Published Online First:
https://doi.org/000000000
P-ISSN: 2078-8665 - E-ISSN: 2411-7986

Baghdad Science Journal

Deep Learning Function Implications For Handwritten Character Recognition Using Improved Mean Square Error Function

Bahera H. Nayef ****[1]***, ***Siti Norul Huda Sheikh Abdullah[2]***, ***Manal Mohammed*** ***[2,3]***

[1]Department of Computer Science, College of Science, Al Nahrain University, Baghdad, Iraq.
[2]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.
[3]Faculty of Administrative Science, Hadhramout University, AL-Mukalla, Yemen.
*Corresponding author.

Received 25/10/2023, Revised 05/07/2024, Accepted 07/05/2024, Published Online First 20/09/2024

© 2022 The Author(s). Published by College of Science for Women, University of Baghdad. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Loss functions are used for estimating the accuracy of the classification models in Machine Learning. The error loss value measures whether the predicted labels match the true labels or are close to them. Handwritten character recognition is a pattern recognition problem. The progression of pattern recognition applications for offline image classification is rapidly improved by using deep learning techniques. Convolution Neural Networks (CNN) and the activation function are the most commonly used in deep learning. In addition, different loss functions are used to evaluate the performance of the model such as Categorical Cross-Entropy (CCE) and Mean Square Error (MSE). In deep learning techniques, the size of the datasets plays an important role in obtaining high performance, but with traditional MSE, the loss values reach zero in the very early stages of the training process when the dataset size is large, and yet the model accuracy is still in progress. This study proposes a developed loss function via enhancing MSE. The proposed improved MSE is based on dividing the square error by the sum of the predicted label probabilities instead of the total of the sample number. Five datasets are used to test the performance of the proposed modified MSE with the proposed CNN pipeline model in addition to the modified VGG16. The datasets are AHCD, AIA9K, HIJJA, Self-collected, and MNIST. The loss rates of the proposed loss function showed a significant improvement in the accuracy rates for all datasets with the improved MSE in comparison to the CCE.

Keywords: Loss functions; Machine learning; pattern recognition; Handwritten character recognition; Deep learning; Mean Square Error.

Introduction

Regression loss functions such as L2 are usually used with regression problems [1]. MSE is a regression loss function that calculates the summation of the squared difference between the probability of the predicted label and the true label. The results of summation are divided by the size of the sample as the penalty factor which is used to update the weights of the neurons. The model performance is good if the error is a small value, but when it is large, the constructed model fails to predict the true label or class. In classification problems, log loss functions such as Cross-Entropy (CE) loss are used to evaluate the classification model performance. Study [1] also mentioned the usability of regression loss functions with classification problems.

In the [2] study, the researchers proposed using MSE with deep learning to generate geometric shapes. The study stated that the MSE function has suffered from a fast approach to zero. So, they proposed an enhancement of MSE by considering the mean of Maximum and Minimum of MSE. However, the proposed method does not take into consideration the effect of the sample size. Another study was presented by [3] to slow down the learning of misclassified samples. The proposed techniques of this study aim to neglect samples that could be correctly classified in the later stages of training. A recent study [4] proposed a balanced MSE to overcome the imbalance in training label distribution for regression problems only. These presented studies motivate us to propose an enhanced MSE loss function that considers all the training samples regardless of the size of the dataset. It also suits the handwritten character classification that contains various handwriting styles.

The characteristic of MSE with handling noisy, small-size data [5], and data with outliers encourages the researchers to apply it to Arabic Handwritten characters classification. In general, character recognition can be investigated in two directions, printed (typed) characters and handwritten character recognition [6]. The printed characters have the same style and size as any given character. On the other hand, with handwritten characters, there are unlimited styles and sizes of characters for the same individual or among different individuals [7]. Character recognition systems are in two categories Offline systems and Online systems. An online character recognition system involves interpreting the entered character via any automated input device such as touch screens and digitizers at the current run time. So, the recognized characters are represented in computer-generated styles. In the offline character recognition system the characters are captured using a scanner or a digital camera and saved in the computer storage digitally [8].

Since the early nineties, a new era of recognition and classification research has started by using Deep Learning Networks (DNN) [3]. DNN proved its superiority in speech recognition, Natural language processing, and image classification over the classical state-of-the-art machine learning techniques [9-11]. The focus of DNN is to minimize the cost function of the model by learning in-depth [10]. Cross entropy and spare cross-entropy are the most common loss functions used for Handwritten character classification with deep learning approaches [7, 12-15]. However, both mentioned loss functions include calculating the logarithmic values for the error rates. That increases the error rates in a way that is not compatible with the resulting accuracy rate. On the other hand, the MSE function suffers from a fast approach to zero from the early training stage.

The current study contributes to our knowledge by addressing four important issues. First, the paper discusses the performance of the MSE as the loss function for classifying Arabic Handwritten characters. Secondly, it proposes an improved MSE to increase the model performance accuracy by overcoming the fast approach to zero problems of the MSE. Third, collect an Arabic handwritten character dataset to test the proposed improved MSE. The last issue is comparing the performance of the proposed improved MSE with CE and MSE functions.

From the above introduction, some challenges are presented as questions as follows:

1- Does the standard MSE perform well with classification problems?
2- Does the proposed improved MSE overcome the approach problem of MSE?
3- Does the proposed improved MSE outperform the Cross-entropy loss function?

Literature Review

This section presents studies that discuss using deep learning techniques for different problems. The proposed solutions are based on either proposing novel loss functions or altering loss functions with others as shown in the next paragraphs.

Recently a group of researchers in [16] proposed a K-means online routing protocol (KMORP) based on a Markov mobility model for UAV ad hoc networks with different sizes. MSE was used to evaluate the routing protocols in predicting the routes chosen by the data packets from the source to the destination. According to the authors, the MSE showed different results for the proposed protocols. However, the study did not discuss the reason for using MSE rather than other loss functions.

Another recent study [17] discussed the lack of remote-sensing images. The image features were extracted using Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU). The study proposed an optimized features augmentation using an iterative genetic algorithm (IGA). The reported results of the accuracy, precision, and recall showed the superiority of the proposed method over the state-of-the-art studies.

Study [18] proposed a novel loss function that includes an improvement to the Categorical Cross-Entropy (CCE) loss function by imposing Mean Absolute Error (MAE). It showed remarkable performance in Deep learning neural networks. In the study, Categorical Cross-Entropy (CCE) was found to be less robust to the large-scale datasets with noise in the training sample labels. The results showed that the proposed loss function has outperformed both the MAE and CCE individually by classifying images from CIFAR-10, CIFAR-100, and Fashion MNIST datasets with different rates of noise. The authors [2] developed the Mean of Maximum Square Errors (MMaSE) and Mean of Max Absolute errors loss functions based on Mean Square Error and Mean Absolute. Deep learning techniques represented by a convolutional autoencoder were used to generate four 2D geometrics with random size and Dirichlet (probability distribution) boundary conditions. Using ImgeNet dataset discrimination the authors [19] proposed a novel loss function called the adversarial loss function. The proposed loss function is composed of the Generator loss (LG), the Discriminator loss (LD), and the Encoder loss (LE).

For facial expression and recognition problems in deep convolutional learning, the authors [20] proposed a margin-based loss function instead of Cross entropy. Another novel loss function was proposed by [21] for object detection. It is composed of two non-decomposable complex loss functions. The first is the Average Precision (AP) and the second is the Normalized Discounted Cumulative Gain (NDCG). The study used a deep learning model with the Pascal VOC 2007 object detection dataset to evaluate the proposed loss function.

In study [22], the researcher claimed that RMSE is weak and inappropriate for normal distribution problems and triangle inequality. On the contrary, [23, 24] studies proved the priority of RMSE over MAE. The study showed that RMSE is not ambiguous and more appropriate to be used in models with a normal distribution. Moreover, it is good with triangle inequality as a distance function metric. Also, the study has proved that the sensitivity of RMSE to outlier probabilities or noise and redundancy were well defined by RMSR. It was explained in the final discussion that both MAE and RMSE are important in different fields that suit their functionality. Moreover, other researchers have been encouraged to use MAE over RMSE for its statistical evaluation as found [25].

Some other authors attempted to reduce the loss rates by manipulating the used activation function like the study proposed by [26] to enhance the performance of the Time Delay Neural Network (TDNN) model for the prediction of the upcoming serial crime. The study proposed an enhanced Nonlinear Autoregressive with Exogenous Input (eNARX) model based on two activation functions the Tanh and RBF at the same hidden layer. According to the authors, using two fusion activation functions led to minimizing the model error and producing a precise prediction for crime spatiotemporal. Moreover, the study by [27] also proposed an Optimized Leak Relu activation function to enhance the model performance for classifying handwriting characters. The proposed activation function fuses both negative and positive feature maps and this led to an improved accuracy rate and a decrease in the loss error.

From reviewing the above studies, many researchers discussed improving or developing a loss function to improve the model discrimination performance. Also, to overcome the outlier data problem. Other studies are presented to prove the ability to use MSE, MAE, and RMSE in recognition problems as well as classification problems. Especially when using a dataset with noise and outliers.

Mean Square Error Loss Function

The performance accuracy of a neural network can be measured by its effectiveness based on the size of the error value. The error value or the loss is calculated from the difference between the true label and the predicted label presented in the MSE.

A small loss value indicates that the predicted label is the true label. Otherwise, it is not, and the weights of the Neural network nodes must be updated to meet the correct prediction. The mean square error cost function formula is as follows in Eq. 4:

$$MSE = \frac{1}{N\times M}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij} - \hat{y_{ij}}\right)^2 \qquad 1$$

Where:

i=1,2,.....N represents the image height

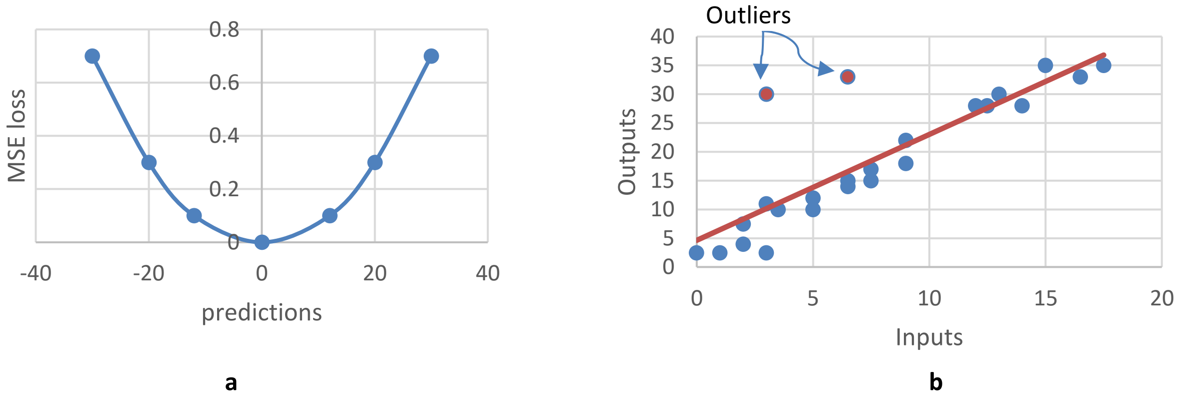
j= 1,2,....M represents the image width

y_{ij} the true label,

$\hat{y_{ij}}$ the predicted label

The modified Mean Square Error loss function

In [28] stated that the MSE function suffers from unboundedness and convexity in both negative and positive infinities. The MSE is great for ensuring that the trained model has no outlier predictions with huge errors since the MSE puts a greater weight on these errors due to the squaring part value of the function as shown in Fig. 1. For that reason, a wrong prediction will magnify the squaring part value. In terms of complexity, this magnification leads to slowing down the learning of misclassified samples.



a

b

Figure 1. (a) The behavior of MSE loss versus prediction error, (b) MSE with outliers [29]

On the other hand, in the classic approach, the denominator $(N \times M)$ is used in the training sample to update the weight of the misclassified samples. This $N \times M$ is a constant value that represents the image size or the total number of pixels in an image. So, suppose different sizes for the same image, assuming the same prediction value the MSE loss decreases and approaches zero when increasing the image size as shown in Fig. 2(a). The loss values fast approach to zero while the model performance still has low accuracy rates.

However, in our research, the result of the square error is a small value ranging between 0 and 1 since it represents the probability difference between the true and the predicted class. So, by dividing the summation of the predicted probabilities of all the class labels in the training epoch as shown in Fig. 2(b), the resulting loss rate and the model performance accuracy start to change Inversely proportional. In other words, at early epochs, the loss rates started relatively high with low accuracy and then decreased with continuing training for many epochs while increasing the accuracy rates proportionally.

$$modified\ MSE\ loss = \frac{\sum_{i=1}^{M} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{M} \hat{y}_i}, \qquad 2$$

where:
$\sum_{i=1}^{M} \hat{y}_i = n\hat{y}_i =$ the fitness penalty applied on the square error and update the unit weights.

y_i = true label of the input sample i

$\hat{y}_i =$ predicted label probability of i$^{\text{th}}$ input sample. Eqs. 3-7, show the first derivative of the modified MSE

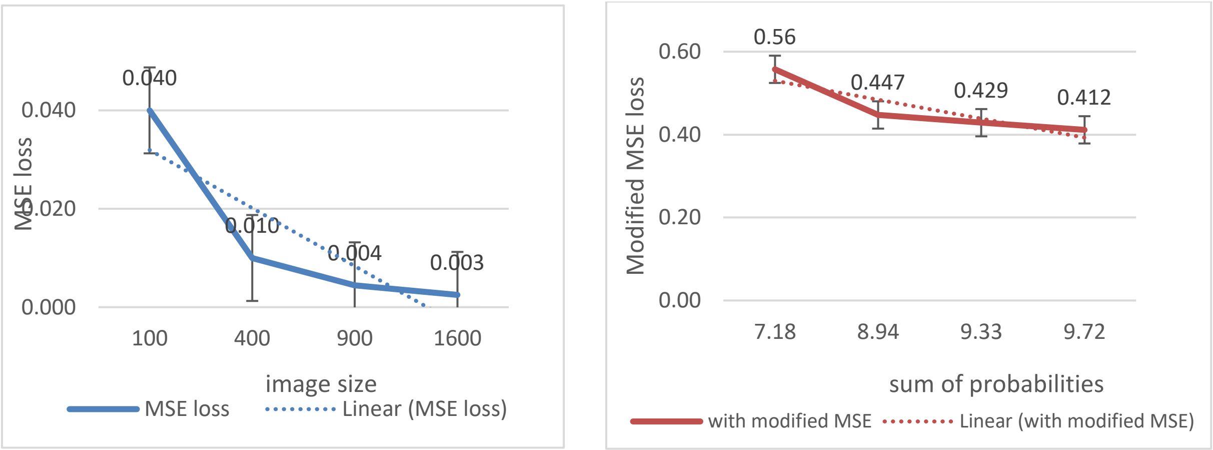
$$\text{Modified } MSE = L = \frac{1}{\sum_{i=1}^{M} \hat{y_i}} \sum_{i=1}^{M} (y_i - \hat{y}_i)^2 \qquad 3$$

$$\frac{\partial L}{\partial y_i} = \frac{\partial}{\partial y_i}\left(\frac{1}{n\hat{y}_i} \sum_{i=1}^{M} (y_i - \hat{y}_i)^2\right) \qquad 4$$

$$\frac{\partial L}{\partial y_i} = \frac{1}{n\hat{y}_i} \sum_{i=1}^{M} \frac{\partial}{\partial y_i} (y_i - \hat{y}_i)^2 \qquad 5$$

$$\frac{\partial L}{\partial y_i} = \frac{1}{n\hat{y}_i} \sum_{i=1}^{M} 2\,(y_i - \hat{y}_i)\, \frac{\partial}{\partial y_i}((y_i - \hat{y}_i)) \qquad 6$$

$$= \frac{2}{n\hat{y}_i} \sum_{i=1}^{M} (y_i - \hat{y}_i) \qquad 7$$



a b

Figure 2. a) The effect of increasing the image size on MSE loss, b) the modified MSE with the sum of the probabilities for 10 training samples.

Algorithm 1. The proposed improved MSE loss function pseudocode

```
The improved MSE function Algorithm
INPUTS:
$y_i$ ←True image class label
$\hat{y_i}$ ← Predicted class label

$SE$ ← 0                # Initial Square error
 $sum$ ← 0
   $for\ i$ ← 1 $to\ M\ do$
       $SE$ ← $SE + (y_i - \hat{y}_i)^2$
      Sum ← $sum + \hat{y}_i$
  end
improved (MSE) = (SE) / Sum
```

The idea is to decrease the denominator by summing small values. These values represent the predicted label probabilities for the training batches. Another objective that can be obtained is to maintain the high performance of the model. Moreover, the proposed approach aims to overcome the fast approach to zero problems associated with MSE, especially with the huge size of samples.

The proposed CNN architecture

The proposed method consists of three parts. The first part is to prepare input images by resizing the width and the height to 28x28x1. The number (1) represents the image channel which is grayscale level. The second part is to divide all the datasets into three sets, the training set (70%), the validation set (20%), and the testing set (10%). Then pass the dataset array to the proposed CNN blocks to build the training model. The model consists of seven CNN layers with Rectified Leaner Unit (Relu) activation function and Batch normalization layer followed by each Relu, one Maxpooling, four dropout layers, three dense layers with 3204, 1024, and 512 nodes, and one dense layer with Softmax classifier with 28 or 10 or 66 classes depending on the dataset in use. MSE and the proposed improved MSE are used with the Softmax classifier for evaluating the model. The learning rate is set to 0.001 with the Adam optimizer. The proposed CNN blocks are presented in Fig. 3.

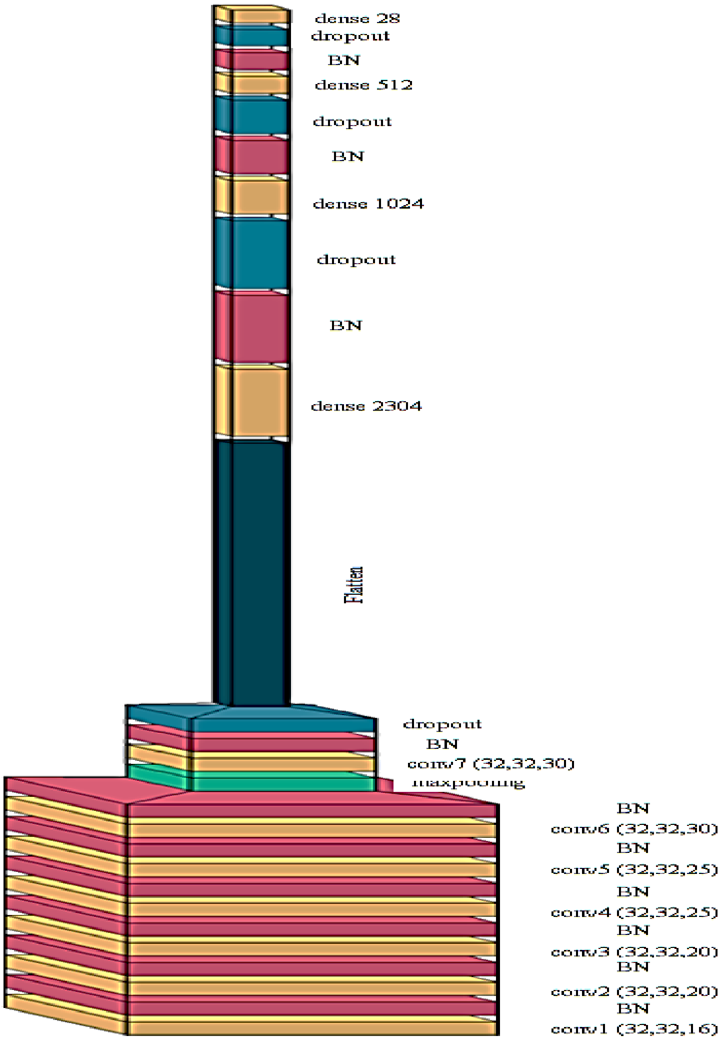


Figure 3. The proposed CNN Pipeline

Fig. 4 presents the proposed methodology flowchart. The first step is reading the images and resizing them to 30×30 then dividing the dataset into three sets the training set, validation set, and the testing set. Next, the training set is passed to the proposed CNN model or VGG16 model. The two dimensions extracted features then flattened to one dimension and passed to the dense layers. (2304, 1024, and 512 units for the proposed model and 1024, and 512 for VGG16). The trained samples are then passed to the classification layer represented by Sofmax classifier with a number of units equal to the number of the dataset class labels. The model is compiled using the enhanced MSE and The standard MSE. The accuracy and the loss rates are calculated for both the training and the validation sets for each epoch. Next is calculating the test accuracy and loss rates in addition to calculating Precision, Recall, and F1 score measures.

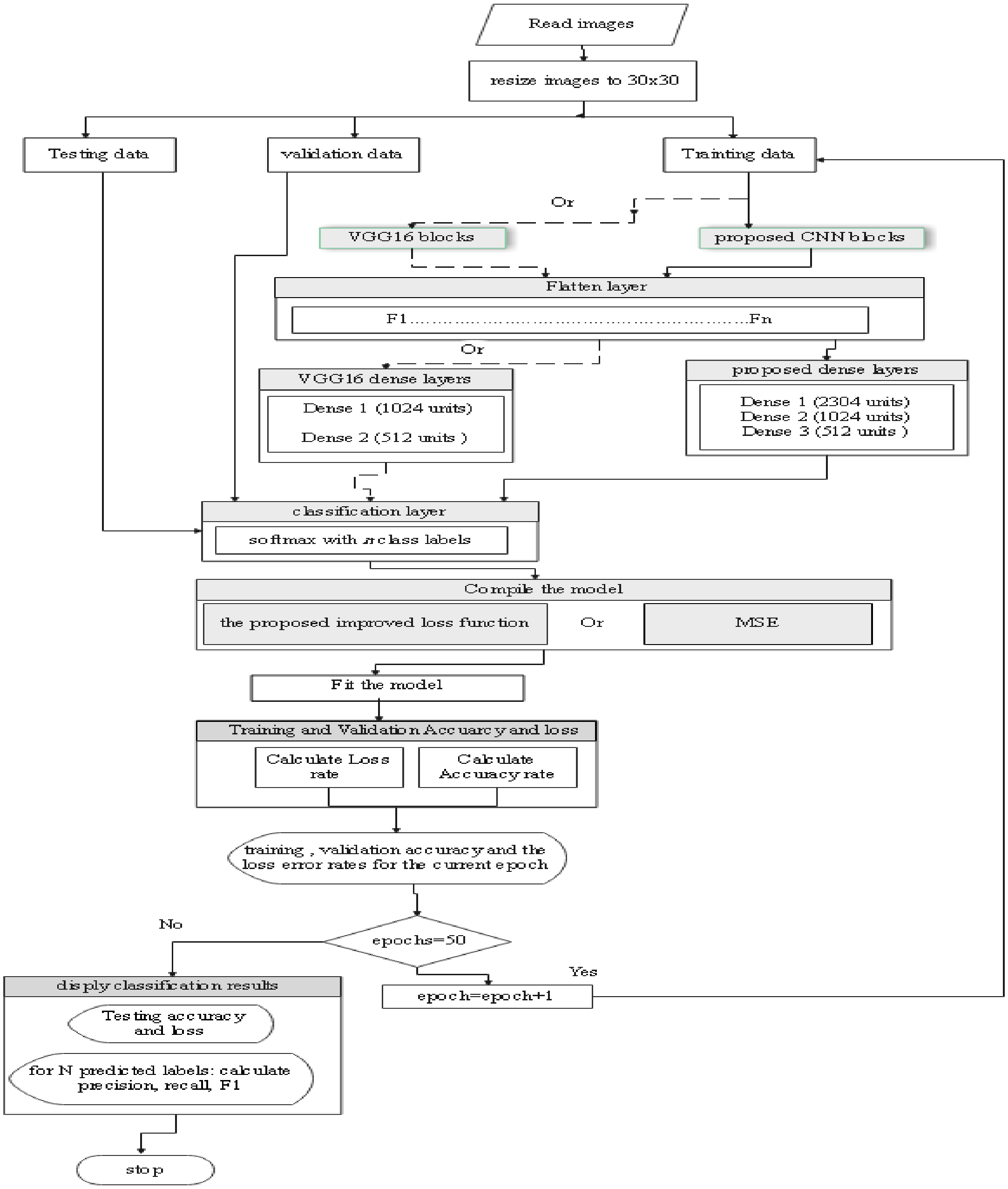


Figure 4. The proposed model flowchart

Published Online First:
https://doi.org/000000000
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Evaluation metrics

More testing approaches are used to show the improvement provided by the proposed improved MSE like the F1-score Eq. 8 and the accuracy rate.

Where:

F1 score = a combination of recall precision

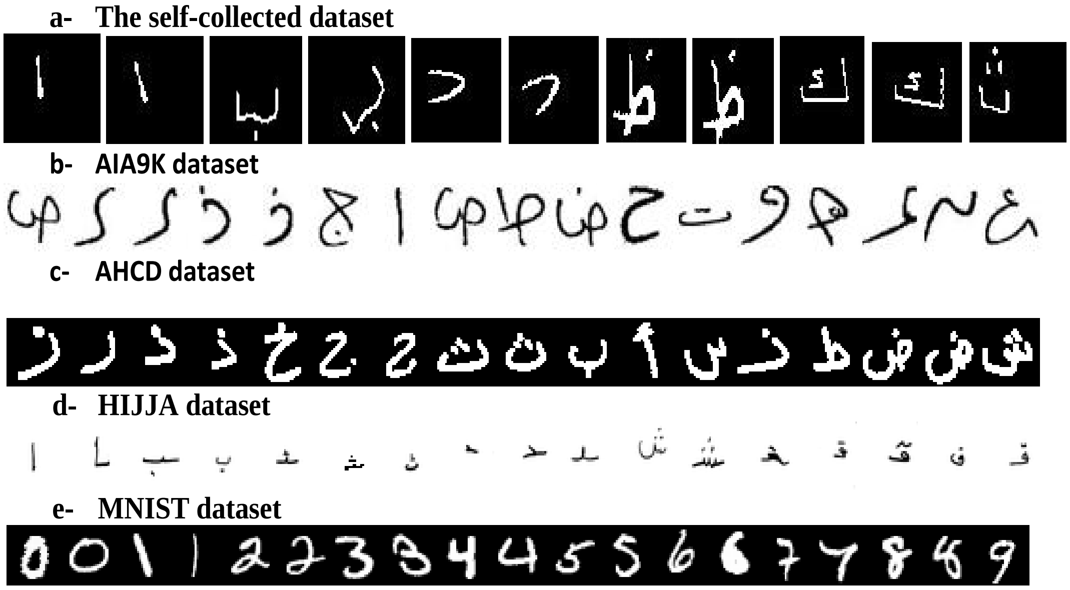
$$\boldsymbol{F1} = \frac{2*precision*recall}{precision+recall} \tag{8}$$

Experiments and results

The datasets and the hardware requirements

The experiments are implemented by using the AHAD (16.8K) [30] with 28 classes, AlexU Isolated Alphabet (AIA9K) [31] with 28 classes, Digits MNIST (70K) with 10 classes, HIJJA (48K) with 29 classes [32], and the self-collected (38K) datasets with 28 classes. Fig. 5 shows samples from the described datasets. All datasets are available for the researcher upon personal request.

The experiments are performed using laptop I5 gen.8, RAM 16 G, GPU of 4G memory, and 1T Hard disk. The proposed model is implemented using Python 3.8 and TensorFlow library. The datasets are divided into three parts training set (0.7), validation set (0.2), and testing set (0.1). These specifications are necessary because of using large-size datasets of images with deep learning.



a- The self-collected dataset

b- AIA9K dataset

c- AHCD dataset

d- HIJJA dataset

e- MNIST dataset

Figure 5. (a,b,c,d,e) Samples from the used datasets

Experiment 1: Evaluating the proposed CNN model and VGG16 with MSE

In this experiment, MSE was applied to the proposed model and VGG16, as shown in Table 1 and Figs. 6 and 7. The parameters of VGG16 model were modified to cope with one-channel images. All datasets are divided into three sets, 70% for the training set, 20% for the validation set, and 10% for the testing set. The results in Table 1 show the performance of each dataset in terms of accuracy and loss rates. As clear from the table, the MSE suffers from a fast approach to zero for all datasets, such as in HIJJA data with VGG16 and the proposed model, the loss rate is 0.00843 and 0.009, which are almost zero, while the accuracy rates are just 85% and 83.4%. On the other hand, the Cross-Entropy loss function was incompatible with accuracy rates.

Table 1. Average test data accuracy and loss rates with VGG16 and the proposed model using MSE

Model	**VGG16 with MSE**				**Propose model with MSE**				**El-Sawy & Cross-Entropy**			
Dataset	**avg test acc**	**Avg test Loss**	**Min acc**	**Max acc**	**avg test acc**	**Avg test Loss**	**Min acc**	**Max acc**	**Avg test acc**	**Avg test loss**	**Min acc**	**Max acc**
AHCD	0.955±0.046	0.003	0.95	0.961	0.964±0.004	0.002	0.958	0.972	0.92±0.01	0.32	0.89	0.915
HIJJA	0.85 ±0.006	0.008	0.84	0.861	0.835±0.009	0.009	0.820	0.843	0.79±0.03	0.9	0.72	0.815
MNIST	0.99±0.0008	0.0006	0.99	0.992	0.992±0.002	0.001	0.988	0.994	0.99±0.00	0.07	0.99	0.994
Self-collect	0.947±0.004	0.003	0.937	0.953	0.924 ±0.01	0.005	0.907	0.939	0.875±0.01	1	0.87	0.926
AIA9K	0.969±0.005	0.002	0.961	0.98	0.978±0.004	0.001	0.972	0.985	0.894±0.01	0.46	0.87	0.907

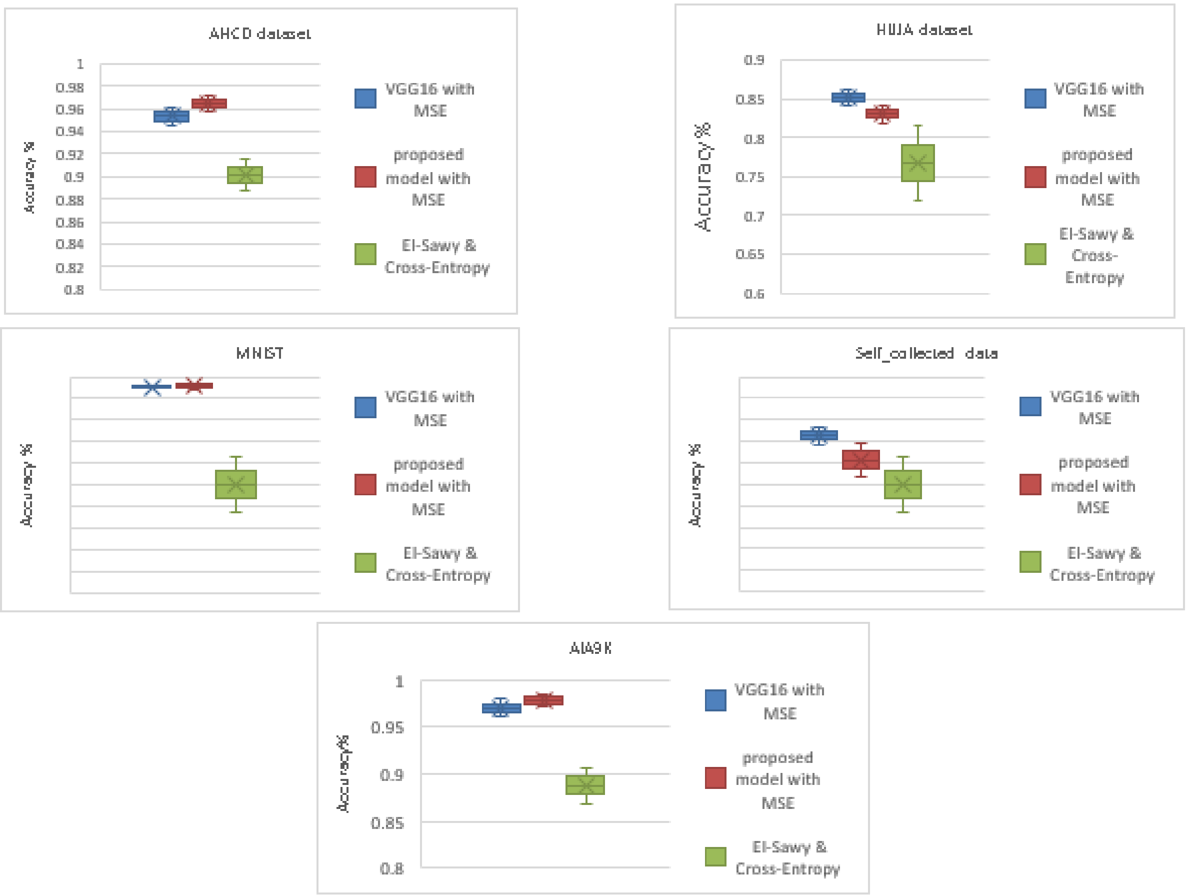


Figure 6. VGG16 model and the proposed model accuracy rate using MSE and CE loss functions.

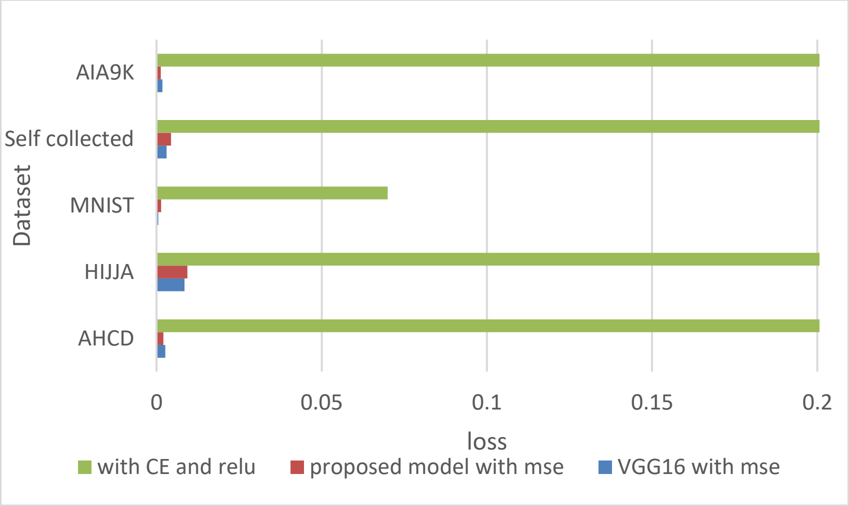


Figure 7. The loss rate for VGG16 and the proposed model using MSE and CE

Experiment 2: Evaluating the proposed CNN model and VGG16 with the improved MSE loss function

For AHCD, the accuracy and the loss rate with VGG16 and the proposed model are 96%, 0.07, 98%, and 0.04 in sequence. While with CE, the performance was 92% and 0.32, respectively. As shown in Table 2, both accuracy and error rates were improved. The models with the HIJJA dataset showed 85.7%, 0.228, 84.3%, and 0.26 accuracy rates and error rates in sequence.

The models showed better performance with the improved MSE than with the CE loss function. Whereas the HIJJA dataset showed 79%, 0.9 accuracy rate, and loss rate, respectively. With MNIST, the models showed very close performance in terms of accuracy and loss rates.

For self-collected data, the models showed significant improvement with the proposed loss function than using CE, as clear from the results in Table 2 and Figs. 9 and 10. The AIA9K dataset also showed promising results with the proposed loss function for both models compared to the result with a CE loss function.

Table 2. VGG16 and the proposed model performance with the Improved MSE loss function and CE

Dataset	**VGG16 model**				**proposed model**					**El-Sawy & CE**		
	avg test acc	**loss**	**Min**	**max**	**avg test acc**	**loss**	**min**	**max**	**Avg test acc**	**Avg test loss**	**Min acc**	**Max acc**
AHCD	0.96±0.0033	0.07	0.95	0.97	0.9759±0.002	0.040	0.972	0.98	0.92±0.01	0.32	0.89	0.915
HIJJA	0.86±0.00425	0.228	0.849	0.864	0.8437±0.003	0.264	0.837	0.849	0.79±0.03	0.9	0.72	0.815
MNIST	0.99±0.0016	0.017	0.988	0.992	0.9922±0.002	0.014	0.987	0.994	0.99±0.00	0.07	0.99	0.99
Self-collected	0.96±0.0045	0.068	0.947	0.963	0.97±0.0009	0.06	0.964	0.967	0.875±0.01	1	0.88	0.926
AIA9K	0.97±0.0037	0.049	0.97	0.975	0.9801±0.006	0.034	0.967	0.986	0.894±0.01	0.46	0.87	0.907

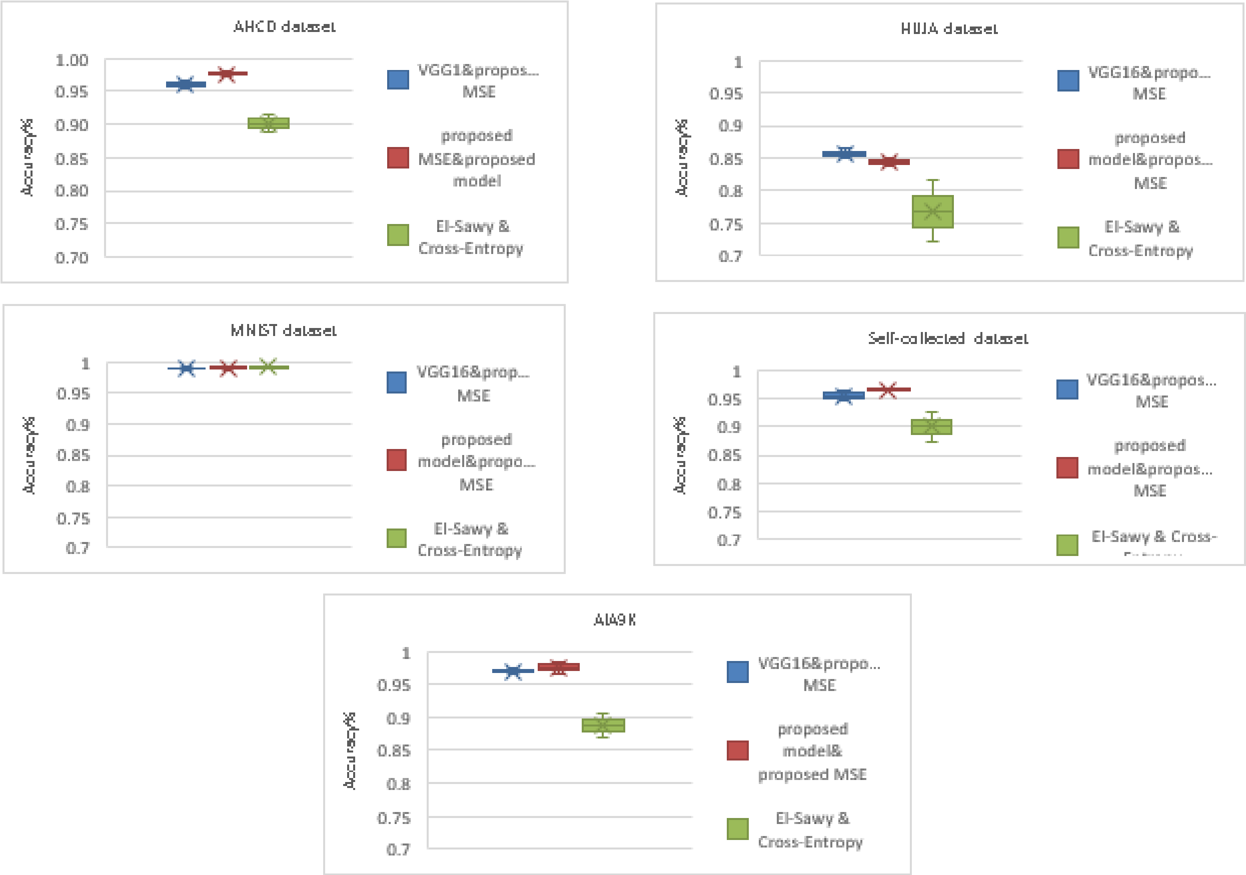


Figure 8. VGG16 and the proposed model accuracy rates with proposed loss and CE

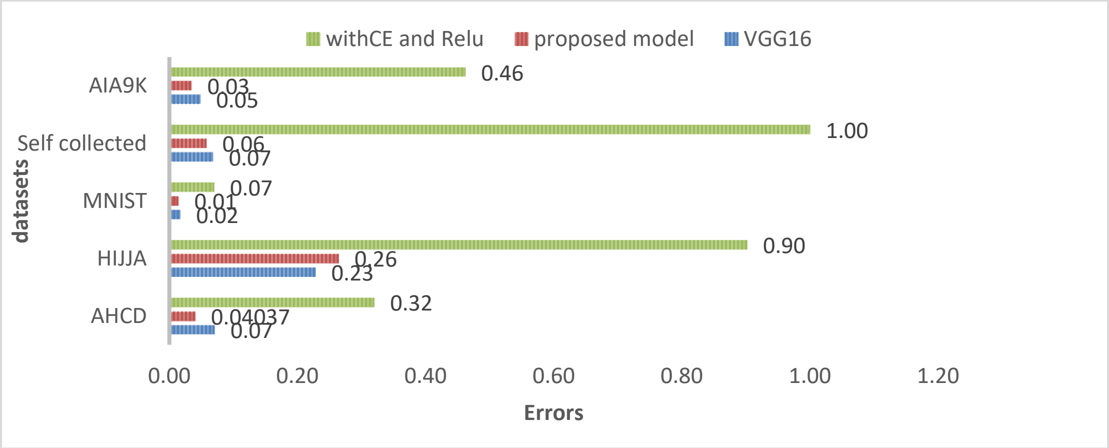


Figure 9. The loss rates for VGG16 and the proposed model with the proposed loss function and CE

Figs. 10 and 11 are presented to illustrate the results per epoch for both the improved MSE and MSE with both VGG16 and the proposed model. the experiments were run for 50 epochs, but only presented 33 epochs since the accuracy and loss rates become unchanged or slightly different. The results showed a significant improvement with the fast approach problem of loss rate gradient. The loss rate showed converge values to the accuracy rate.

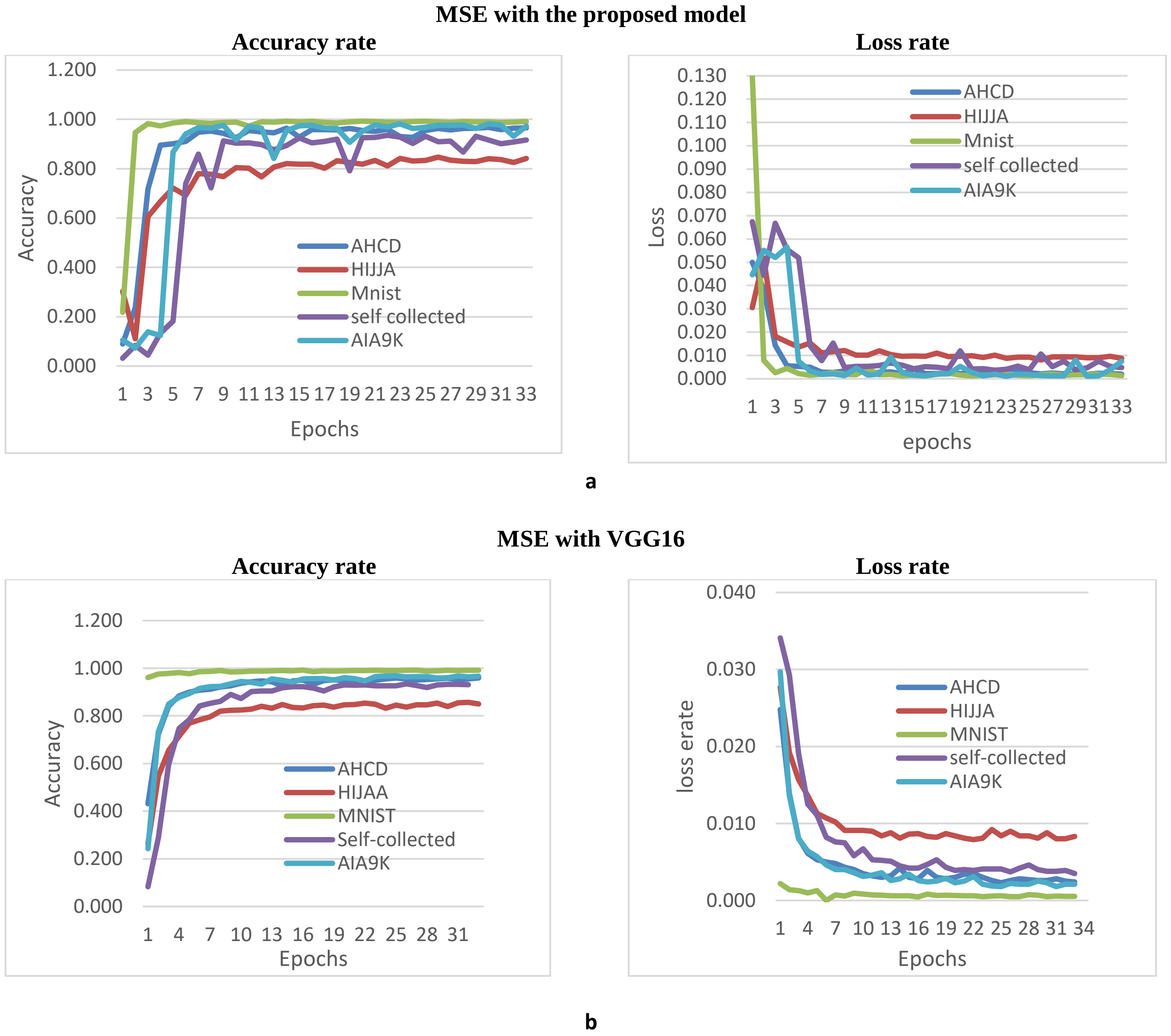
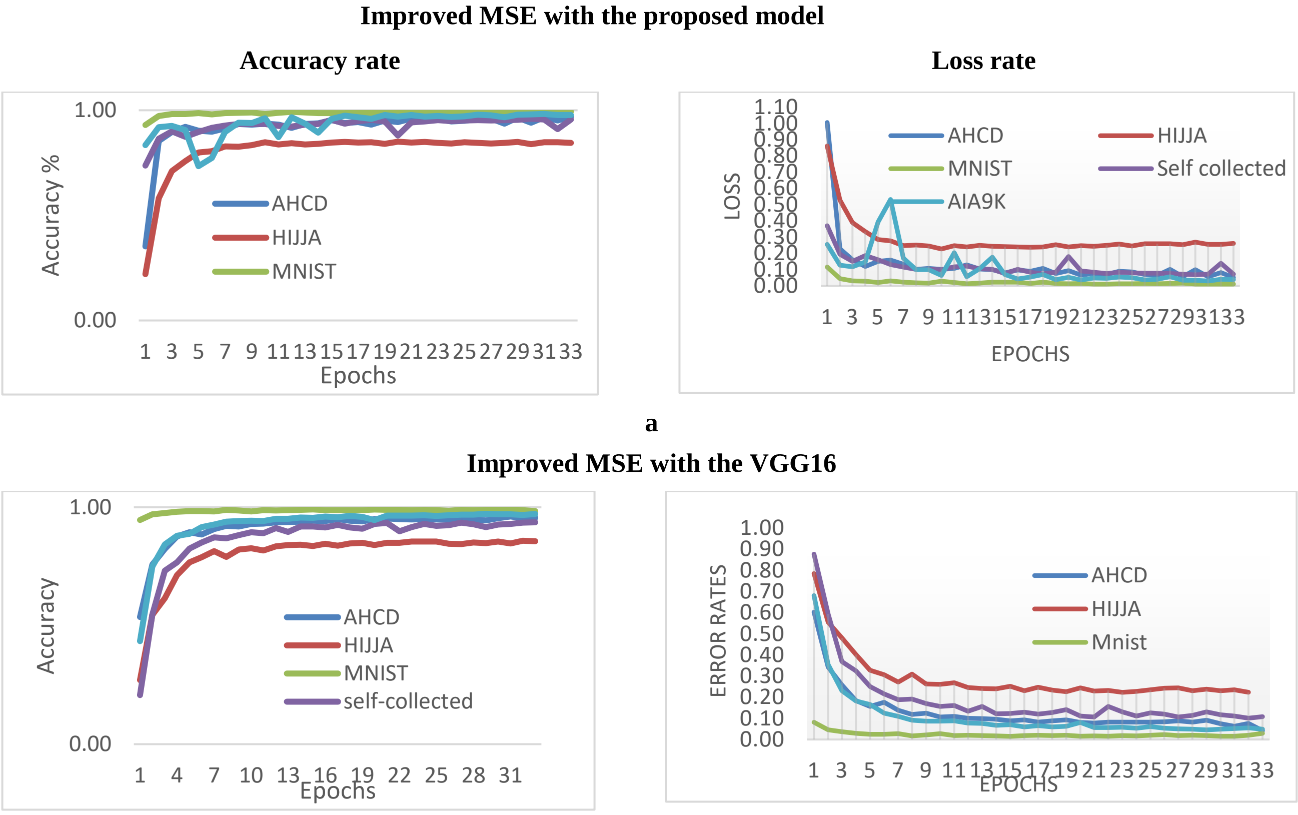


Figure 10. (a,b) The performance of MSE with VGG16 and the proposed model

From Fig. 11 (a), the result of MSE with the proposed model at epoch 3 with AHCD is 71.8% and 0.014 for Accuracy and loss rates. For HIJJA dataset the accuracy and loss rate for epoch 7 is 77.9% and 0.011. At epoch 2 the MNIST dataset showed 4.7% and 0.008 for accuracy and loss rates in turn. The self-collected data accuracy and loss rates are 72.3% and 0.015. The last dataset is the AIA9K which showed 94% and 0.003 at epoch 6.

The datasets' results with MSE and VGG16 model are presented in Fig. 11(b). As clear, all the loss rates are fast approaching and reached a small value close to zero from the early epochs. On the other hand, the accuracy rates are not compatible with the loss rates. For example, in AHCD at epoch 3 the accuracy rate is 84% and the loss rate is 0.008 which is very close to zero. Also, for HIJJA dataset the accuracy rate at epoch 8 is 82% while the error rate is 0.009 also it is not compatible with the performance accuracy rate. For MNIST at epoch 1, the accuracy is 96% and the error rate is 0.002 which is very small and close to zero. The same observation is found with AIA9K and self-collected data. This issue is due to the functionality of the MSE. It calculates the mean loss rates of all samples for each epoch.

From the above-presented result, MSE showed a fast approach to zero with the proposed model. The accuracy rates are not compatible with loss rates, and this is due to the functionality of the MSE algorithm.


Improved MSE with the proposed model

Accuracy rate

Loss rate

a

Improved MSE with the VGG16

b

Figure 11. (a,b) accuracy and loss rates of the improved MSE with VGG16 and the proposed model

Similarly, The results of the first 33 epochs of the best run with all datasets using VGG16 model and the improved MSE are presented in Fig. 11(a,b). The results of both accuracy rates and error rates have shown a clear improvement for all datasets.

Fig. 11(a) shows that the progress of both the accuracy rate and the loss rate was not smooth with the proposed model, but they still converged. When the accuracy rate decreases the error rate is increased. Such as in epochs 8 and 20 for the self-collected dataset and epoch 13 for AIA9k dataset.

For AHCD the accuracy rate at epoch 3 is 82%, and the error rate shows a compatible value which is 0.258. for HIJJA the accuracy and error rates are 79% and 0.3. Also, the MNIST showed a 94.7% accuracy rate and 0.08 error rate at epoch 1. While self-collected rates for epoch 7 are 87% and 0.188. In regards to AIA9k the rates of epoch 2 are 75% and 0.35. All datasets showed smooth progress for the accuracy performance. The loss rate for VGG16 model was better than with the proposed model.

F1 measure Factor

In addition to the accuracy and loss rate, the F1 measure factor is calculated to compare the proposed model and loss function with VGG16 and MSE. The results are presented in Table 3. It is clear from the results in bold that the F1 measure value with the proposed model and the improved MSE is better than with VGG16, MSE, and Cross entropy.

Table 3. F1 factor for all datasets with the proposed model, VGG16, MSE, and improved MSE

model	with MSE & VGG16	with improved MSE& VGG16	with MSE &proposed model	with an improved MSE&proposed model	with CE&relu
dataset	**F1**	**F1**	**F1**	**F1**	**F1**
AHCD	0.96	0.96	0.96	**0.98**	0.92
HIJJA	0.86	0.81	0.84	**0.85**	0.79
MNIST	0.99	0.99	0.99	**0.99**	0.99
AIA9K	0.98	0.96	0.97	**0.98**	0.89
self-collected	0.96	0.94	0.96	**0.97**	0.87

State of Art comparison

MSE and Cross entropy are loss functions used for Arabic handwritten characters classification by [14], [30], [33-36] using Deep learning techniques in various CNN architectures. Their main objective was to reduce the loss rate and increase the accuracy rate.

To compare with [33], the authors reimplementated the proposed model(VGG12) and datasets with MSE, the proposed improved MSE, and the proposed model. They used MSE as a loss function, reported only the data validation results, and ignored the testing result because the model did not hold for testing data [33]. The study used two datasets the Arabic digits database (Adbase), which contains 70000 samples, and Handwritten Arabic characters (HACB), which contains 6600 samples. Fig. 12 shows some samples from these two datasets. The description of these datasets is mentioned in the same study.

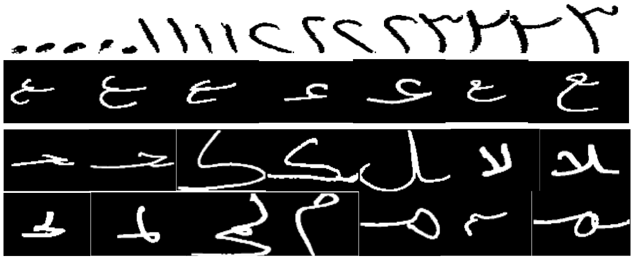


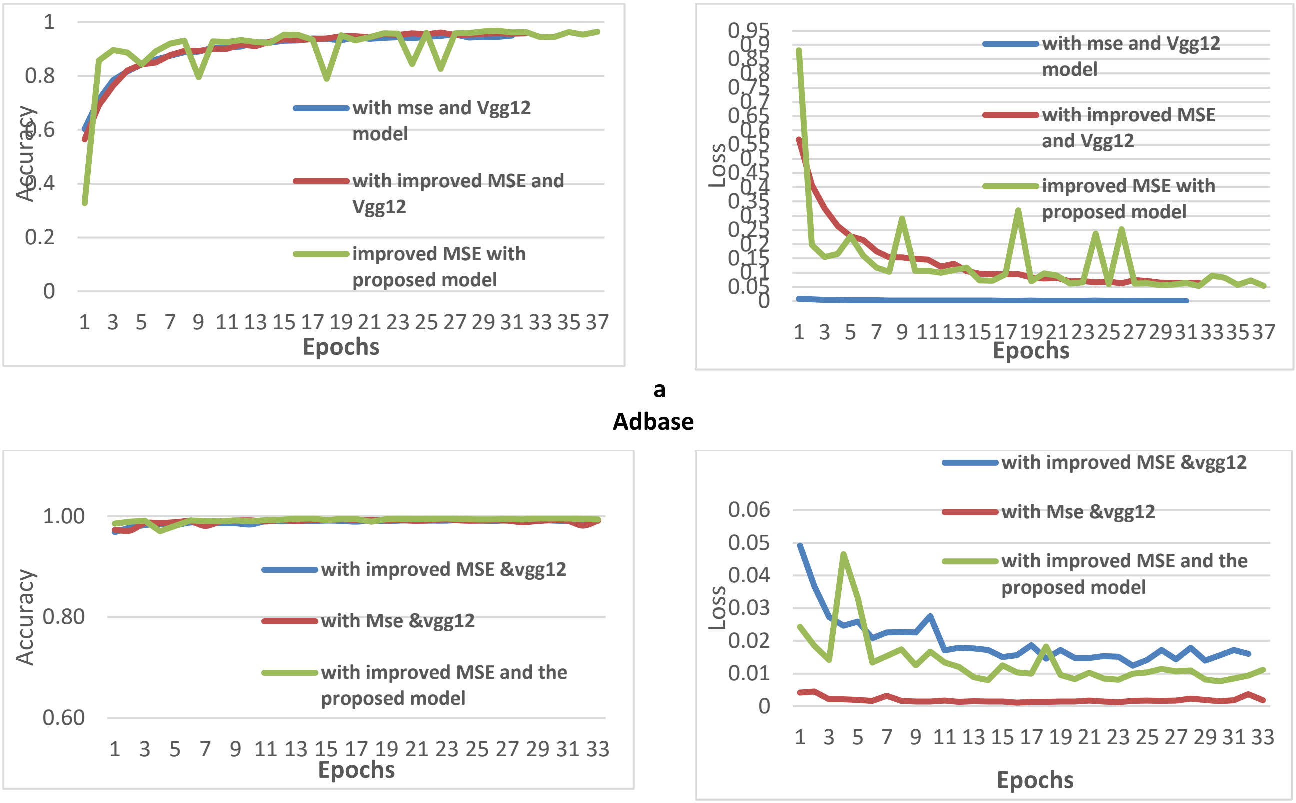
Figure 12. Samples from Adbase and AHCDB datasets

Ten runs were performed for both datasets using MSE and our proposed improved MSE and model. The average results of testing accuracy and loss rates are presented in Table 4 and Fig. 13. Also, some true and predicted labels are presented in Table 5. The results showed a slight improvement in the accuracy rates of both datasets, but the error rate gradient showed less approach speed.

Table 4. The Average, Min, Max accuracy, and loss rates of 10 runs for Adbase and AHCDB dataset with Vgg12 and the proposed model with improved MSE

Model	vgg12				with the proposed model	
Dataset	**Adbase**		**AHCDB**		**Adbase**	**AHCDB**
Method	**With improved MSE**	**with MSE**	**with improved MSE**	**with MSE**	**with improved MSE**	
avg accuracy	0.99 ±0.0011	0.994±0.001	0.969±0.00196	0.959±0.00197	0.9959±0.0004	0.9695±0.0014
avg loss	0.0113	0.0011	0.049	0.00099	0.0076	0.0485
Max	0.9953	0.9958	0.9729	0.9624	0.9964	0.9714
Min	0.991	0.993	0.9661	0.9572	0.9954	0.9672
avg F1 measure	1	0.99	0.96	0.99	1.0	0.97

AHCDB



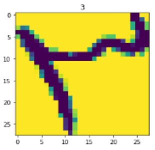
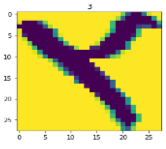
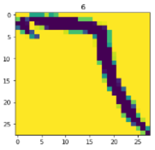
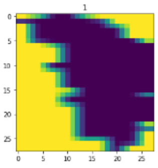
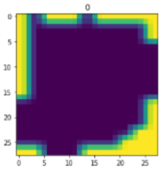
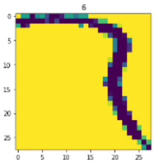
a

Adbase

b

Figure 13. (a,b) Accuracy and loss rates for Adbase and AHCDB datasets with VGG12 , the proposed model, and the improved MSE loss

Table 5. Samples from the tested samples of Adbase and HACDB (B for beginning, E for Ending, and m for middle)

Adbase dataset test samples		
True label: 3 Predicted label: 3	True label: 3 Predicted label: 3	True label: 6 Predicted label: 6
Class: 3 	Class: 3 	Class: 6 
True label: 1 Predicted label: 1	True label: 0 Predicted label: 0	True label: 6 Predicted label: 6
Class: 1 	Class: 0 	Class: 6 

HACDB dataset test samples

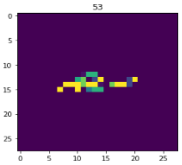
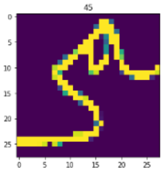
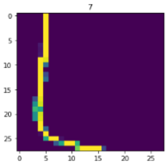
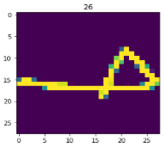
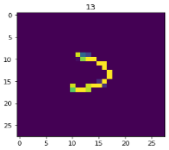
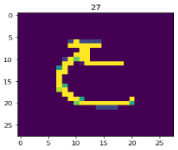
True label: Meem, M Predicted label: Meem, M	True label: Meem_Jeem Predicted label: Meem_Jeem	True label: Alef_E Predicted label: Alef_E
Class: 53 	Class: 45 	Class: 7 
True label: Jeem_B	True label: Daal	True label: Aeen_E
Predicted label: Jeem_B	Predicted label: Daal	Predicted label: Aeen_E
Class: 26 	Class: 13 	Class: 27 

Table 6 presents only the accuracy rates reported by the state of Art studies. The authors of this study concluded that the improved MSE overcomes the fast approach to zero problems of the MSE with the classification problems. The resulting accuracy rates are improved in a way that converged reasonably with error rates. this improvement overcomes all the results of the state of art studies reported in Table 6. Except for the HIJJA dataset which showed less accuracy rate because it contains samples with high similarity.

Table 6. The comparison between state-of-the-art and the proposed improved MSE results

Research	Dataset	Method	Accuracy rate
[33]	ADBase	MSE/RMSprop	99.66% best val. acc
		CCE/Adam	99.57% best val. acc
	HACDB	VGG/CE	97.32% best val. Acc
		VGG/MSE	96% best val. Acc
[34]	CMATERDB 3.3.1	RBM-CNN	98.59%
[35]	MADbase AHCD	Monte Carlo Cross-Validation (MCCV)& KCC/CE	99.52% 98.42%
[30]	AHCD	CNN+ CE	94.8%
[36]	AIA9K AHCD	CNN+ CE Without augmentation With augmentation	94.9% 94.7% 97.6%
[14]	AHCD	CNN+ CE	97%
[32]	AHCD Hijja	CNN+ CE	97% 88%
Our proposed method	**AHCD** **AIA9K**	**CNN+ Improved MSE**	**97%** **98.56%**

Self-collected	**97%**
Hijja	**85.4%**
D. MNIST	**99.45%**
Adbase	**99.64%**
AHCDB	**96.72%**

Conclusion

The Mean Square Error loss function is usually used for regression problems. While for classification problems MSE showed unsatisfactory results due to its structure.

This study proposed an improved MSE loss function to classify handwritten characters and digits. The idea is to replace the sample's mean value with a lower value represented by summing the predicted probability values of the predicted labels. The improved MSE is tested using two models 1) VGG16 and 2) the proposed CNN model. VGG16 model was modified to perform on one-channel images. Also, the study presented in detail the improved MSE using equations and algorithms.

AHCD, HIJJA, AIA9K, Self_collected, and MNIST datasets were used to evaluate the performance of the proposed improved MSE. The results also showed a high-performance accuracy rate with a converged loss rate and overcame the fast approach to zero problems associated with MSE, especially with the huge size of samples as shown in Fig. 14. The results of the Improved MSE are compared to the performance of MSE and cross-entropy (CE) loss functions from the state-of-art. The results showed outstanding performance for the Improved MSE over the MSE and CE.

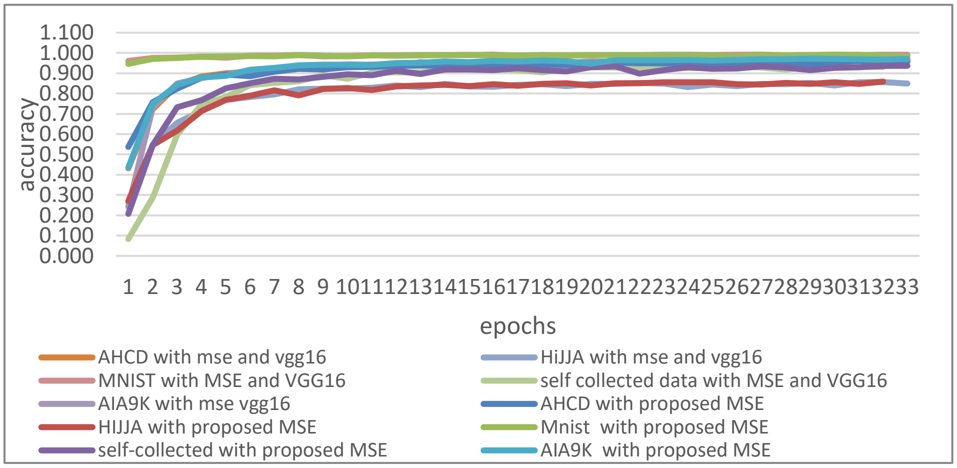


Figure 14. The accuracy rates of all datasets with improved MSE and MSE loss functions for VGG16 model

The results of the improved MSE are compared to the performance of MSE and CE loss functions from the state-of-the-art. The results showed an outstanding slower approach to zero for the proposed MSE over the MSE and more reliable loss rates than with CE as shown in Fig. 15. In addition, two more datasets were used, Adbase and AHCDB, to compare the performance of the proposed loss function with state-of-the-art studies. The main objective is to reach an accuracy rate that is compatible with the loss rate. The results of our improved MSE with our proposed model outperformed the MSE and CE loss functions and VGG16 models.

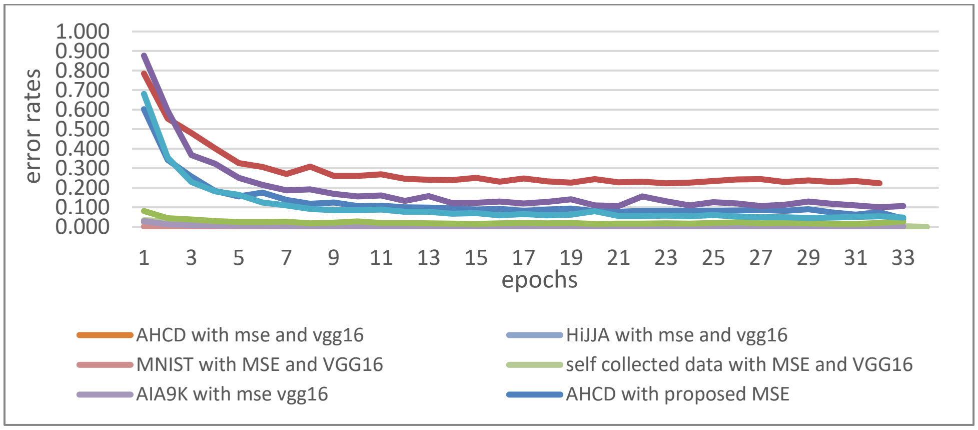


Figure 15. The loss rates of all datasets with improved MSE and MSE loss functions for VGG16 model

Acknowledgment

The author would like to thank UKM for providing the facilities and financial support under No.UKM-AP-2017-005-2 entitled “Using STEM data through Smart self-crime prevention at schools for open data readiness” and FRGS/1/2019/ICT02/UKM/02/9 entitled “Convolution neural network enhancement based on adaptive convexity and regularization functions for fake video analytics”. Not forgetting the research team. members at the Digital Forensic Lab and Medical and Health Informatics Lab at the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, and Tier Mobility GmbH from Berlin, Germany.

Author’s Declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for re-publication, which is attached to the manuscript.
- No animal studies are present in the manuscript.
- No human studies are present in the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at University of Al Nahrain.

Author’s Contribution Statement

B. H., S. N. H. Sh. A., and M. M. contributed to the design and implementation of the research, the analysis of the results, and the writing of the manuscript.

References

1. Janocha K, Czarnecki WM. On loss functions for deep neural networks in classification. arXiv preprint arXiv: 170205659. 2017 Feb 18; 25. https://doi.org/10.4467/20838476SI.16.004.6185
2. Edalatifar M, Tavakoli MB, Ghalambaz M, Setoudeh F. Using deep learning to learn physics of conduction heat transfer. J Therm Anal Calorim. 2020 July 31; 146: 1435–1452. https://doi.org/10.1007/s10973-020-09875-6
3. Pham H. A New Criterion for Model Selection. Math. 2019 Dec 10; 7(12): 1215. https://doi.org/10.3390/math7121215
4. Ren J, Zhang M, Yu C, Liu Z, editors. Balanced mse for imbalanced visual regression. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022 Mar 30. https://doi.org/10.48550/arXiv.2203.16427

5. Mazaal AR, Karam NS, Karam GS. Comparing Weibull Stress – Strength Reliability Bayesian Estimators for Singly Type II Censored Data under Different loss Functions. Baghdad Sci J. 2021; 18(2): 0306. https://doi.org/10.21123/bsj.2021.18.2.0306
6. Lawgali A, Bouridane A, Angelova M, Ghassemlooy Z. Handwritten Arabic character recognition: Which feature extraction method?. Int J Adv Sci Technol. 2011; 34: 1-8.
7. Hasasneh N, Hasasneh A, Salman N, Eleyan D. Towards offline Arabic handwritten character recognition based on unsupervised machine learning methods: A perspective study. 2019.
8. Lorigo LM, Govindaraju V. Offline Arabic handwriting recognition: a survey. IEEE Trans Pattern Anal Mach Intell. (TPAMI). 2006; 28(5): 712-24. https://doi.org/10.1109/TPAMI.2006.102
9. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural Comput (NECO). 2006; 18(7): 1527-54. https://doi.org/10.1162/neco.2006.18.7.1527
10. Goyal P, Pandey S, Jain K. Deep learning for natural language processing. SpringerLink. 2018; 138-43. https://doi.org/10.1007/978-1-4842-3685-7
11. Chen Z. Deep-learning Approaches to Object Recognition from 3D Data. Case Western Reserve University. 2017.
12. Rosasco L, Vito ED, Caponnetto A, Piana M, Verri A. Are loss functions all the same? Neural Comput (NECO).2004; 16(5): 1063-76. https://doi.org/10.1162/089976604773135104
13. Deng L, Gong Y, Lu X, Lin Y, Ma Z, Xie M. STELA: A Real-Time Scene Text Detector With Learned Anchor. IEEE Access. 2019;7:153400-7. https://doi.org/10.1109/ACCESS.2019.2948405
14. Najadat HM, Alshboul AA, Alabed AF, editors. Arabic handwritten characters recognition using convolutional neural network. 2019 10th International Conference on Information and Communication Systems (ICICS). 2019. https://doi.org/10.1109/IACS.2019.8809122
15. El Atillah M, El Fazazy K, Riffi J. Classification of Arabic Alphabets Using a Combination of a Convolutional Neural Network and the Morphological Gradient Method. Baghdad Sci J. 2024; 21(1): 0252. https://doi.org/10.21123/bsj.2023.7877
16. Saifullah, Ren Z, Hussain K, Faheem M. K-means online-learning routing protocol (K-MORP) for unmanned aerial vehicles (UAV) adhoc networks. Ad Hoc Networks. 2024; 154: 103354. https://doi.org/10.1016/j.adhoc.2023.103354
17. Saini D, Garg R, Malik R, Prashar D, Faheem M. HFRAS: design of a high-density feature representation model for effective augmentation of satellite images. Signal, Image and Video Processing. 2023 Nov 11. https://doi.org/10.1007/s11760-023-02859-7
18. Zhang Z, Sabuncu M, editors. Generalized cross entropy loss for training deep neural networks with noisy labels. Adv Neural Inf Process Syst. 2018; May 20: 8792–8802. https://doi.org/10.48550/arXiv.1805.07836
19. Zhu X, Zhou H, Yang C, Shi J, Lin D, editors. Penalizing top performers: Conservative loss for semantic segmentation adaptation. Proceedings of the European Conference on Computer Vision (ECCV); Springer, Cham; 2018; 11211: 587–603. https://doi.org/10.1007/978-3-030-01234-2_35
20. Tang Y. Deep learning using linear support vector machines. arXiv preprint arXiv: 13060239. 2013 Jun 2. https://doi.org/10.48550/arXiv.1306.0239
21. Mohapatra P, Rolinek M, Jawahar C, Kolmogorov V, Pawan Kumar M, editors. Efficient optimization for rank-based loss functions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).2018 Jun 28; 3693-3701. https://doi.org/10.1109/cvpr.2018.00389
22. Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res. 2005; 30(1): 79-82. https://doi.org/10.3354/cr030079
23. Willmott CJ, Matsuura K, Robeson SM. Ambiguities inherent in sums-of-squares-based error statistics. Atmos. Environ.2009 Jan; 43(3): 749-52. https://doi.org/10.1016/j.atmosenv.2008.10.005
24. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geosci Model Dev. 2014 Jun 30; 7(3): 1247-50. https://doi.org/10.5194/gmd-7-1247-2014
25. Chatterjee S, Hammad A, Katzin EN, Hua J. Virtual wallet card selection apparatuses.methods and systems. 2013 Nov 5.
26. Ghazvini A, Abdullah SNHS, Hasan MK, Kasim DZAB. Crime spatiotemporal prediction with fused objective function in time delay neural network. IEEE Access. 2020 Jun 18; 8: 115167-83. https://doi.org/10.1109/ACCESS.2020.3002766
27. Nayef BH, Abdullah SNHS, Sulaiman R, Alyasseri ZAA. Optimized leaky ReLU for handwritten Arabic character recognition using convolution neural networks. Multimed Tools Appl. 2022 Oct 19; 81(2): 2065-94. https://doi.org/10.1007/s11042-021-11593-6
28. Pandit V, Schuller B. On Many-to-Many Mapping Between Concordance Correlation Coefficient and Mean Square Error. arXiv preprint arXiv: 190205180. 2019 Feb 14. https://doi.org/10.48550/arXiv.1902.05180

29. Sharma N. A Beginner's Guide to Loss functions for Regression Algorithms November 14, 2021 .
30. El-Sawy A, Loey M, El-Bakry H. Arabic handwritten characters recognition using convolutional neural network. WSEAS Trans Comput Res. 2017; 5: 11-9.
31. Torki M, Hussein ME, Elsallamy A, Fayyaz M, Yaser S. Window-based descriptors for arabic handwritten alphabet recognition: A comparative study on a novel dataset. arXiv preprint arXiv: 14113519. 2014 Nov. https://doi.org/10.48550/arXiv.1411.3519
32. Altwaijry N, Al-Turaiki I. Arabic handwriting recognition system using convolutional neural network. Neural Comput Appl. 2020 Jun 28 ; 1-13. https://doi.org/10.1007/s00521-020-05070-8
33. Mudhsh M, Almodfer R. Arabic handwritten alphanumeric character recognition using very deep neural network. info. 2017 Aug 31; 8(3): 105. https://doi.org/10.3390/info8030105
34. Alani AA. Arabic handwritten digit recognition based on restricted Boltzmann machine and convolutional neural networks. info. 2017 Nov 9 ; 8(4): 142. https://doi.org/10.3390/info8040142
35. de Sousa IP. Convolutional ensembles for Arabic handwritten character and digit recognition. PeerJ Comput Sci. 2018 Oct 15 ; 4: e167. https://doi.org/10.7717/peerj-cs.167.
36. Younis KS. Arabic handwritten character recognition based on deep convolutional neural networks. Jordanian J Comput Inf Techno (JJCIT). 2017 Feb 18; 3(3): 186-200. https://doi.org/10.48550/arXiv.1702.05659

أثارالتعلم العميق لتمييزالأحرف المكتوبة بخط اليد باستخدام دالة متوسط مربع الخطأ المحسّنة

باهرة هاني نايف [1]، ستي نورالهدى الشيخ عبدالله [2]، منال محمد [2،3]

[1]قسم الحاسوب،كلية العلوم، جامعة النهرين، بغداد، العراق.
[2]كلية علم وتكنولوجيا المعلومات ، الجامعة الوطنية الماليزية، بانجي-سلانجور، ماليزيا.
[3]كلية العلوم الادارية،لجامعة حضرموت، المكلا، اليمن.

الخلاصة

تستخدم دوال الخسارة لتقييم دقة أداء نموذج التصنيف في تعلم الماكنة. قيمة معدل الخسارة يقيس اذا ما كان قيم التسميات المتوقعة مطابقة الى التسميات الحقيقة او قريبة منها. تعد تمييزالحروف المكتوبة بخط اليد من مشاكل تمييز الانماط. استخدام التعلم العميق سرع التطور الحاصل في تطبيقات تمييز الانماط لتصنيف الصورالمخزونة.تعد الشبكات العصبية التلافيفية ودالة التفعيل الاكثر استخداما في التعلم العميق. اضافة الى استخدام دوال خسارة مختلفة لتقييم أداء النموذج مثل Categorical Cross-Entropy (CCE) و Mean Square Error (MSE). في تقنيات التعلم العميق, حجم قاعدة البيانات يلعب دور مهم للحصول على أداء عالي, ولكن مع MSE التقليدية قيم الخسارة تصل الى الصفر في مراحل مبكرة من عملية التدريب مع قاعدة البيانات الكبيرة الحجم, بالرغم من كون دقة النموذج ما زالت تتحسن. هذه الدراسة تقترح دالة خسارة مطورة عن طريق تحسين MSE. تستند MSE المقترحة على تقسيم مربع الخطأ على مجموع احتمالات التسمية المتوقعة بدل من مجموع عدد العينات. تم استخدام خمسة قواعد بيانات لاختبار اداء النموذج المقترح الشبكات العصبية التلافيفية مع MSE المطوره بالاضافة الى استخدام نموذج VGG16 المعدل. وقواعد البيانات هم AHCD, AIA9K, HIJJA, Self-collected وMNIST . اظهرت نتائج الدراسة تطور واضح في دقة اداء النموذج المقترح مع متوسط مربع الخطأ MSE المطورة مقارنة ب CCE.

الكلمات المفتاحية: تعلم الماكنة، تمييز الأنماط، التعلم العميق، الشبكات العصبية التلافيفية، ودالة التفعيل، متوسط مربع الخطأ.