International Journal of Electrical and Computer Engineering (IJECE)
Vol. 12, No. 2, April 2022, pp. 1904~1909
ISSN: 2088-8708, DOI: 10.11591/ijece.v12i2.pp1904-1909

Expansion dataset COVID-19 chest X-ray using data augmentation and histogram equalization

Farah Flayyeh Alkhalid[1], Abdulhakeem Qusay Albayati[2], Ahmed Ali Alhammad[3]
[1]Control and Systems Engineering Department, University of Technology-Iraq, Baghdad, Iraq
[2]Computer Engineering Department, University of Technology-Iraq, Baghdad, Iraq
[3]Computer Engineering Department, College of Engineering, Al-Nahrain University, Baghdad, Iraq

Article Info

Article history:

Received Jul 30, 2020
Revised Sep 8, 2021
Accepted Oct 2, 2021

Keywords:

CLAHE
Convolutional neural network
COVID-19
Data augmentation
Histogram equalization

ABSTRACT

The main important factor that plays vital role in success the deep learning is the deep training by many and many images, if neural networks are getting bigger and bigger but the training datasets are not, then it sounds like going to hit an accuracy wall. Briefly, this paper investigates the current state of the art of approaches used for a data augmentation for expansion the corona virus disease 2019 (COVID-19) chest X-ray images using different data augmentation methods (transformation and enhancement) the dataset expansion helps to rise numbers of images from 138 to 5520, the increasing rate is 3,900%, this proposed model can be used to expand any type of image dataset, in addition, the dataset have used with convolutional neural network (CNN) model to make classification if detected infection with COVID-19 in X-ray, the results have gotten high training accuracy=99%.

Corresponding Author:

Farah Flayyeh Alkhalid
Control and Systems Engineering Department, University of Technology-Iraq
Baghdad, Iraq
Email: farah.f.alkhalid@uotechnology.edu.iq

1. INTRODUCTION

With the development of technology, artificial intelligence and machine learning have become interested in most fields and applications including medical, so the computer was used in medical area like skin cancer [1] breast cancer [2], thyroid nodules [3] and diabetic retinopathy [4] are some of these studies. Deep learning approaches and neural networks run well when there is a big amount of images in dataset. In the case of image recognition, the recognition model always benefits from a large amount of image data [5]. Data augmentation procedures used for expanding data set, such as translation in [6], [7] while histogram equalization is used to enhance the x-ray image [8]. Mikołajczyk and Grochowski [9] studied and compared numerous approaches of data augmentation for image classification in order to progress the training process proficiency. Engistrom *et al.* [10] used adversarial rotations and translations to improve robustness of dataset. Zheng *et al.* [11] used a filled phase data augmentation to expand the precision of convolutional neural networks (CNN), it can be used to understood model collective without presenting extra training costs. Concurrent data augmentation throughout the stages of training and testing, can guarantee network enhance and optimization its full ability.

On the other hand, there are many different training models impedes its real goal when using neural augmentation [12]. Other research work trained their models on a huge dataset such as [13]–[15]. This study focuses on how can make expansion corona virus disease 2019 (COVID-19) dataset from some tens of pictures to some thousands using image enhancement and data augmentation. Which can be used later in deep learning model to make deep training and then high accuracy recognition.

2. PROPOSED METHOD

There are multi layers of preprocessing and enhancement are used in this system, as highlighted in Figure 1, the processes are listed in steps in Figure 1. Collect raw dataset for healthy and COVID-19 state with limited number (70 and 68 respectively), these images are X-ray images so by using 5. Histogram equalization (HE) the contrast of images are enhanced and the intensities are distributed equally, another layer of contrast limited adaptive histogram equalization (CLAHE) is applied to generate double dataset and provide different levels of contrast, after that make augmentation to expand the dataset, finally remove the duplicate images Table 1 demonstrates the statistics of each step. Each step is defined in detail in the next.

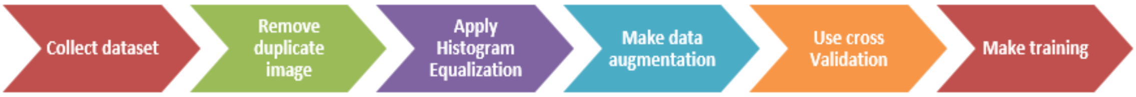


Figure 1. System sequence

Table1. Steps statistics

Step #	Operation	Number of images		
		Covid-19	Healthy	Total
Step 1	Collect data	68	70	138
Step 2	HE	136	140	276
	CLAHE	272	280	552
Step 3	Augmentation	2720	2800	5520
Step 4	Hashing	2720	2800	5520

3. COLLECT DATASET

The 2019 novel COVID-19 grants numerous unique features [9], [16]. Although the analysis is established consuming polymerase chain reaction (PCR), the diseased patients with pneumonia might appear on chest X-ray and computed tomography (CT) images with a design that is only moderately characteristic for the doctor glimpse, so the dataset can help significantly in systems based deep learning to make diagnostic of infection with (COVID-19) or not by computer. Collecting dataset is the first step of the proposed method in this paper, the raw dataset which is used in this paper is downloaded from [10], [11], [17], [18].

4. IMAGE HASHING

In any training model, we need to ensure the used dataset does not contain repeated images, to ensure variety in learn when make forward and backward propagation, so, by using image hashing [19] which each image is compact to a small hash one or fingerprint size by recognizing noticeable features in the main image and hashing a compressed of those features (instead of hashing the image data directly). Then, count the number of different bits in the same position using hamming distance (HD). Classically, use the HD to compare hashes. The HD measures the number of bits in two hashes that are different. If two hashes with a HD of zero indicates that the two hashes are matching (no bits are different), that the two images are perceptually similar also. Perez and Wang [12] suggests that hashes with differences more than 10 bits are most likely different while HD between 1 and 10 are possibly a simple difference for the equivalent image [13].

5. HISTOGRAM EQUALIZATION (HE)

The contrast is referred as the contouring in brightness of color or difference in color and intensity between any adjacent objects in an image [20]. If the contrast is too low, it is impossible to distinguish between two objects and they are seen as a single object, HE is a popular approach for regulating image intensities to improve contrast. The HE method is built on a distribution an intensity along the image; the aim is to get a high contrast for the new image. Assume the r is a random factor which shows the level of gray of an image. First of all, it can be assumed that r is series and involves in the locked interval [0:1] where when r is 0 then meaning black and when r is 1 meaning white:

$$s=T(r)$$

First thing, it must be transforming the value from black to white with respect to the gray level, and the other one ensures that the range is including within the allowable values of gray pixels range. The transformation of inverse levels factor s to r could be denoted by:

r=T–1(s)

Consider the original and inversed gray contrast be categorized by the probability of density functions pr(r) and ps(s) separately. So, as of basic probability function, if pr(r) and ps(s) are acknowledged and also T-1(s) then the probability function of density for the transformed gray level is shown by (1):

$$ps(s) = \left[pr(r)\frac{dr}{ds}\right]^{T^{-1}(s)} \qquad (1)$$

The transform function is denoted by (2):

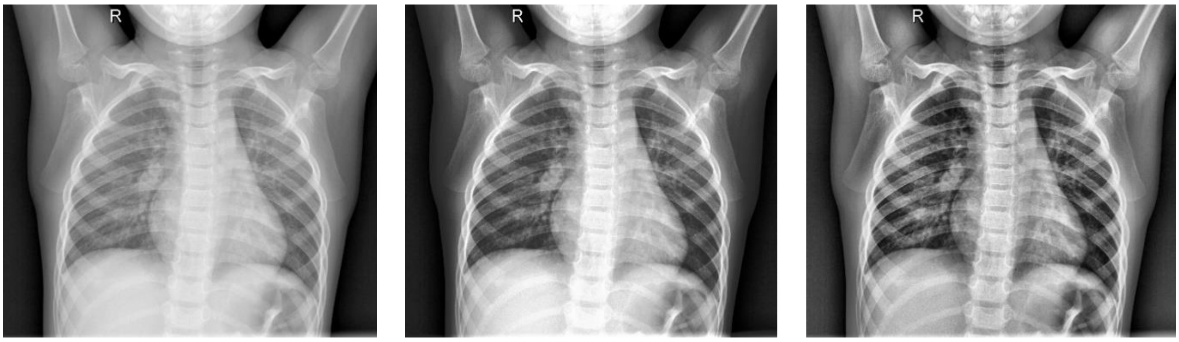
$$s = T(r) = \int_0^T P_r\ (w)\ dw \qquad (2)$$

For all pixels in the images that with gray level in parameter i, hist[i]=hist[i+1], if i from 0 to L-1 for an L value image, from the histogram sequence, presents increasing contrast of histogram histcf[i]=histcf (i-1)+hist[i], generate the equalized histogram as (3):

$$equhist[i]\left\lfloor\frac{[(L*hist_{cf}[i])-N^2}{N^2}\right\rfloor \qquad (3)$$

6. CONTRAST LIMITED ADAPTIVE HISTOGRAM EQUALIZATION (CLAHE)

The CLAHE enhances images in which the noise features of an image do not improved. Images deal with CLAHE get a natural image and simplify the evaluation of many parts of an image. Nevertheless, the reduced contrast enhancement of CLAHE might hinder the facility of a viewer to sense the presence of some weighty gray level density [14], Figure 2 shows the effect of HE and CLAHE:



(a) (b) (c)

Figure 2. The effect of HE and CLAHE where (a) Normal image, (b) HE, and (c) CLAHE

7. DATA AUGMENTATION

Data augmentation in brief is expanding dataset, by making several processing on image to generate multi-images [20], this is popular way that used in deep learning especially if the dataset is not as huge as wanted, where the original images and generated images are used in training and testing, Figure 3 shows the main diagram of data augmentation. Deep learning requires a huge amount of training data in order to successfully learn [21], but sometimes, collecting huge data can be very expensive and unrealistic, preserving transformations by increasing the sample amount. One of the most common tricks for improving the recognition performance is to augment the training data in an intelligent way. There are multiple strategies to achieve this effect. For the network to learn translation as well as rotation invariances, it is often suggested to augment a training dataset of images with the different perspective transformation of images. For instance, you can take an input image and flip it horizontally and add it to the training dataset. Along with horizontal flips, you can translate them by a few pixels among other possible transformations. In this paper, the transformation methods that used is as shown in Figure 4.

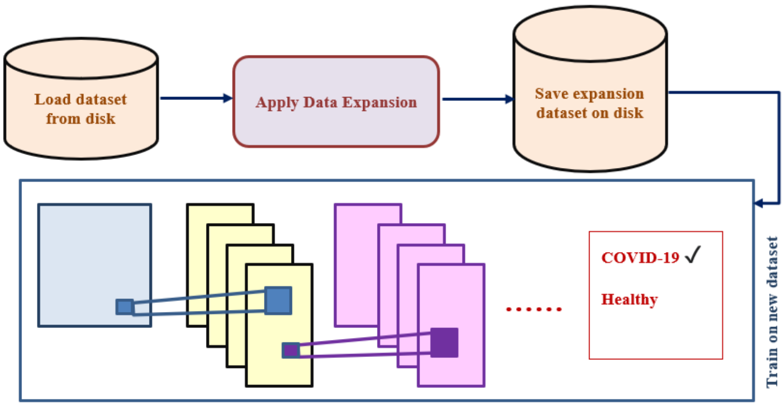


Figure 3. Data augmentation

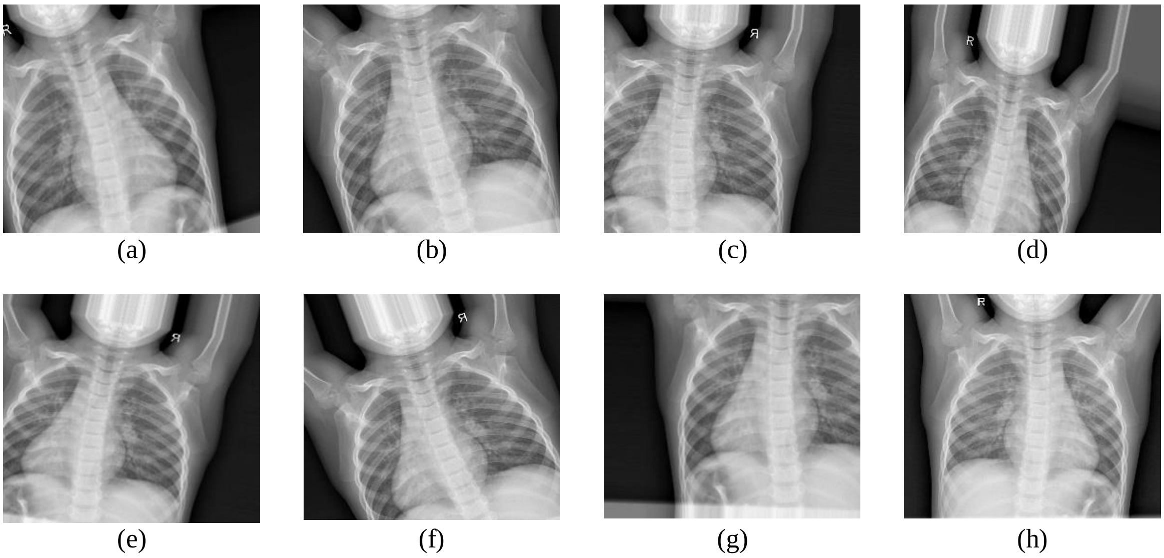


Figure 4. Data augmentation using (a) rotation range=20, (b) zoom range=0.15, (c) shift of width=0.2, (d) shift of height=0.2, (e) shear range=0.15, (f) horizontal flip=true, (g) fill mode=”nearest”, and (h) original image

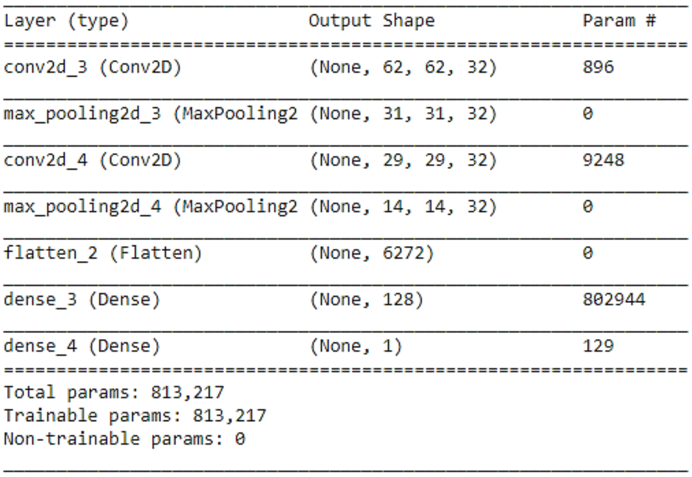
8. CONVOLUTIONAL NEURAL NETWORK

A CNN is a deep learning approach that may income an entered image, allocate the most important (learnable weights and biases) to numerous aspects and objects in the image in order to be able to distinguish one from the other [15]. Involves an input, and output layers in addition to several hidden layers. These hidden layers classically involve a sequence of convolved layers which are convolving with a multiplication or may other dot product [21]. The activation function is generally a rectified linear unit (ReLU) layer, and is consequently followed by other convolutions such as pooling layers, fully connected layers and normalization layers, referred to as hidden layers because their inputs and outputs are masked by the activation function and final convolution [22], [23].

9. DATA TRAINIG

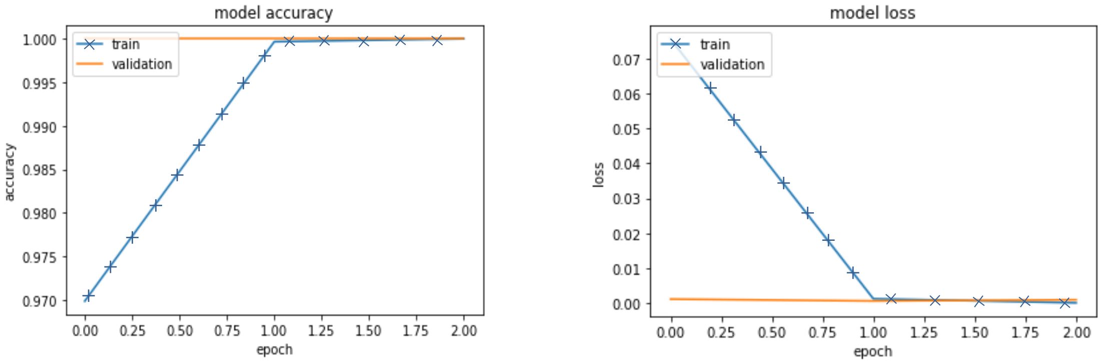
The proposed model that based to train dataset in order to make right decision to judge if the X-ray is infected with COVID-19 or not. First of all, for first convolutional layer, 32 (3x3) filters are applied with ReLU activation, then another 32 (3x3) filters with activation max pooling size (2,2), full connection and

Adam optimizer as denoted in Figure 5, the program of training data is available in [24], [25]. The output accuracy for training after whole epochs is 99% as shown in Figure 6.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
conv2d_3 (Conv2D)            (None, 62, 62, 32)        896       
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 31, 31, 32)        0         
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 29, 29, 32)        9248      
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 14, 14, 32)        0         
_________________________________________________________________
flatten_2 (Flatten)          (None, 6272)              0         
_________________________________________________________________
dense_3 (Dense)              (None, 128)               802944    
_________________________________________________________________
dense_4 (Dense)              (None, 1)                 129       
=================================================================
Total params: 813,217
Trainable params: 813,217
Non-trainable params: 0
_________________________________________________________________
```

Figure 5. Data training model



(a) (b)

Figure 6. Accuracy-loss model (a) model accuracy and (b) model loss

10. CONCLUSION

In this paper, different popular approaches are used for data expansion (data transformation and Image enhancement), where seven layers of data transformation are used also two layers of HE are used, but at the first we uses conceptual hashing algorithm to remove duplicate images, so finally, we can create big variety dataset with increasing rate ((5520-138)/138)*100%=3900%. And also used neural network based deep learning CNN model to train the output dataset in order to make decision if the x-ray image is infected with COVID-19 or not with high accuracy equal 99%.

REFERENCES

[1] N. Razmjooy *et al.*, "Computer-aided diagnosis of skin cancer: a review," *Current Medical Imaging Formerly Current Medical Imaging Reviews*, vol. 16, no. 7, pp. 781–793, Sep. 2020, doi: 10.2174/1573405616666200129095242.

[2] J. Y. Kim *et al.*, "Kinetic heterogeneity of breast cancer determined using computer-aided diagnosis of preoperative MRI scans: relationship to distant metastasis-free survival," *Radiology*, vol. 295, no. 3, pp. 517–526, Jun. 2020, doi: 10.1148/radiol.2020192039.

[3] D. Fresilli *et al.*, "Computer-aided diagnostic system for thyroid nodule sonographic evaluation outperforms the specificity of less experienced examiners," *Journal of Ultrasound*, vol. 23, no. 2, pp. 169–174, Jun. 2020, doi: 10.1007/s40477-020-00453-y.

[4] A. S. Krishnan, D. Clive R., V. Bhat, P. B. Ramteke, and S. G. Koolagudi, "A transfer learning approach for diabetic retinopathy classification using deep convolutional neural networks," in *2018 15th IEEE India Council International Conference (INDICON)*, Dec. 2018, pp. 1–6, doi: 10.1109/INDICON45594.2018.8987131.

[5] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, “Recent advances in recurrent neural networks,” *arXiv preprint arXiv:1801.01078*, Dec. 2017.

[6] Q. Zheng, X. Tian, M. Yang, and H. Wang, “Differential learning: A powerful tool for interactive content-based image retrieval,” *Engineering Letters*, vol. 27, no. 1, pp. 202–215, 2019.

[7] S. Phiphiphatphaisit and O. Surinta, “Food image classification with improved MobileNet architecture and data augmentation,” in *Proceedings of the 2020 The 3rd International Conference on Information Science and System*, Mar. 2020, pp. 51–56, doi: 10.1145/3388176.3388179.

[8] F. F. Alkhalid, “The effect of optimizers in fingerprint classification model utilizing deep learning,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 20, no. 2, pp. 1098–1102, Nov. 2020, doi: 10.11591/ijeecs.v20.i2.pp1098-1102.

[9] A. Mikolajczyk and M. Grochowski, “Data augmentation for improving deep learning in image classification problem,” in *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, May 2018, pp. 117–122, doi: 10.1109/IIPHDW.2018.8388338.

[10] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, “Exploring the landscape of spatial robustness,” in *36th International Conference on Machine Learning, ICML 2019*, 2019, vol. 2019-June, pp. 3218–3238.

[11] Q. Zheng, M. Yang, X. Tian, N. Jiang, and D. Wang, “A full stage data augmentation method in deep convolutional neural network for natural image classification,” *Discrete Dynamics in Nature and Society*, vol. 2020, pp. 1–11, Jan. 2020, doi: 10.1155/2020/4706576.

[12] L. Perez and J. Wang, “The effectiveness of data augmentation in image classification using deep learning,” *Computer Vision and Pattern Recognition*, Dec. 2017.

[13] T. Kauppi *et al.*, “The DIARETDB1 diabetic retinopathy database and evaluation protocol,” in *Procedings of the British Machine Vision Conference 2007*, 2007, pp. 15.1-15.10, doi: 10.5244/C.21.15.

[14] P. R. and A. P., “Diagnosis of diabetic retinopathy using machine learning techniques,” *ICTACT Journal on Soft Computing*, vol. 03, no. 04, pp. 563–575, Jul. 2013, doi: 10.21917/ijsc.2013.0083.

[15] D. Doshi, A. Shenoy, D. Sidhpura, and P. Gharpure, “Diabetic retinopathy detection using deep convolutional neural networks,” in *2016 International Conference on Computing, Analytics and Security Trends (CAST)*, Dec. 2016, pp. 261–266, doi: 10.1109/CAST.2016.7914977.

[16] T. Ai *et al.*, “Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases,” *Radiology*, vol. 296, no. 2, pp. E32--E40, Aug. 2020, doi: 10.1148/radiol.2020200642.

[17] L. L. Wang *et al.*, “CORD-19: The COVID-19 open research dataset,” *Digital Libraries*, Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.10706.

[18] J. P. C, P. M., and L. Dao, “Chest X-ray COVID-19,” Github, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset (accessed Feb. 5, 2020).

[19] X. Luo, D. Wu, C. Chen, M. Deng, J. Huang, and X.-S. Hua, “A survey on deep hashing methods,” *Unpublished*, Mar. 2020.

[20] T. Acharya and A. K. Ray, *Image Processing: Principles and Applications*. Hoboken, NJ, USA, NJ, USA: John Wiley & Sons, Inc., 2005.

[21] Y. Sun, B. Xue, M. Zhang, and G. G. Yen, “Evolving deep convolutional neural networks for image classification,” *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 2, pp. 394–407, 2020, doi: 10.1109/TEVC.2019.2916183.

[22] M. V Valueva, N. N. Nagornov, P. A. Lyakhov, G. V Valuev, and N. I. Chervyakov, “Application of the residue number system to reduce hardware costs of the convolutional neural network implementation,” *Mathematics and Computers in Simulation*, vol. 177, pp. 232–243, Nov. 2020, doi: 10.1016/j.matcom.2020.04.031.

[23] A. Rosebrock, *Deep learning for computer vision with python: Practitioner bundle*. PyImageSearch, 2017.

[24] F. F. Alkhalid, “Data training code.” Github. [Online]. Available: https://github.com/farahuot/COVID-19-Chest-X-Ray/blob/main/Farah-Covid.ipynb (Accessed: Nov. 30, 2021).

[25] A. Muhsin, M. Bashra, B. K. Oleiwi, and F. Alkhaled, “Recognize Arabic handwritten using CNN model,” *Journal of University of Babylon for Pure and Applied Sciences*, vol. 27, no. 6, pp. 359–367, 2019.