscript, and the calligraphers' style. It is also possible to identify the period of time that the manuscript was written at, through analyzing its handwritten style.

Automating the process of predicting and retrieving the Arabic handwriting style eliminates the need for calligraphy experts to manually process and handle manuscripts [10]. Thus, the proposed solution can be the step-stone for calligraphy analysis, simulation, or generation applications.

Rest of the paper organized as follows: section 2 discusses related work and current existing solutions. Section 3 explains the development of the deep learning model. Section 4 highlights the conducted experiments with the analysis of their test results. Finally, section 5 concludes the paper.

2. Related Work

The trend in most Computer vision research papers is toward deep learning technology because it has proven success in many different domains; one of them is image classifications. The classification task includes categorizing inputs into specific groups. For instance, we might need to classify research papers into a particular genre based on used terminologies, such as the study done in [6]. In addition, Chan *et al*. [8] propose a deep learning image classification model based on texture features. The architecture utilizes cascaded Principal Component Analysis (PCA) to classify input images like a face, handwritten text, or digits. The authors concluded that leveraging the PCANet-2 deep neural network recorded between 83.74% and 86.66% accuracy of classifying texture images.

Mohamed *et al*. [16] believe that selecting the perfect functioning techniques for feature extraction and similarity measurement; will improve the performance of image classification. Hence, they propose combining the CNN for automatic feature extraction with a Support Vector Machine (SVM) classifier. The selected dataset for their study is the Caltech256 database, which is part of the publicly available "ImagNet" dataset. The authors initially converted the images into a smaller dimension feature. Afterward, they extracted the features one time from each individual image. Finally, they were able to record 90% of image classification accuracy utilizing 1000 images. Moreover, they found-out that increasing the evaluated images number to 6000 images had also increased the classification accuracy up-to 96%.

Wang *et al*. [21] recommend a fusion CNN and RNN framework to classify multiple-label images

accurately. For this purpose, they utilized many datasets like NUS-WIDE, Microsoft COCO, and PASCALVOC2007. They pre-processed their images to visualize them better. Then, they computed the similarity measurement using the beam search algorithm to be able to predict the nearest labels in images successfully. For the evaluation purpose, the authors used Caffe deep learning framework to experiment their proposed model. Hence, they calculated the mean Average Precision (mAP), and reached 84%.

Liu *et al*. [15] investigate the ability of deep neural networks to classify images accurately. They used the Corel dataset to test their model. The authors suggested using the AlextNet convolutional neural network to extract global semantic features from their dataset images. AlexNet activated through the ReLU activation function, which assists in generalizing the network to classify more new images. Afterward, they used the distance metric function to measure the similarity between the query image and the rest of the Corel dataset images. The authors evaluated their model using the mean average precision. Eventually, they figured-out that AlexNet-Fc8 image features extraction recorded 0.9277 mAP.

Seddati *et al*. [19] explore the rigidness of utilizing convolutional neural networks for extracting visual features from images. They used four well-known images datasets to conduct their study. The datasets were IN-RIA Holidays, University of Kentucky Benchmark, Oxford5k, and Paris6k datasets. The authors used ResNet101 CNN feature extractor. Their proposed approach based on the Multi-Scale Regional Maximum Activation of Convolutions (MS-RMAC) descriptor, which utilizes the resulting fully connected layer to extract images features. The authors measured the similarity using the K-nearest neighbour algorithm to find-out the nearest four images to the original queried image. They evaluated their model by computing the mean average precision, and they accomplished an accuracy of 72.3 using Oxford5k, 87.1 using Paris6k, and 94.0 using the "INRIA"[1] Holidays benchmark dataset.

The studies in [2, 7, 10, 23] addressed the classification and prediction of handwriting styles. For instance, Allaf and Al-Hmouz [2] predicted the Arabic calligraphy types utilizing an offline neural network recognition system. They used two different datasets for the classification and prediction task as following: local dataset that consists of three Arabic handwriting styles written by calligraphers and public dataset that consists of ten Arabic handwriting styles generated by the Computer. To pre-process the images of the dataset, the authors converted all the images into binary version and removed the noise. Afterward, they extracted the visual features from the pre-processed images using the Genetic Algorithm. Finally, the authors reached a recognition error rate that equals 8.02% for the local dataset and 7.55% for the public dataset.

Bataineh *et al*. [7] proposed using a backpropagation neural network to predict the Arabic handwriting styles. Considering that the backpropagation neural network consists of one input layer, one hidden layer, and one output layer to perform the classification task. They started by pre-processing fourteen images, including seven handwriting styles by converting them into binary versions and removing their edges and skews. Afterward, they extracted the features using Edges Direction Matrices (EDMs), which is a statistical algorithm for interpreting the texture features in images. The authors eventually accomplished 43.7% recognition accuracy.

Ezz *et al*. [10] classified and predicted only two Arabic handwriting styles, which are Naskh and Reqaa. The authors employed the static Scale-Invariant Feature Transform (SIFT) and Speeded-up Robust Feature (SURF) algorithms to do the features extraction from two-hundred images. Then, they experimented four different machine learning classifiers as following: gaussian naive bayes, decision tree, random forest, and the K-nearest neighbor. They concluded that the best method for predicting the Arabic handwriting styles is utilizing the SIFT with the gaussian naive bayes classifier since it recorded 92% accuracy.

Yu-Sheng *et al*. [23] classified and predicted five Chinese handwriting styles included within a dataset of two-thousands images. Each image in the dataset includes only one Chinese character. The authors experimented four various machine learning algorithms looking for the algorithm that best classifies their dataset images. The experimented algorithms are Softmax regression, support vector machine, K-nearest neighbors, and random forests. Moreover, they tuned the learning hyper-parameters using the k-fold cross-validation method. The authors reached that using the HOG descriptor with the Softmax regression algorithm outperforms other algorithms since it recorded 95.55% accuracy.

After reading previous literature, we found that many efforts implemented deep learning in various domains. At the same time, there is a lack of implementing it on the Arabic handwriting styles in particular. Thus, we aim to tackle the problem of classifying and predicting the Arabic handwriting styles using the deep learning technique. Because once the model is trained then, it will be able to predict the Arabic handwriting style of any unseen input image to the model since it already learned the way to classify the Arabic handwriting styles.

[1]http://lear.inrialpes.fr/people/jegou/data.php

3. Methodology

The proposed method begins by collecting the historical Arabic manuscripts in a dataset and pre-process its images to prepare them for entering the model. Afterward, we developed a customized deep learning model that takes the dataset images as input and outputs their predicted handwriting styles. The transfer learning technique is employed in developing the model to enable it to learn and classify the Arabic handwriting styles successfully. Finally, we conducted experiments to increase the accuracy and prove the success of the proposed method in predicting the Arabic handwriting styles.

3.1. Dataset Collections

We collected the required dataset by randomly selecting historical Arabic manuscripts from the "wqf"[2] Online website. The dataset has a total of (37) Arabic manuscripts/classes, including (2653) images, as illustrated in Table 1.

There are many different types of handwriting styles, but, in our dataset, we found only six of them. Hence, the considered handwriting styles in this study are Al-Nask, Al-Thulth, Al-Reqaa, Al-Hur, Al-Diwani, and Al-Farsi.

We ensured that each handwriting style includes approximately the same number of images to give credence to the implementation of the deep learning model on them and to confirm its fairness of operation. Hence, we have chosen around (450) images under each handwriting style.

By analyzing the manuscripts' images under each handwriting style, we found that all of them are having RGB color representation. Furthermore, we noticed that most of the images are having (2160) pixels for the width dimension and (1440) for the height dimension, which is like a book dimension. However, there are many other images of different sizes. These images might be indices, appendices, or cover pages since they appear mostly either in the beginning or at the ending of the manuscript. Thereby, we resized all the images into (224x224) pixels to prepare them for the MobileNetV1 deep learning model.
Figure 1 illustrates samples of the handwriting styles presented in our dataset.

3.2. Model Development

To develop our model, we transferred learning utilizing the pre-trained (MobileNet-V1-100-244) deep learning model, which was initially trained on "ILSVRC-2012-CLS" portion under ImageNet dataset to classify images. The model is eligible to extract the features from the input images by reading their pixels and transform them into features.

MobileNetV1 considered a small deep neural network. That is because as it operates, it reduces the spatial dimensions between its convolutional tensors. The advantage of this reduction is the faster training and execution of the model. Hence, comparing MobileNetV1 with other larger types of deep neural networks such as VGG19, InceptionV3, and NasNetLarge, …etc., The MobileNetV1 model is quicker in its operation and computation, even though there might be a slight decrease in its measurement metrics.

MobileNetV1 deep neural network accepts a fixed input size of images, which is (224x224). The accepted number of channels is (3), which refers to the representation of the RGB color of the input images to the model. The model performs different types of filters on its processed images to be able to analyze and comprehend them. MobileNetV1 comprises 4.2 million learned parameters, which points-out the model's ability to learning. Another factor that denotes the model's competency is the Multiply Accumulates Compute (MAC), which determines the required amount of computations, and it is equal to 569 million [11]. MobileNetV1 pre-trained model consists of thirteen main convolutional layers ordered linearly without any residual connections between them to keep them simple. Each main convolutional layer is having five other hidden layers attached to it to perform the zero paddings, depth-wise, batch normalization, ReLU activation function, and finally, the two-dimensional feature maps output. After the convolutional layers, there is the Average Pooling (AP) layer, which also assists in extracting the features while reducing the spatial dimensions departed from the preceding feature maps. Finally, it is the Fully Connected (FC) layer that comes with the original model to calculate the output loss function that solves the classification task.

We reused all the layers of the MobileNetV1 pre-trained model for extracting the needed visual features, as well as for the training purposes, except the last fully connected layer. We removed it, and we added instead of it two different layers. The primary purpose behind this step is to be able to load and utilize all the available pre-trained model's weights and, at the same time being able to modify its final layer to fit with our dataset parameters. Hence, we appended one flatten layer followed by the final output dense layer (Softmax layer), as illustrated in Figure 2.

[2]http://wqf.me/?p=15619

Table 1. List of handwriting styles.

Handwriting Style ID	Arabic Name	English Name	Manuscripts Details		Images Number	Total No. of images
			Numbers	Titles		
1	النسخ	Al-Nask	8	قطعة من شرح معاني الآثار	80	421
				الزواجر في الكبائر	16	
				العقد الفريد لبيان الراجح في جواز التقليد	27	
				خزانة الروايات	49	
				تحفة التحرير	7	
				كنز الدقائق	104	
				الدرة المنيفة على مذهب أبي حنيفة	40	
				الاشاعة لاشراط الساعة	98	
2	الثلث	Al-Thulth	8	الأعمال الموجبة	12	447
				مقدمة عن الصلاة و شروطها	25	
				القول البليغ في حكم التبليغ	9	
				أحكام الناطفي	34	
				الفوائد الزينية في مذهب الحنفية	50	
				تعليق الفواصل على إعراب العوامل	82	
				مقامات الحريري	132	
				شرح الصدور في شرح حال الموتى في القبور	103	
3	الرقعة	Al-Reqaa	7	ذكر أسباب إصلاح البيوت	16	468
				ثبت الأمير	30	
				حاشية على متن السمرقندية	9	
				شرح الرسالة العضدية	6	
				حاشية على شرح الكافي	96	
				تبيين الحقائق شرح كنز الدقائق	141	
				فتح القدير	170	
4	الحر	Al-Hur	6	شرح الجامع الصغير	42	475
				شرح الأربعين، المسمى الفتح المبين	101	
				المربع في حكم العقد على المذاهب الأربع	6	
				عمدة الحكام ومرجع القضاة في الأحكام	65	
				درر الحكام شرح غرر الأحكام	82	
				شرح الاجرومية	179	
5	الديواني	Al-Diwani	4	الأجوبة المكية على الأسئلة الحفظية	10	473
				ملتقى الأبحر	158	
				رسالة في القنوت في النوازل	8	
				الأشباه والنظائر الفقهية	297	
6	الفارسي	Al-Farsi	4	الهداية في علم الرواية	16	369
				إجادة الجدة بمنع القصر في طريق جدة	11	
				شرح التسهيل	71	
				ريحانة الألبا وزهرة الحياة الدنيا	271	
Total			37			2653

a) “Al-Nask”.

b) “Al-Thulth”.

C) “Al-Reqaa”.

d) “Al-Hur”.

e) “Al-Diwani”.

f) “Al-Farsi”.

Figure 1. Manuscript images written using different Arabic handwriting styles.

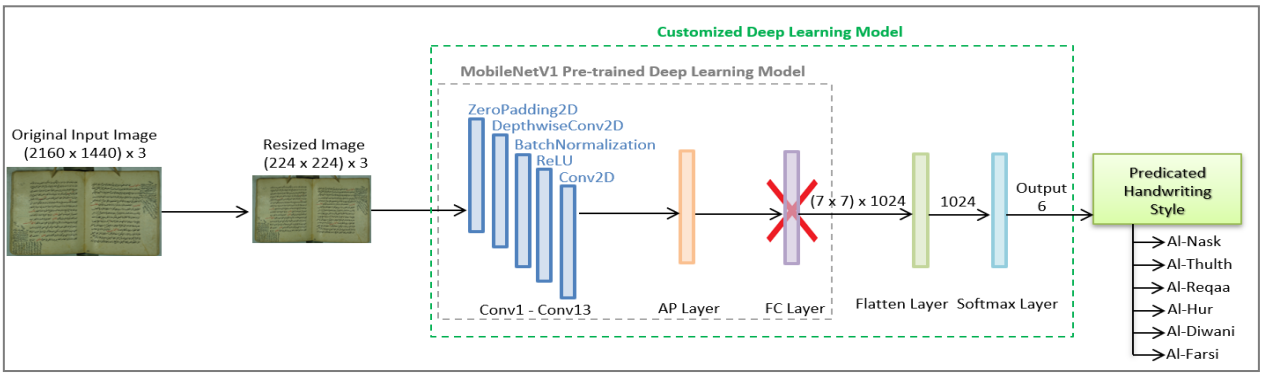


Figure 2. Architecture of the developed model.

The main function of the added flatten layer is to convert the data extracted from the previous convolutional and pooling layers from two-Dimensional matrix (2D) Into one Vector (1V) to prepare them for entering the final dense fully connected layer. The final added dense layer accepts "Softmax" activation function, which is adapted to calculate the six handwriting styles in our dataset instead of the original function with the pre-trained MobileNetV1 model. The Softmax function is presented in Equation (1) [13] :

$$p(s_t|s_{t-k},\dots s_{t+k}) = \frac{e^{ywt}}{\sum_i e^{y_i}} \quad (1)$$

Where p is the Softmax function computed probability, $\{S_1, S_2, S_3 \dots S_T\}$ represents a sequence of training samples, and y_i is the non-normalized log-probability for each output sample i. Equation (2) clarifies the computations of y_i.

$$y_i = b + Uh(s_{t-k},\dots s_{t+k}; S) \quad (2)$$

Where b *and* U represent the parameters of the Softmax function, S represents the matrix that is mapped with the input samples to predict their associated labels, and h is the average of the sample vectors produced from S.

4. Experiments and Tests Results

To conduct our experiments, we implemented the model using the Python programming language version 3.7 and Pycharm API on Ubuntu 16.04 Operating System. Note that "Tensorflow" and "Keras" were two of the main deep learning libraries that we used at the backend.

Concerning the used hardware device, it is the "ABS Battelbox" PC, including Intel Core i7-9700K 3.60 GHz with 8 core processors and Nvidia Gefore RTX 2080.

To evaluate the developed model, we calculated the validation accuracy and the validation loss for ten epochs (learning cycles). Afterward, we also calculated the precision (P), recall (R), and F-score of each predicated handwriting style. The following Equations (3), (4) and (5) [1] represent their calculations.

$$P = \frac{number\ of\ correctly\ predicted\ handwriting\ styles}{total\ number\ of\ predicted\ handwriting\ styles} \quad (3)$$

$$R = \frac{number\ of\ correctly\ predicted\ handwriting\ styles}{total\ number\ of\ relavent\ handwriting\ styles\ in\ the\ dataset} \quad (4)$$

$$F\text{-}score = 2 * (P * R)|\ (P + R0) \quad (5)$$

The dataset was divided into three main categories as the following: training, testing, and validation. The training portion of the dataset should always contain the largest amount of data to train the model successfully. Therefore, we assigned 70% (1857) from the original size of the dataset, which is (2653) images for the training purpose. The rest 30% of the remaining data that has never been seen by the model were divided equally between the testing and the validation sub-sets. Hence, 398 images used for the testing purpose, as well as; another 398 images used for the validation purpose.

This categorization performed based upon the default settings used for running most deep learning models. However, we wanted to check the impact of changing the datasets ratios on the performance of the learning algorithm. Thus, we repeated the experiment with two more different datasets categorizations. The first used categorization is 60% for the training purpose, and the remaining 40% is divided equally between the testing and the validation sub-sets. Whereas, the second used ratio is 80% for the training and the rest 20% is for the testing and validation purposes.

After preparing the dataset, we assigned a small learning rate that equals "1e-3" to allow the model to learn the extracted features more efficiently. Then, we trained the model by running the learning algorithm ten times (10 epochs).

We ran the same MobileNetV1 customized pre-trained model 10 times and re-used the same hyperparameters on all the three datasets ratios to ensure the fairness of implementation. Eventually, we recorded the precision, recall, and the F-score per each handwriting style and for the three experimented datasets ratios, as presented in Table 2.

Table 2. Evaluation parameters per each handwriting style.

Handwriting Style ID	60% Training, 40% Testing and Validation			70% Training, 30% Testing and Validation			80% Training, 20% Testing and Validation		
	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
1	0.0000	0.0000	0.0000	**1.0000**	0.8750	0.9333	**1.0000**	0.9726	**0.9861**
2	0.6042	0.9355	0.7342	0.9821	0.9821	0.9821	0.9375	0.8955	0.9160
3	0.6610	0.9512	0.7800	0.8452	0.9861	0.9103	0.9275	0.9697	0.9481
4	0.7295	**0.9674**	0.8318	0.9868	**1.0000**	**0.9934**	0.9571	0.9853	0.9710
5	**1.0000**	0.8750	0.9333	**1.0000**	0.9697	0.9846	0.9853	0.9571	0.9710
6	0.9714	0.9315	**0.9510**	**1.0000**	0.9538	0.9764	0.9643	**1.0000**	0.9818

*The highest generated results were highlighted by bold and the lowest generated results were highlighted by red colour for easier visualization.

From Table 2, we notice that the 60% training and 40% testing and validation was the worst-performing ratios of datasets. That is because the first handwriting style was having (0.0000) for all the evaluation parameters, which means that the model was not able to predict the first handwriting style. That is most likely due to the small size of the training dataset, which didn't allow the model to see any images written using the first handwriting style, and hence, it was not capable of recognizing it.

However, there was a fluctuation between the highest recorded evaluation metrics. Since the highest recorded precision was by the fifth handwriting style, which was 100% recognized and successfully predicted by the model. While the highest recorded recall was by the fourth handwriting style as (0.9674), and the highest recorded F-score was by the last handwriting style as (0.9510). We conclude that the best predicted Arabic handwriting styles using the 60% training and 40% testing and validation were the last three handwriting styles because they generated the highest metrics. While the worst predicted Arabic handwriting style was the first one named "Al-Nask".

The second experimented ratios of the datasets, which are 70% training and 30% testing and validation performed very well since it recorded (1.0000), which is the best result we can achieve, under the precision evaluation parameter for three handwriting styles. As well as, it recorded (1.0000) under the recall for the fourth handwriting style. The model also generated the best F-score for the fourth handwriting style as (0.9934).

Regarding the lowest recorded results, they were all above 80%, which affirms the effectiveness of the deep learning model in accurately learning and predicting the handwriting styles using the 70% training and 30% testing and validation datasets ratios.

We conclude that the best predicted Arabic handwriting style using the 70% training and 30% testing and validation was the fourth one since it recorded the highest recall and F-score. In contrast, the worst predicted Arabic handwriting styles were the first and the third because it includes the smallest metrics.

Increasing the size of the training portion from the Arabic manuscripts dataset to become 80% instead of 70%, it also generated good satisfying results. Sine, the first handwriting style recorded the highest precision as (1.0000) and the highest F-score as (0.9861). Moreover, the last handwriting style recorded the highest recall as (1.0000). In contrast, the lowest recorded precision was by the third handwriting style as (0.9275). While the lowest recorded recall and F-score was by the second handwriting style as (0.8955) and (0.9160), respectively.

Hence, we conclude that the best predicted Arabic handwriting styles using the 80% training and 20% testing and validation were the first one named "Al-Nask" and the last one named "Al-Farsi". On the other hand, the worst predicted Arabic handwriting style was the second one named "Al-Thulth".

Eventually and after analyzing the results generated using the three different ratios of datasets, we admit that the 60% training and 40% testing and validation was the worst-performing categorization. On the other hand, both the other categorizations of datasets were generating better results.

To confirm our reached conclusion and to more accurately assess the performance of our developed model, we computed the final validation accuracy and the validation loss of the last tenth learning cycle. In addition, we recorded the AP, Average Recall (AR), and Average F-score (AF) for each one of the three experimented datasets ratios as illustrated in Table 3.

Table 3. Computed metrics for different datasets ratios.

Metric	60% Training, 40% Testing, and Validation	70% Training, 30% Testing, and Validation	80% Training, 20% Testing, and Validation
Validation Accuracy	0.8164	0.9583	0.9375
Validation Loss	2.8124	0.5664	0.8353
Average Precision	0.6610	0.9690	0.9619
Average Recall	0.7768	0.9611	0.9634
Average F-Score	0.7051	0.9633	0.9623

Considering that the accuracy should be high, and the loss should be low for the model to achieve good results, and based on the results in Table 3, we notice that the 60% training and 40% testing and validation generated the lowest results for all the evaluation parameters. In contrast, the 70% training and 30%

testing and validation generated the highest results for all the metrics except the average recall. That is because the highest recorded average recall was using the 80% training and 20% testing and validation. Therefore, the generated results in Table 3 proved that the ultimate dataset categorization is 70% for the training portion and 30% for both the testing and the validation portions.

To further assess the performance of the developed model, we generated three confusion matrices illustrated in Figures 3, 4, and 5. The matrices predict learned handwriting styles for all manuscripts' images and compare them with the ground truth ones. Thus, the rows in the matrices represent the six actual or true handwriting styles. While, the columns represent the same six handwriting styles but, the predicted and not the true ones. The confusion matrices include the test portion of the original dataset. Therefore, the sum of the total numbers written inside the confusion matrix in Figure 3 equals (531) images, which constitute only 20% from the complete original dataset (2653). In other words, the classifier made (531) predictions. Similarly, the total numbers inside the confusion matrix in Figure 4 constituting 15% only from the original dataset, and it equals to (398) images. While, the sum of the total numbers inside the generated confusion matrix in Figure 5 equals (266), which constitutes 10% of the size of the original dataset.

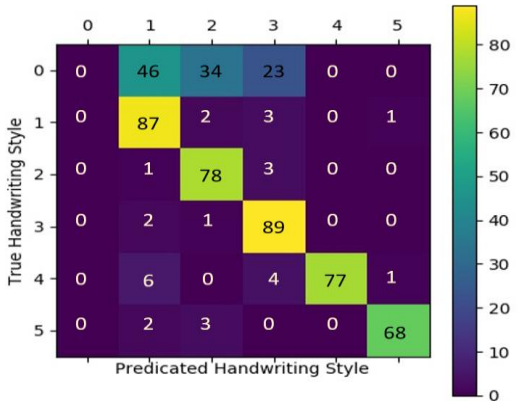


Figure 3. Generated confusion matrix by the 60% training, 40% testing, and validation datasets ratios.

It is clear from Figure 3 that the model performed the worst in predicting the first handwriting style, which is numbered (0) in the confusion matrix and named "Al-Nask" in our dataset. That is because the model couldn't predict any image written using "Al-Nask" successfully. On the other hand, the model was performing well in successfully predicting the rest five handwriting styles. That is because the diagonal is appearing clearly inside the confusion matrix, and it includes all the big numbers, which represent the correct predictions of the Arabic handwriting styles.

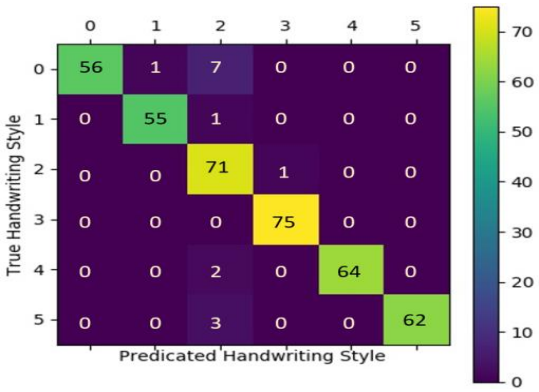


Figure 4. Generated confusion matrix by the 70% training, 30% testing and validation datasets ratios.

Figure 4 illustrates that the deep learning model found 75 images written by the fourth handwriting style, which is denoted by number (3) in the confusion matrix and named "Al-Hur" without any miss-prediction of images written using "Al-Hur" handwriting style. This result illustrates that the model was performing the best in predicting the fourth handwriting style. However, the maximum number of miss-predicted images found under the first handwriting style numbered (0) and named "Al-Nask". Because seven images were miss-predicted as written using "Al-Reqaa" handwriting style and one image was miss-predicted as written using "Al-Thulth" handwriting style.

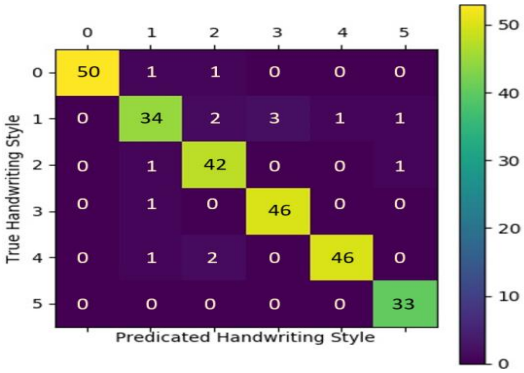
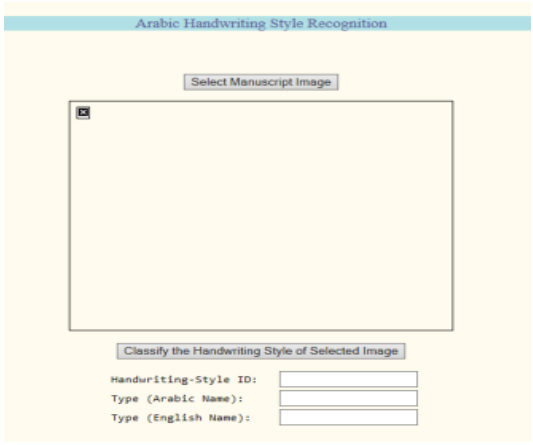


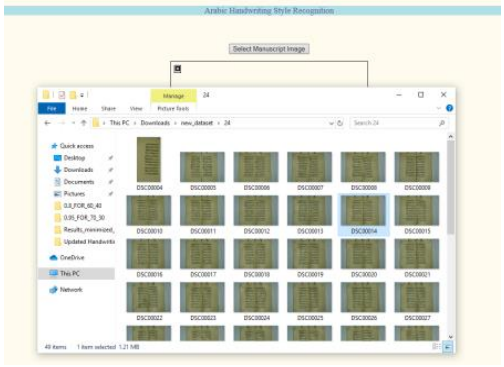
Figure 5. Generated confusion matrix by the 80% training, 20% testing, and validation datasets ratios.

From Figure 5, we notice that the highest number of successfully predicted images found under the first handwriting style, which is denoted by (0) in the confusion matrix and named "Al-Nask". That is because the model correctly classified 50 images to be written using it. Another successful result generated by the model found under the last handwriting style that is denoted as (5) in the confusion matrix and named "Al-Farsi". Since the model classified all found 33 images successfully without any miss-predictions. About the worst predicted handwriting style, it is found under the second handwriting style that is denoted as (1) and named "Al-Thulth". Because the model made the maximum number of miss-predictions under it reaching seven miss-predicted images.

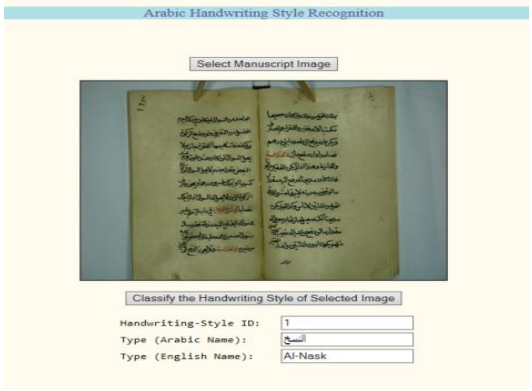
To clarify the environment of the graphical user interface of the proposed model, we developed the application illustrated in Figure 6.



a) Main interface of the model.



b) Arabic manuscript image selection.



c) Handwriting style classification result.

Figure 6. Graphical user interface of the proposed model.

Figure 6-a illustrates the main interface of the developed application to predict the handwriting styles of the Arabic manuscripts' images. From the main interface, the user selects the image that s/he intends to retrieve its handwriting style as shown in Figure 6-b Finally, the user should press on the "Classify the Handwriting Style of Selected Image" button to be able to visualize the results as shown in Figure 6-c.

Table 4 compares the proposed method with some state-of-the-art methods discussed in the related work section of the study. The chosen methods are all classify and predict handwriting styles. The comparison is according to the approach used for the feature extraction and classification. It is also according to the selected dataset language and size, as well as, the number and types of the predicted handwriting styles.

Table 4. Comparison with state-of-the-art methods.

Reference	Feature Extraction	Classification	Dataset			Handwriting Styles		Results
			Name	Size	Lang.	No.	Type	
Allaf and Al-Hmouz [2]	Genetic algorithm	Neural network module	Local dataset written by calligraphers	89 images	Arabic	3	• Thuluth • Reqaa • Kufi	8.02% recognition error rate
			Public dataset generated by the Computer	113284 images	Arabic	10	• AdvertisingBold • Andalus • ArabicTransparent • DecoTypeNaskh • DecoTypeThuluth • DiwaniLetter • MUnicodeSara • SimplifiedArabic • Tahoma • TraditionalArabic	7.55% recognition error rate
Bataineh *et al.*[7]	Edge direction Matrices (EDMS)	Backpropagation neural network	Selected Arabic images	14 images	Arabic	7	• Thuluth • Andalusi • Diwani • Persian • Kufi • Naskh • Roqaa	43.7% recognition accuracy
Ezz *et al.* [10]	SIFT algorithm	Gaussian Naive Bayes (GNB)	Two historical Islamic Arabic books	200 images	Arabic	2	• Naskh • Reqaa	92% recognition accuracy
Yu-Sheng *et al.* [23]	HOG descriptor	Softmax Regression	Single character images	2000 images	Chinese	5	• Seal • Clerical • Regular • Running • Cursive	95.55% recognition accuracy
Proposed method	Transfer learning from MobileNet_V1_100_244 deep learning model	Softmax dense layer	Collected historical Arabic manuscripts	2653 images	Arabic	6	• Al-Nask • Al-Thulth • Al-Reqaa • Al-Hur • Al-Diwani • Al-Farsi	95.83% recognition accuracy

From Table 4, we notice that the proposed method is outperforming other existing methods since it recorded the highest evaluation parameters. That is because the proposed method is utilizing deep learning features that are automatically recognized and classified by the model using its deep convolutional layers. Whereas, the other state-of-the-art methods are using classical learning techniques that require manual extraction of the hand-crafted features.

5. Conclusions

This paper introduces a novel approach that applies trending technology to preserve our significant Arabic cultural values. Since we developed a deep learning model to classify and predict the handwriting styles of Arabic manuscripts' images.

The study started by collecting the dataset manually and considering the found six handwriting styles that exist in the collected dataset, which are: Al-Nask, Al-Thulth, Al-Reqaa, Al-Hur, Al-Diwani, and Al-Farsi.

The original dataset categorized into three main subsets, and we experimented three different ratios of the datasets as follows:

- Allocate 60% from the original size of the main dataset for the training purpose and divide the rest 40% equally between the testing and the validation subsets.
- Assign 70% from the data for the training and split the rest 30% equally between the testing and the validation subsets.
- Allocate 80% from the data for the training and split the remaining 20% evenly between the testing and validation.

Afterward, we transferred learning from the pre-trained MobileNetV1 deep learning model to extract and classify the visual features in the images automatically.

To evaluate the proposed method, both the validation accuracy and the validation loss calculated for each dataset ratio, as well as the precision, recall, and the F-score computed for each predicted handwriting style.

We concluded that assigning 60% from the size of the original dataset for the training purpose didn't perform well because it generated the lowest evaluation parameters among the three tested dataset ratios. Since it recorded (0.8164) accuracy and (0.7051) average F-score.

Even though increasing the size of the training portion from the original dataset to become 80% might bias the learning operation, we experimented this option, and we were expecting the learning performance to rise as we increase the ratio of the training portion from the original dataset. Instead, we found that the prediction problem generated lower results since the model recorded (0.9375) accuracy and (0.9623) average F-score. On the other hand, the model reached the highest results using the 70% portion of the data for the training and 30% of the data for the testing and validation since it recorded (0.9583) accuracy and (0.9633) F-score. Therefore, we admit that the deep learning model performed well with the image's classification and prediction process and was able to predict the handwriting styles successfully utilizing the 70% training and 30% testing and validation datasets ratios.

Future work might include comparing more than one deep learning models looking for the most accurate model in predicting the Arabic handwriting styles. We can also try to do the training from scratch instead of transfer learning from pre-trained models and compare the results.

References

[1] Alaei F., Alaei A., Pal U., and Blumenstein M., "A Comparative Study of Different Texture Features for Document Image Retrieval," *Expert Systems with Applications*, vol. 121, pp. 97-114, 2018.

[2] Allaf S. and Al-Hmouz R., "Automatic Recognition of Artistic Arabic Calligraphy Types," *King Abdulaziz University Scientific Publishing Center*, vol. 27, no. 1, pp. 3-17, 2016.

[3] Altoe P., "Class Lecture, Topic: "Fundamentals of Deep Learning for Computer Vision," Supercomputing Laboratory, King Abdullah University of Science and Technology, KAUST, Jeddah, 2019.

[4] Al-Ayyoub M., Nuseir A., Alsmearat K., Jararweh Y., and Gupta B., "Deep learning for Arabic NLP: A Survey," *Journal of Computational Science*, vol. 26, pp. 522-531, 2018.

[5] Al-Jawfi R., "Handwriting Arabic Character Recognition LeNet Using Neural Network," *The International Arab Journal of Information Technology*, vol. 6, no. 3, pp. 304-309, 2009.

[6] Al-Yahya M., "Stylometric Analysis of Classical Arabic Texts for Genre Detection," *The Electronic Library*, vol. 36, no. 5, pp. 842-855, 2018.

[7] Bataineh B., Abdullah S., and Omar K., "Arabic Calligraphy Recognition Based on Binarization Methods and Degraded Images," *in Proceedings of International Conference on Pattern Analysis and Intelligent Robotics*, Putrajaya, pp. 65-70, 2011.

[8] Chan T., Jia K., Gao S., Lu J., Zeng Z., and Ma Y., "PCANet: A Simple Deep Learning Baseline for Image Classification," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5017-5032, 2015.

[9] Das S., "CNN Architectures: LeNet, AlexNet, VGG, GoogLeNet, ResNet and more…,"

https://medium.com/analytics-vidhya/cnns-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more-666091488df5, Last Visited 2017.

[10] Ezz M., Sharaf M., and Hassan A., "Classification of Arabic Writing Styles in Ancient Arabic Manuscripts," *International Journal of Advanced Computer Science and Applications*, vol. 10, no.10, pp. 409-414, 2019.

[11] Howard A., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., and Adam H., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," pp. 1-9, 2017.

[12] Huang K., Hussain A., Wang Q., and Zhang R., *Deep Learning: Fundamentals, Theory and Applications*, Springer International Publishing, 2019.

[13] Le Q. and Mikolov T., "Distributed Representations of Sentences and Documents," *in Proceedings of the 31st International Conference on Machine Learning*, Beijing, pp. 1-5, 2014.

[14] Liang H., Sun X., Sun Y., and Gao Y., "Text Feature Extraction Based On Deep Learning: A Review," *Eurasip Journal on Wireless Communications and Networking*, vol. 211, no. 1, pp. 1-12, 2017.

[15] Liu H., Li B., Lv X., and Huang Y., "Image Retrieval Algorithm Based on Convolutional Neural Network," *Current Trends in Computer Science and Mechanical Automation*, vol. 133, pp. 304-314, 2017.

[16] Mohamed O., Khalid E., Mohammed O., and Brahim A., *Content-Based Image Retrieval Using Convolutional Neural Networks*, Springer International Publishing, 2019.

[17] Reddy A. and Krishna C., "A Survey on Applications and Performance of Deep Convolution Neural Network Architecture for Image Segmentation," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 19, pp. 43-60, 2018.

[18] Saleh A., *تاريخ الخط العربي عبر العصور المتعاقبة*, Scientific Books Publishing, 2017.

[19] Seddati O., Dupont S., Mahmoudi S., and Parian M., "Towards Good Practices for Image Retrieval Based on CNN Features," *in Proceedings of IEEE International Conference on Computer Vision Workshops*, Venice, pp. 1246-1255, 2018.

[20] Tyagi V., *Research Issues for Next Generation Content-Based Image Retrieval*, Springer Singapore, 2017.

[21] Wang J., Yang Y., Mao J., Huang Z., Huang C., and Xu W., "CNN-RNN: A Unified Framework for Multi-Label Image Classification," *Journal of the Japanese Society for Cancer Therapy*, vol. 13, pp. 245-246, 2016.

[22] Yahia M., Content-Based Retrieval of Arabic Historical Manuscripts Using Latent Semantic Indexing, Thesis, King Fahd University of Petroleum and Minerals, 2011.

[23] Yu-Sheng C., Guangjun S., and Haihong L., "Machine Learning for Calligraphy Styles Recognition,"

[24] Zhou W. and Jia J., "A Learning Framework for Shape Retrieval Based on Multilayer Perceptrons," *Pattern Recognition Letters*, vol. 117, pp. 119-130, 2018.

Manal Khayyat received the B.Sc. degree (Hons.) in Computer Science from King Abdulaziz University, Saudi Arabia, in 2007 and received M.Sc. degree of Applied Science in Quality Systems Engineering from Concordia University, Canada, in 2015. She is currently a PhD student in the Department of Computer Science at King Abdulaziz University. She worked at the IT department of Effat University, Saudi Arabia, from 2007 to 2010. Then, she worked as a lecturer at King Abdulaziz University, from 2012 to 2019 and she is currently working as a lecturer at Umm Al-Qura University, Saudi Arabia. Her research interests include computer vision, image processing, natural language recognition, and deep learning.

Lamiaa Elrefaei received the B.Sc. degree (Hons.) in electrical engineering (electronics and telecommunications), and the M.Sc. and Ph.D. degrees in electrical engineering (electronics) from the Faculty of Engineering at Shoubra, Benha University, Egypt, in 1997, 2003, and 2008, respectively. She held a number of faculty positions at Benha University, as a Teaching Assistant, from 1998 to 2003, as an Assistant Lecturer, from 2003 to 2008, and has been a Lecturer, since 2008. She is currently an Associate Professor with the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. Her research interests include computational intelligence, biometrics, multimedia security, wireless networks, and nano networks.