Journal of Studies in Science and Engineering

EngiScience Publisher

Research Article

Comparison of Handwritten Recognition Methods on Arabic and Latin Characters

Mehmet Tutar [1,*]

[1] Department of Computer Engineering. Faculty of Engineering, Harran University, Şanlıurfa, 63050, Turkey

*Corresponding Author: Mehmet Tutar, E-mail: mtutar27@hotmail.com

Article Info

Article History

Received Jul 31, 2022

Revised Aug 12, 2022

Accepted Aug 14, 2022

Keywords

OCR

MNIST

MADBase

Arabic character recognition

Latin character recognition

Abstract

In this article, machine learning techniques and deep learning methods were applied to the digit datasets created using the Arabic and Latin alphabets, and the methods' performances were compared. Furthermore, each method was tested with various parameters, and the results were analyzed. In addition, this study also observed the recognizability of handwritten numeral datasets created using different alphabets. For experiments, an Arabic alphabet handwritten digit dataset (60,000 training and 10,000 testings) and a Latin alphabet handwritten digit dataset (60,000 training and 10,000 testings) were used. When the results of the experiment are examined, it is seen that successful results are obtained in the classification made with the MADBase dataset in some methods, and the classification made with the MNIST dataset in some methods. As a result, it can be stated that a method's handwriting character recognition success cannot be measured only by the classification made on a dataset. Also, the digits in the Arabic alphabet appear to be almost more recognizable than those in the Latin alphabet.

1. Introduction

Throughout history, handwriting recognition and comprehension have been seen as a problem to be solved, and different languages and alphabets have been developed for this. However, many algorithms and learning methods for handwriting recognition have been developed and continue to be developed today. While developing handwriting recognition algorithms and methods, datasets can be used where needed, and in this context, datasets were created for commonly used alphabets, although not for all alphabets.

Pattern recognition, which is the most important stage of character recognition, was first taken into account by Grimsdale, and a study was conducted in this context [1]. However, most research on handwritten character recognition is based on a technique known as analysis by synthesis, proposed by Eden in 1968 [2]. This technique has been used in all methods, and structural character recognition approaches over time.

The creation of the CEDAR dataset, one of the handwritten Latin character datasets, dates back to 1994. It contains characters for handwritten words, letters, and numbers, such as city names and postal codes [3].

In 1995, NIST published the "Special Dataset 19," which consists of Latin script and cursive characters (letters and numbers). The NIST dataset has 62 labeled classes for numbers and letters [4]. In 1998, a 28x28 pixel MNIST (Modified-NIST) dataset, consisting of numbers (10,000 tests, 60,000 trainings), was published using the dataset published by NIST [5]. MNIST, consisting of Latin letters, is frequently used by researchers in many studies as a handwritten digit dataset.

Like the Latin alphabet, the Arabic alphabet is a universal one known worldwide. Arabic is officially used in 25 more than 300 million countries [6-8]. For this reason, the Arabic alphabet, like the Latin alphabet, is frequently the subject of research in handwritten character classification studies, and many studies are carried out in this field.

In this article, both machine learning techniques and deep learning methods were applied to handwritten digit datasets created using Arabic and Latin alphabets, and the recognition performances of the methods were compared. Some machine learning methods were applied with various parameters on handwritten datasets, and the results were analyzed. Likewise, the parameters specified by the authors in Deep learning methods were applied to handwritten datasets, and the results were examined. In addition, this study also compared the recognizability of handwritten digit datasets created using different alphabets (Latin and Arabic). The Arabic handwritten digit dataset MADBase [10] (60,000 training and 10,000 tests) and the Latin handwritten digit dataset MNIST [5] (60,000training and 10,000 tests) were used.

2. Related Works

Character datasets such as Arabic [9-12] and Latin [4, 13, 14] were created from different alphabets and languages to be used in handwriting character recognition. There are handwritten character recognition studies on many languages and alphabets in the literature. The literature review reveals some studies on handwritten character recognition and development.

It is reported that a 94% accuracy rate was obtained in the CNN-based handwritten digit recognition study conducted by the researchers using the MNIST dataset [15]. K-nearest neighbor (KNN) is a basic classification method that can be used for classification when the data distribution is unknown [16].

Karakaya states that in her study, where machine learning methods are applied to the MNIST dataset, the methods that provide the highest efficiency are the K-nearest neighbor algorithm and the K-mean algorithm [17]. In another study using Convolutional Neural Networks (CNN) with an MNIST dataset, 98.51% accuracy was reported for handwritten digit recognition [18]. Support Vector Machine (SVM) is a method proposed by Vapnik and attracting much attention in machine learning research. Numerous recent studies have shown that SVM can generally provide better overall performance, achieving better type accuracy than alternative data classification algorithms [19].

In this study, in which the MNIST dataset was used, it is stated that the highest accuracy rate of 95.88% among all machine learning algorithms was achieved with SVM, one of the machine learning algorithms, for the identification of handwritten digits [20].

In this study, in which a CNN-SVM hybrid model is proposed for handwritten digit recognition, it is reported that 99.28% classification accuracy is achieved for the MNIST dataset by combining the advantage of CNN and SVM classifiers in handwriting digit recognition [21]. Furthermore, it is stated that SVM, MLP (Multilayer Perceptron), and CNN methods based on Machine learning and Deep learning algorithms are applied by using MNIST, and CNN gives more successful results for handwritten digit recognition [22].

3. Materials and Methods

Many machine learning and Deep learning methods are used in handwriting character recognition studies. Therefore, the literature review decided to use our experiments' KNN and SVM models from machine learning methods, LeNet5 [5] and Effective CNN [17] models from deep learning methods. In addition, the Arabic alphabet handwritten digit dataset MADBase [10] (60,000 training and 10,000 tests) and the Latin alphabet handwritten digit dataset MNIST [5] (60,000 training and 10,000 tests) were used.

3.1. MNIST Database

MNIST dataset; It was published as 28x28 pixels, formatted and preprocessed, composed of only numbers (60,000 training, 10,000 tests), written in the Latin alphabet by approximately 500 participants [5]. Handwriting is frequently used in number recognition applications. Some character examples from the MNIST dataset are shown in Figure 1.

Figure 1. Random character samples from MNIST [5] dataset (28x28 pixels)

3.2. MADBase Database

ADBase dataset; It was created by writing in the Arabic alphabet by approximately 700 participants, consisting of digits (60,000 training examples, 10,000 test examples). In order to have the same format and size as the MNIST dataset, the ADBase dataset was modified to obtain the MADBase dataset [10]. Some character examples of the MADBase dataset are shown in Figure 2.

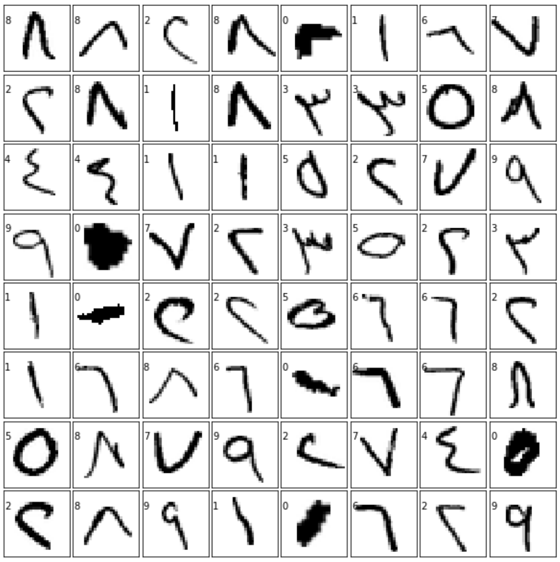

Figure 2. Random character samples from the MADBase [10] dataset (28x28 pixels)

3.3. Methodology

Machine learning (ML) methods KNN and SVM, deep learning (DL) methods LeNet5 [5], and Effective CNN [17] were applied on MNIST (60,000 training examples, 10,000 test examples) and MADBase (60,000 training examples, 10,000 test examples) handwritten digit datasets.

3.3.1 Machine Learning Methods

KNN (k-Nearest Neighbor) is one of the simple classification methods that can be used for classification when there is no prior knowledge about the data distribution [15]. Cosine and Euclidean distance metrics were used in experiments with KNN. In our experiments with KNN, we used the k metric value as 3, 5, and 7, considering the literature studies [23].

SVM is a frequently used method in machine learning research. It provides both easy-to-apply and successful results on linear and non-linear data. In our experiments with SVM, RBF, Linear, and Poly kernel metrics were used.

3.3.2 Deep Learning Methods

LeNet5 [5] is the oldest convolutional neural network, which takes the grayscale image as input and consists of a multilayer structure with learnable parameters. LeNet5 [5] model was applied to MNIST and MADBase datasets. In experiments with LeNet5, epochs size 30 and batch size 128 were used. The LeNet5 Model summary is shown in Figure 3.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 28, 28, 6)         156

 max_pooling2d (MaxPooling2D  (None, 14, 14, 6)        0
 )

 conv2d_1 (Conv2D)           (None, 10, 10, 16)        2416

 max_pooling2d_1 (MaxPooling  (None, 5, 5, 16)         0
 2D)

 flatten (Flatten)           (None, 400)               0

 dense (Dense)               (None, 120)               48120

 dense_1 (Dense)             (None, 84)                10164

 dense_2 (Dense)             (None, 10)                850

=================================================================
Total params: 61,706
Trainable params: 61,706
Non-trainable params: 0
_________________________________________________________________
```

Figure 3. LeNet5 model summary

The Effective CNN [17] model consists of feature extraction with binary classification and convolution. In experiments with Effective CNN, epochs size 100 was used. The Effective CNN Model summary is shown in Figure 4.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 26, 26, 32)        320

 max_pooling2d (MaxPooling2D  (None, 13, 13, 32)       0
 )

 conv2d_1 (Conv2D)           (None, 11, 11, 64)        18496

 max_pooling2d_1 (MaxPooling  (None, 5, 5, 64)         0
 2D)

 flatten (Flatten)           (None, 1600)              0

 dense (Dense)               (None, 128)               204928

 dense_1 (Dense)             (None, 10)                1290

=================================================================
Total params: 225,034
Trainable params: 225,034
Non-trainable params: 0
_________________________________________________________________
```

Figure 4. Effective CNN model summary

Experiments with KNN, SVM, LeNet5, and Effective CNN are performed in a local machine (Core i5-9300H 2.40GHz, 16 GB DDR4, and GPU NVIDIA GeForce).

4. Results and Discussions

Handwriting digit recognition experiments were carried out by using ML and DL methods on MNIST datasets created with the Latin alphabet and MADBase created with the Arabic alphabet, the performance of the methods was measured, and the results were compared. The results of the experiments with the KNN method are shown in Table 1, the results of the experiments with the SVM method in Table 2, and the results of the experiments with the DL (LeNet5 and Effective CNN) methods are shown in Table 3.

Table 1. Classification results with the KNN method

Datasets	Metrics	k Values	Test Accuracy (%)
MNIST	Cosine	3	97.33
		5	97.30
		7	97.27
	Euclidean	3	97.05
		5	96.88
		7	96.94
MADBase	Cosine	3	98.20
		5	98.20
		7	98.20
	Euclidean	3	98.13
		5	98.14
		7	98.10

When Table 1 is examined, it is seen that the MADBase dataset written in the Arabic alphabet has a better recognition rate than the MNIST dataset written in the Latin alphabet. Likewise, when examined in terms of KNN metrics, it is understood that the Cosine distance metric is more successful in both data sets than the Euclidean distance metric.

Table 2. Classification results with the SVM method

Datasets	Metrics	Test Accuracy (%)
MNIST	Gaussian	97.92
	Linear	76.49
	Poly	97.71
MADBase	Gaussian	98.52
	Linear	86.76
	Poly	97.68

Table 2. shows that the MADBase dataset written in the Arabic alphabet has a better recognition rate than the MNIST dataset written in the Latin alphabet. Likewise, when analyzed in terms of SVM metrics, it is seen that the Gaussian kernel metric is more successful than the Linear and Poly metrics in both data sets.When Table 1 and Table 2 are examined together, the MADBase dataset written in the Arabic alphabet exhibits better recognition performance in Machine learning methods compared to the MNIST dataset written in the Latin alphabet.

Table 3. Classification results with DL (LeNet5 [5] and Effective CNN [17]) methods

Datasets	Methods	Test Accuracy (%)
MNIST	LeNet5	98.49
	Effective CNN	99.33
MADBase	LeNet5	98.79
	Effective CNN	99.11

When Table 3 is examined, it is seen that the MADBase dataset written in the Arabic alphabet exhibits better classification performance with the LeNet5 method. Likewise, it is observed that the MNIST dataset, written in the Latin alphabet, exhibits better classification performance with the Effective CNN method. Considering the parameter numbers of the LeNet5 and Effective CNN methods (Figure 3 and Figure 4), it is thought that the MADBase dataset shows better classification success with the methods containing fewer parameters.

When Table 1, Table 2, and Table 3 are analyzed together, it is seen that the MADBase dataset written in the Arabic alphabet has better classification success in many methods. This may be due to the diversity of participants (MNIST 500 participants, MADBase 700 participants). In addition, it can be said that the Arabic alphabet is more recognizable than the Latin alphabet. However, to make the final decision, it is thought that the number of participants should be the same and the participants should consist of the same people, with the number of images being equal.

When all the experimental results are examined, it is seen that Deep learning methods provide better classification results than Machine learning methods. However, it is understood that a model or method does not show the same recognition success in datasets written in different alphabets. It is thought that this problem can be overcome by considering datasets written in different alphabets while developing the model.

5. Conclusions

In this article, some of the known machine learning and deep learning methods are applied to handwritten digit datasets created with Arabic and Latin alphabets, and the results are compared. KNN and SVM were applied from machine learning methods, LeNet5 [5], and Effective CNN [17] models from deep learning. Experiments show that Arabic numerals are almost more recognizable than Latin numerals. When the results obtained are examined, it is thought that the recognition success of a method or a developed model cannot be measured with a dataset written using only one alphabet, and its real performance can be revealed by classification with datasets written in different alphabets. Therefore, in our future study, it is considered to compare the performances of frequently used methods with datasets composed of handwritten letters written in different alphabets and to analyze the results.

Declaration of Competing Interest: The author declares that he has no conflict of interest.

References

[1] Grimsdale, R. L., Sumner, F. H., Tunis, C. J., & Kilburn, T., 1959. A system for the automatic recognition of patterns. Proceedings of the IEE-Part B: Radio and Electronic Engineering, 106(26), 210-221.

[2] Megha Agarwal, Shalika, Vinam Tomar, Priyanka Gupta., 2019. "Handwritten Character Recognition using Neural Network and Tensor Flow", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue- 6S4.

[3] Hull, J. J., 1994. A database for handwritten text recognition research. IEEE Transactions on pattern analysis and machine intelligence, 16(5), 550-554.

[4] Grother, P. J., 1995. Nist special database 19-hand-printed forms and characters database. Technical Report, National Institute of Standards and Technology.

[5] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324. http://yann.lecun.com/exdb/mnist/

[6] AlKhateeb, J. H., Ren, J., Ipson, S., and Jiang, J., 2008. Knowledge-based baseline detection and optimal thresholding for words segmentation in efficient preprocessing of handwritten Arabic text. Fifth international conference on information technology: new generations. IEEE computer society. pp. 1158-1159.

[7] Khorsheed, M.S., 2002. Off-line Arabic character recognition a review. Pattern Analysis & Applications. 5(1): 31-45.

[8] Liana, M., and Venu, G., 2006. Offline Arabic Handwriting Recognition: A Survey. IEEE, Transactions on Pattern Analysis and Machine Intelligence. 28: 712-724.

[9] Al-Ohali, Y., Cheriet, M., & Suen, C., 2003. Databases for recognition of handwritten Arabic cheques. Pattern Recognition, 36(1), 111-121.

[10] El-Sherif, E. A., & Abdelazeem, S., 2007. A Two-Stage System for Arabic Handwritten Digit Recognition Tested on a New Large Database. In Artificial intelligence and pattern recognition (pp. 237-242).

[11] Mahmoud, S. A., Ahmad, I., Alshayeb, M., Al-Khatib, W. G., Parvez, M. T., Fink, G. A., ... & El Abed, H., 2012. Khatt: Arabic offline handwritten text database. In 2012 International conference on frontiers in handwriting recognition (pp. 449-454). IEEE.

[12] Jbrail, M.W. and Tenekeci, M.E., 2022. Character Recognition of Arabic Handwritten Characters Using Deep Learning. Journal of Studies in Science and Engineering, 2(1), pp.32-40.

[13] Bartos, G. E., Hoşcan, Y., Kauer, A., & Hajnal, É. N., 2020. A Multilingual Handwritten Character Dataset: THE Dataset. Acta Polytechnica Hungarica, 17(9).

[14] Marti, U. V., & Bunke, H., 2002. The IAM-database: an English sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition, 5(1), 39-46.

[15] Wu, H., 2018. CNN-Based Recognition of Handwritten Digits in MNIST Database. Research School of Computer Science. The Australia National University, Canberra.

[16] Peterson, L. E., 2009. K-nearest neighbor. Scholarpedia, 4(2), 1883.

[17] Karakaya, R., 2020. Makine öğrenmesi yöntemleriyle el yazısı tanıma (Master's thesis, Sakarya Üniversitesi).

[18] Bharadwaj, Y. S., Rajaram, P., Sriram, V. P., Sudhakar, S., & Prakash, K. B., 2020. Effective handwritten digit recognition using deep convolution neural network. International Journal of Advanced Trends in Computer Science and Engineering, 9(2), 1335-1339.

[19] Vapnik Vladimir, N., 1995. The nature of statistical learning theory.

[20] Gope, B., Pande, S., Karale, N., Dharmale, S., & Umekar, P., 2021. Handwritten digits identification using MNIST database via machine learning models. In IOP Conference Series: Materials Science and Engineering (Vol. 1022, No. 1, p. 012108). IOP Publishing.

[21] Ahlawat, S., & Choudhary, A., 2020. Hybrid CNN-SVM classifier for handwritten digit recognition. Procedia Computer Science, 167, 2554-2560.

[22] Pashine, S., Dixit, R., & Kushwah, R., 2021. Handwritten Digit Recognition using Machine and Deep Learning Algorithms. arXiv preprint arXiv:2106.12614.

[23] Grover, D., & Toghi, B., 2019. MNIST dataset classification utilizing k-NN classifier with modified sliding-window metric. In Science and Information Conference (pp. 583-591). Springer, Cham.