www.dergipark.gov.tr
ISSN:2148-3736
El-Cezerî Fen ve Mühendislik Dergisi
Cilt: 9, No: 3, 2022 (1051-1060)

El-Cezerî Journal of Science and Engineering
Vol: 9, No: 3, 2022 (1051-1060)
DOI : 10.31202/ecjse.1061270

ECJSE

Research Paper / Makale

Detection of Face Mask Wearing Condition for COVID-19 Using Mask R-CNN

Ahsen BATTAL[1a*], Adem TUNCER[2b]

[1]Computer Engineering Department, Institute of Graduate Studies, Yalova University, Yalova, Turkiye
[2]Computer Engineering Department, Engineering Faculty, Yalova University, Yalova, Turkiye
[*]battalahsen@gmail.com

Received/Geliş: 21.01.2022 **Accepted/Kabul:** 24.07.2022

Abstract: Due to the COVID-19 pandemic, which has affected the whole world, countries have made it mandatory for people to wear face masks. Because wearing a mask is considered one of the most effective methods to reduce the risk of transmission of the virus. However, it is difficult to manually check whether people are wearing masks. It is aimed to develop a model that detects all kinds of face masks in crowded environments using a deep neural network in this study. Mask R-CNN, which is one of the deep learning algorithms and used for object detection was used to detect and classify people's mask states. The proposed deep learning model was trained and tested with *k*-fold cross-validation using a dataset of 853 images containing three classes (with mask, without a mask, incorrect use of mask). ResNet101 backbone was chosen as the backbone architecture and transfer learning was performed using the MS COCO model. The proposed Mask R-CNN model achieves a mAP of 83%, a mAR of 90%, and an F1 score of 86%. These results reveal that the proposed model is successful in mask detection.

Keywords: Covid-19, deep learning, mask r-cnn, mask detection

Mask R-CNN Kullanarak COVID-19 için Yüz Maskesi Takma Durumunun Tespiti

Öz: Tüm dünyayı etkisi altına alan COVID-19 salgını nedeniyle ülkeler insanların yüz maskesi takmasını zorunlu hale getirdi. Çünkü maske takmak virüsün bulaşma riskini azaltmak için en etkili yöntemlerden biri olarak kabul edilmektedir. Ancak insanların maske takıp takmadığını manuel olarak kontrol etmek zordur. Bu çalışmada derin bir sinir ağı kullanılarak kalabalık ortamlarda her türlü yüz maskesini algılayan bir modelin geliştirilmesi amaçlanmıştır. Derin öğrenme algoritmalarından biri olan ve nesne tespiti için kullanılan Mask R-CNN, insanların maske durumlarını tespit etmek ve sınıflandırmak için kullanıldı. Önerilen derin öğrenme modeli, üç sınıf (maskeli, maskesiz, yanlış maske kullanımı) içeren 853 görüntüden oluşan bir veri seti kullanılarak *k*-kat çapraz doğrulama ile eğitildi ve test edildi. Omurga mimarisi olarak ResNet101 seçildi ve MS COCO modeli kullanılarak transfer öğrenmesi gerçekleştirildi. Önerilen Mask R-CNN modeli, %83'lük bir mAP, %90'lık bir mAR ve %86'lık bir F1 puanına ulaşmıştır. Bu sonuçlar önerilen modelin maske tespitinde başarılı olduğunu ortaya koymaktadır.

Anahtar Kelimeler: Covid-19, derin öğrenme, mask r-cnn, maske tespiti

1. Introduction

COVID-19 virus was first detected in Wuhan, China, in late 2019. Since this virus is easily transmitted to people through droplets and contact, it has taken the whole world under its influence in a short time. The World Health Organization (WHO) declared this epidemic, which affected the whole world, as a pandemic. Then, shopping malls, restaurants, cafes were closed, and curfews

How to cite this article
Battal A., Tuncer A., "Detection of Face Mask Wearing Condition for COVID-19 using Mask R-CNN", El-Cezerî Journal of Science and Engineering, 2022, 9 (3); 1051-1060.

Bu makaleye atıf yapmak için
Battal A., Tuncer A., "Mask R-CNN Kullanarak COVID-19 için Yüz Maskesi Takma Durumunun Tespiti", El-Cezerî Fen ve Mühendislik Dergisi, 2022, 9 (3); 1051-1060.
ORCID: [a]0000-0002-4824-5889; [b]0000-0001-7305-1886

were declared in many countries. Due to the increase in vaccine studies and the decrease in the number of cases since the beginning of the epidemic, many countries have entered a period of normalization. In this period when normalization begins, wearing a mask in public and crowded areas is one of the most effective methods to reduce the risk of transmission of the virus. Therefore, wearing face masks has become one of the priority agendas of the countries and has been made compulsory. Correspondingly, especially in crowded environments, it has emerged as a new problem to determine whether people wear masks and whether they use the mask correctly.

The latest developments in science and technology show that many problems that seem impossible have been overcome. Studies made in many areas with artificial intelligence have started to make people's lives easier. A recent example of this is the fight against COVID-19. According to WHO, it is important for people to wear masks, maintain social distance and avoid contact to prevent the spread of the virus. It is difficult and time-consuming to manually identify people wearing masks, not wearing masks, or using the wrong mask. Deep learning methods, which are frequently used in object detection applications in different fields, and which are a subset of artificial intelligence, are also successfully applied in the mask detection problems.

A convolutional neural network (CNN) is one of the most widely used deep learning approaches. In the literature, there are studies carried out in many different fields, including in the field of agriculture [1] and in the field of health [2], using CNN. Standard CNN models are well-suited to extracting generic descriptions from a dataset and categorizing the input data, but they are insufficient for object detection networks to locate specified objects on an image [3]. To overcome this problem, Region-Based CNN (R-CNN) [4], Faster R-CNN [5], You Only Look Once (YOLO) [6], Mask R-CNN [7] have been developed for object detection. R-CNN proposes about 2.000 regions using Selective Search Algorithm and applies a separate CNN to each region. Faster R-CNN uses a single model instead of 2.000 CNN models and Regional Proposal Network (RPN) instead of Selective Search. YOLO performs object detection in single steps, rather than using a two-step recommendation and classification approach as in R-CNN models. It divides the resulting image into several regions and creates and bounding boxes around each region to estimate probabilities by applying a single neural network to an image [8]. Mask R-CNN, different from the others, generates masks for instance object segmentation and is a slightly expanded version of Faster R-CNN. Instance segmentation defines a class for all pixels of each object in an image.

There are many studies on face mask detection. Bhuiyan et al. [9] used the YOLOv3 architecture to determine whether a person is wearing a mask or not. They indicated that the study achieved an accuracy rate of 96% after training with 4.000 epochs. Similarly, Liu et al. [10] used YOLOv3 for the detection of masks. In addition to mask detection, they also considered the incorrect use of masks in the study. Susanto et al. [11] developed a detector that can detect a variety of face masks. They aimed to detect face masks using the YOLOv4 algorithm. Abbasi et al. [12] proposed a dataset and two different methods for mask detection. In the first method, an object detection model was applied to find and classify masked or unmasked faces. In the second method, the face detector created with YOLO detected faces and was then categorized as masked and unmasked with CNN. They stated that the model achieved a 99.5% accuracy rate is achieved. Gawde [13] used YOLOv3 to detect whether a person is wearing the mask. In the case of not wearing a mask, it was tried to determine how risky a person is by estimating his age. Amin et al. [8], divided their study into two stages as face mask detection and crowd counting. The model gave a warning if there are rule breaks in the face mask. The control of whether people were wearing masks was provided by YOLOv3.

In this study, Mask R-CNN was used for detecting the face mask and determining whether correct use if there is a mask. Mask R-CNN is a deep learning network with sample object segmentation developed by the Facebook Intelligence Research (FAIR) team [7]. In the literature review on Mask

R-CNN, it has been seen that this network is used in many different areas. Singh and Shekhar [14] carried out the road damage detection classification project using Mask R-CNN. Çakıroğlu et al. [15] trained the Mask R-CNN deep learning network for face detection purposes. Bayram et al. [16] developed a system that automatically recognizes license plates from camera images.

In this study, the Mask R-CNN model based on the transfer learning approach with the MS COCO model has been proposed to detect people who wear masks or not, and whether the incorrect wearing of the mask in public and crowded environments. The MS COCO [17] is a pre-trained model on the COCO dataset which has been used for object detection and segmentation with 80 classes. Since in the Mask R-CNN models are generally used Residual Network-101 (ResNet101) [18] as the backbone, ResNet101 was selected in the study.

The rest of the paper is organized as follows: The methodology consisting of the Mask R-CNN, transfer learning, the dataset, and evaluation criteria were presented in Section 2. The experimental results were discussed in Section 3, and the conclusion was presented in Section 4.

2. Methodology

Figure 1 illustrates the block diagram of the face mask detection system. After the dataset is divided into training and test subsets, the Mask R-CNN model is trained with the training subset. The model, whose training has been completed, performs mask detection using the data in the test subset. Detection is carried out for 3 different classes: with mask, without the mask, and incorrect use of the mask. In addition, location information is given for all faces detected in the images.

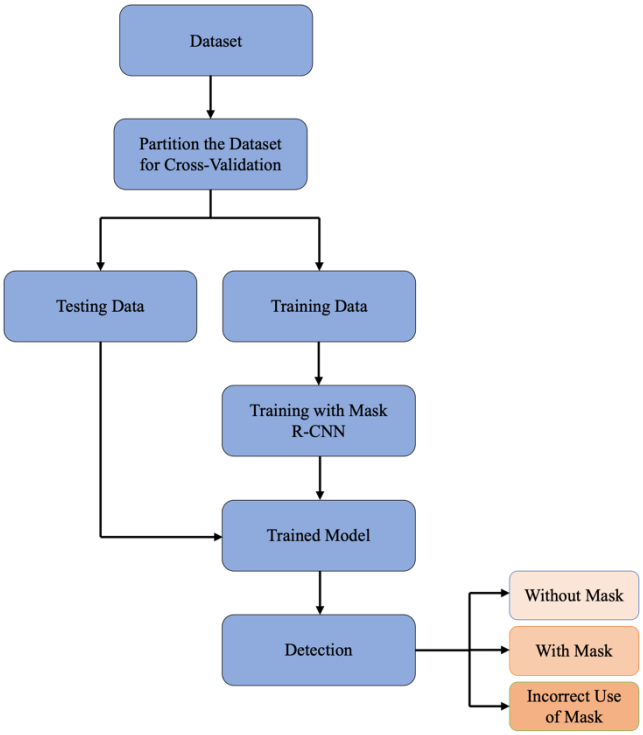

Figure 1. The block diagram of the proposed detection system

2.1. Mask R-CNN

In this section, the two-stage object detection model realized on the input image with Mask R-CNN deep neural network is explained. Mask R-CNN deep neural network aims to solve the sample segmentation problem in computer vision. Instance segmentation assigns a class to each object's

pixels in a picture. As seen in Figure 2, it makes separate definitions for more than one object belonging to the same class.

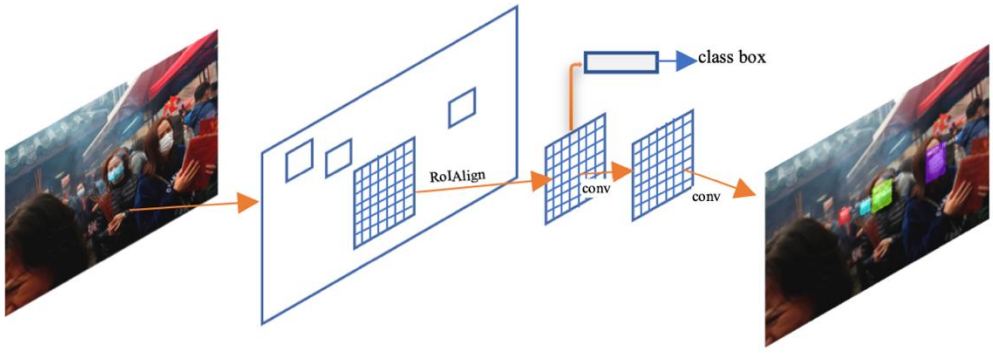

Figure 2. The Mask R-CNN framework for instance segmentation

The Mask R-CNN algorithm used in the study also performs an instance segmentation. The purpose of the algorithm is to detect certain objects on the images of the data as input and to create a mask on each detected object. Mask R-CNN consists of two stages which are suggestion and classification in object detection. In the Mask R-CNN architecture [14], as shown in Figure 3, a feature map is first extracted using a CNN when an image is given as input. The extracted feature map generates candidate frameworks using the Region Proposal Network (RPN). Since the dimensions of the candidate frames created can be different from each other, all frames are brought to the same size by making adjustments in the dimensions of the candidate frames with Region of Interest (RoI) Align. Candidate frames brought to the same size are passed through the fully connected layer to classify and output the frames of the objects in the images. Then the objects inside the box are masked.

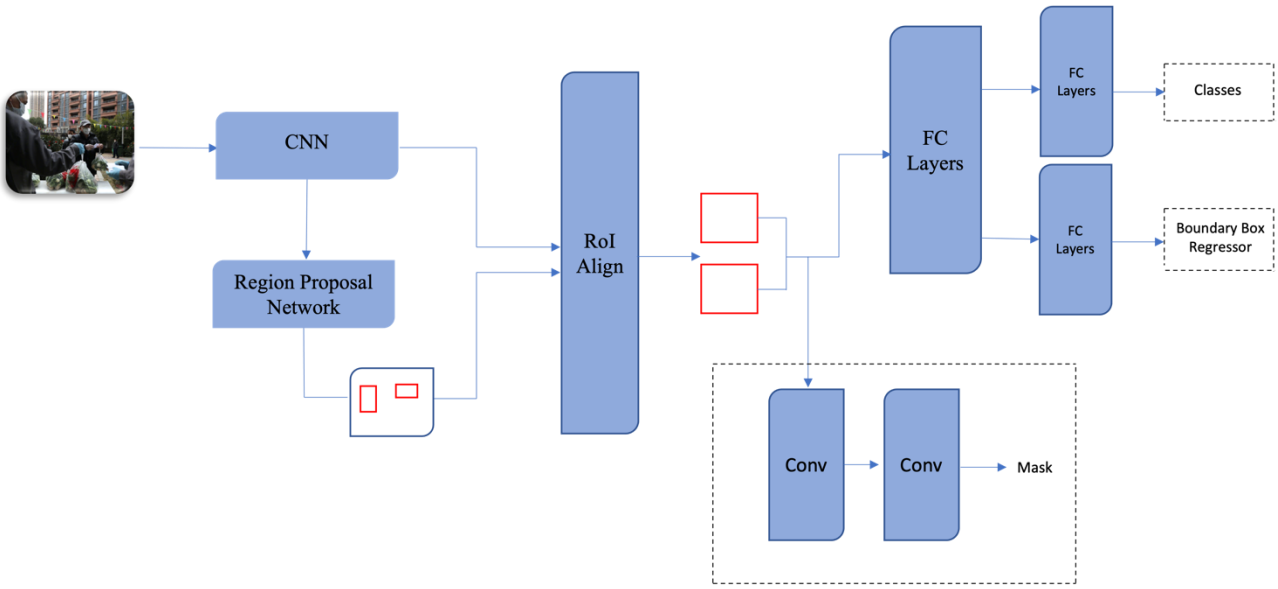

Figure 3. General architecture of Mask R-CNN

2.1.1 Region Proposal Network

The RPN is used to identify objects by scanning the given image and delimiting the identified object with the classifier box. In other words, RPN creates RoIs by sliding windows in the feature map between anchor points of different scales ratios [19]. The regions that the RPN scans are called anchors. The RPN structure is shown in Figure 4.

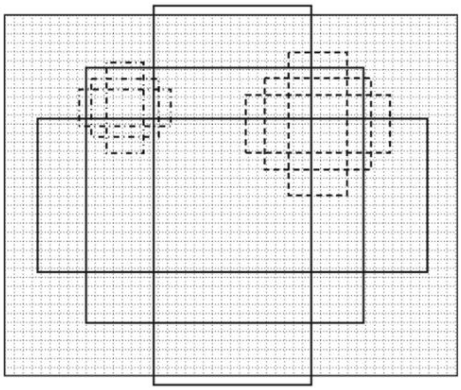

Figure 4. Illustration of RPN structure [19]

2.1.2. Backbone

The backbone network is a standard CNN used as a feature extractor. In the Mask R-CNN backbone, ResNet50 and ResNet101 are known CNN models. In this study, ResNet101 was used as a backbone. As can be seen in Table 1, the ResNet101 architecture consists of 101 layers and classifies with the softmax function. These layers are divided into 6 parts. conv1 layer contains a 7×7 convolution kernel. conv2_x, conv3_x, conv4_x, conv5_x consists of 3, 4, 23, and 3 residual units, respectively. The network ends an average pooling layer, a fully connected layer with an output size of 1000, and a softmax activation function.

Table 1. ResNet101 architecture

layer name	output size	101-layer
conv1	112×112	7×7, 64, stride 2
conv2_x	56×56	3×3 max poll, stride 2
		$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix}\times4$
conv4_x	14×14	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1024 \end{bmatrix}\times23$
conv5_x	7×7	$\begin{bmatrix} 1\times1, & 512 \\ 3\times3, & 512 \\ 1\times1, & 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax
FLOPs		7.6×10^9

2.2. Transfer Learning

Training a model from scratch is generally not preferred because it requires high computer performance. Instead, a previously trained model is retrained by adapting it to the new data set. This solution is called transfer learning. Transfer learning is one of the approaches that increase performance. Transfer learning stores the knowledge obtained while solving any problem and is

used to apply it to other related problems. In this study, transfer learning was successfully applied to the proposed model using the MS COCO model. Instead of training the network from scratch, the weights obtained from pre-training were used in the MS COCO dataset. The configuration file of the pre-trained MS COCO model, which detects 80 classes, was reconfigured to detect the 3 classes which are with mask, without the mask, and incorrect use of mask in this study. Therefore, the hyper-parameters which are the number of classes, maximum/minimum image size, number of epochs, steps per epoch, and validation steps in the configuration file were updated. The training hyper-parameters and their values are shown in Table 2.

Table 2. Training hyper-parameters and their values

Parameters	Values
Backbone	ResNet101
Num of classes	3
Image max dim	512
Image min dim	512
Number of epochs	10
Steps per epoch	640
Validation steps	213
Learning rate	0.001

2.3. Dataset

The dataset named Face Mask Detection used in the study was taken from Kaggle [20]. Labelled images are in PASCAL VOC format. The files of the labelled images contain information such as the name of the image, the size of the image, and the bounding box coordinates. The *k*-fold cross-validation method was applied to evaluate the performance of the model. This method divides the dataset randomly into a number of *k* parts for classification problems. In this study, *k* was taken as 4 and the dataset was divided into 4 parts. 3 parts were used for training and the remaining part was used for testing. A few examples of images from the dataset are shown in Figure 5.

(a) (b) (c)

Figure 5. Example of images from the dataset

2.4. Evaluation Criteria

It is known that it is insufficient to consider only the accuracy values to decide which model is the most accurate in object detection studies. Therefore, precision, recall, and F1 score metrics were also examined in addition to the accuracy in the study. The Precision and Recall values were calculated using the intersection over union (IoU) value for a given IoU threshold. The IoU is the truth value of the bounding boxes of the detected objects. In other words, it is the value resulting from dividing the area of overlap by the area of union. The calculation of the IoU is shown in Figure 6. The precision is the ratio of true positives to all predicted positives. Average precision (AP) is the average among all precision values at various IoU thresholds. Mean average precision (mAP) is the average over the threshold of more than one IoU.

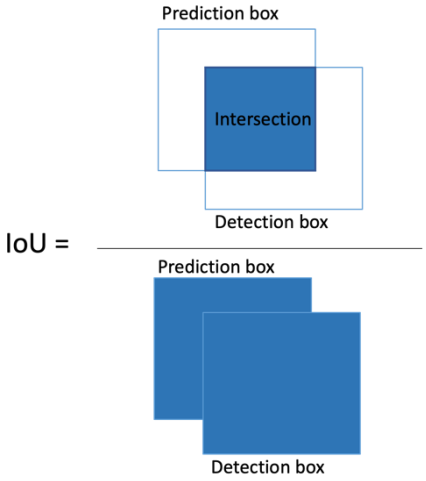

Figure 6. Calculation of IoU

The calculation of the precision is shown in Equation 1.

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

The recall is the ratio of true positives to all actual positives. Average recall (AR) is the average among all recall values at various IoU thresholds. Mean average recall (mAR) is the average over the threshold of more than one IoU. The calculation of the recall is shown in Equation 2.

$$recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

AP and AR are calculated for each class individually while mAP and mAR are the averages of APs and ARs calculated for all classes.

F1 score value shows the harmonic mean of precision and recall values. In other words, it is used when a balance needs to be sought between precision and recall or when there is an imbalance in class distribution [21]. The calculation of the F1 score value is shown in Equation 3.

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

3. Experimental Results

The proposed model in this study is based on Mask R-CNN with ResNet101 architecture. The experiments were conducted on images containing face masks to evaluate the performance of the proposed model. In the training and testing phase of the model, a total of 853 images belonging to three classes, including with mask, without the mask, and incorrect use of mask, were used. In the images in the dataset, there is often more than one class, although there is a small number of single classes. For example, there may be people who do not use masks or who use them incorrectly, together with people who use masks in the same image. The experiments were developed in Google Colaboratory using the Python programming language.

k-fold cross-validation technique with *k*=4 was applied to assess the accuracy of the model. In this technique, the dataset was divided into 4 different subsets. One of the subsets was used for testing and the remaining 3 subsets were used for training. In each training phase, the test dataset was

changed and thus 4 different training processes were performed on the whole dataset. The images used in the testing phase are shown in Figure 7.

(a) (b) (c)

Figure 7. Examples from the testing phase

Figure 8 shows the loss graphics of the proposed Mask R-CNN model with 4-fold cross-validation. Besides, mAP, mAR, and F1 score values from test datasets are shown in Table 3.

Table 3. The performance results from test datasets

Tests	mAP	mAR	F1 score
4-Fold #1	0.81	0.90	0.85
4-Fold #2	0.84	0.92	0.88
4-Fold #3	0.85	0.91	0.88
4-Fold #4	0.81	0.89	0.84
Average	0.83	0.90	0.86

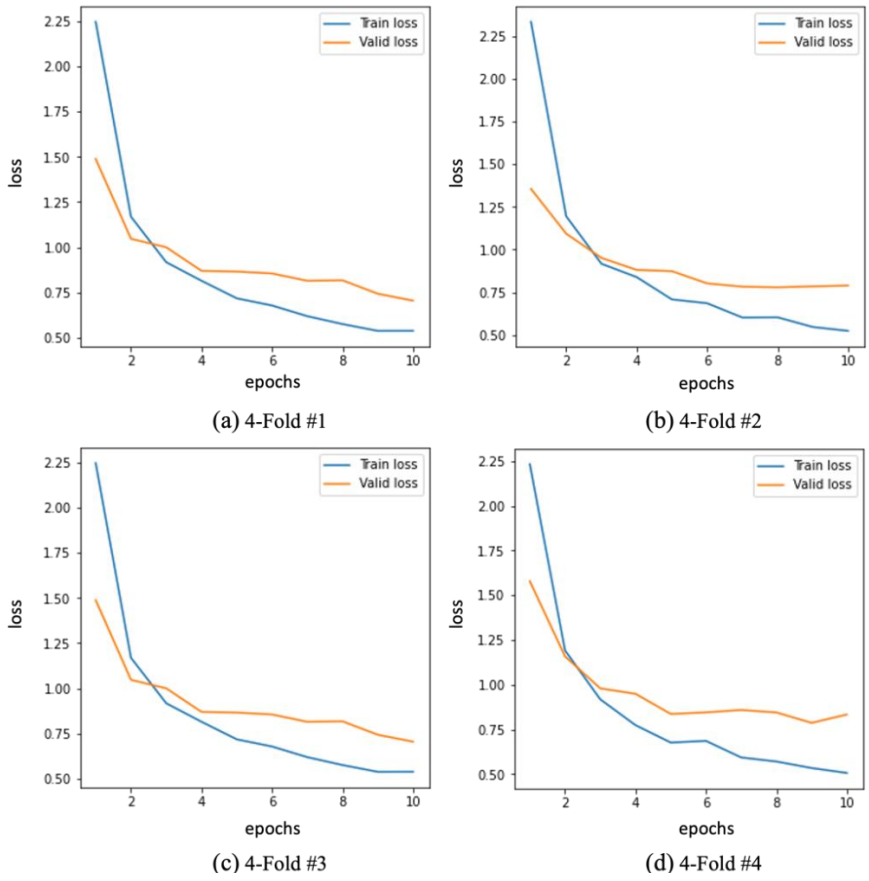

Figure 8. Loss graphics of the proposed Mask R-CNN model with 4-fold cross-validation

4. Conclusion

In this study, a Mask R-CNN model has been proposed to detect people who wear masks or not, and whether the incorrect use of mask of the mask in public and crowded environments. Studies in the literature are generally carried out on datasets containing 2 classes with the mask or without the mask. In this study, in addition to detecting the mask, it was also tried to determine whether it was used incorrectly. It was applied transfer learning to the proposed model using pre-trained MS COCO. The model's performance was assessed using the *k*-fold cross-validation approach. The proposed Mask R-CNN model with a ResNet101 backbone achieves a mAP of 83%, a mAR of 90%, and an F1 score 86%.

Authors' Contributions

AB and AT proposed the idea of the study together. AB carried out the experimental studies, simulations and wrote the initial draft. AT contributed to the study by supervising and interpreting. AB and AT wrote the final version of the paper. All authors read and approved the final version of the paper.

Conflict of Interests

The authors declare that they have no conflict of interest.

References

[1]. Sardogan, M., Tuncer, A., and Ozen, Y., "Plant Leaf Disease Detection and Classification Based on CNN with LVQ Algorithm", In 2018 3rd International Conference on Computer Science and Engineering (UBMK), IEEE, 382-385, (2018).

[2]. Orman, A., Köse, U., and Yiğit, T., "Açıklanabilir Evrişimsel Sinir Ağları ile Beyin Tümörü Tespiti", El-Cezeri Fen ve Mühendislik Dergisi, 2021, 8(3): 1323-1337.

[3]. Sardogan, M., Özen, Y., and Tuncer, A., "Detection of Apple Leaf Diseases using Faster R-CNN", Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 2020, 8(1): 1110-1117.

[4]. Girshick, R., Donahue, J., Darrell, T., and Malik, J., "Region-Based Convolutional Networks for Accurate Object Detection and Segmentation", IEEE Trans. Pattern Anal. Mach. Intell., 2015, 38(1): 142-158.

[5]. Ren, S., He, K., Girshick, R., and Sun, J., "Faster R-CNN: Towards Realtime Object Detection with Region Proposal Networks", IEEE Trans. Pattern Anal. Mach. Intell., 2017, 39(6), 1137- 1149.

[6]. Redmon, J., Farhadi, A., "YOLOv3: An Incremental Improvement", 2018, arXiv preprint arXiv:1804.02767.

[7]. He, K., Gkioxari, G., Dollár, P., and Girshick, R., "Mask R_CNN, Proceedings of the IEEE International Conference on Computer Vision (ICCV)", 2961-2969, (2017).

[8]. Amin, P. N., Moghe, S. S., Prabhakar, S. N., and Nehete, C. M., "Deep Learning Based Face Mask Detection and Crowd Counting", In 2021 6th International Conference for Convergence in Technology (I2CT), IEEE, 1-5, (2021).

[9]. Bhuiyan, M. R., Khushbu, S. A., and Islam, M. S., "A Deep Learning Based Assistive System to Classify Covid-19 Face Mask for Human Safety with YOLOv3", In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, 1-5, (2020).

[10]. Liu, R., and Ren, Z., "Application of Yolo on Mask Detection Task", In 2021 IEEE 13th International Conference on Computer Research and Development (ICCRD), IEEE, 130-136, (2021).

[11]. Susanto, S., Putra, F. A., Analia, R., and Suciningtyas, I. K. L. N., "The Face Mask Detection for Preventing the Spread of COVID-19 at Politeknik Negeri Batam", In 2020 3rd International Conference on Applied Engineering (ICAE), IEEE, 1-5, (2020).
[12]. Abbasi, S., Abdi, H., and Ahmadi, A., "A Face-Mask Detection Approach based on YOLO Applied for a New Collected Dataset", In 2021 26th International Computer Conference, Computer Society of Iran (CSICC), IEEE, 1-6, (2021).
[13]. Gawde, B. B., "A Fast, Automatic Risk Detector for COVID-19", In 2020 IEEE Pune Section International Conference (PuneCon), IEEE, 146-151, (2020).
[14]. Singh, J., and Shekhar, S., "Road damage detection and classification in smartphone captured images using mask r-cnn", arXiv preprint arXiv:1811.04535, (2018).
[15]. Cakiroglu, O., Ozer, C., and Gunsel, B., "Design of a deep face detector by mask r-cnn", In 2019 27th Signal Processing and Communications Applications Conference (SIU), IEEE, 1-4, (2019).
[16]. Bayram, F., "Derin öğrenme tabanlı otomatik plaka tanıma", Politeknik Dergisi, 2020, 23(4): 955-960.
[17]. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., and Zitnick, C. L., "Microsoft coco: Common objects in context, European conference on computer vision", Springer, Cham, 740-755, (2014).
[18]. He, K., Zhang, X., Ren, S., Sun, J., "Deep residual learning for image recognition", Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778, (2016)
[19]. Lin, K., Zhao, H., Lv, J., Li, C., Liu, X., Chen, R., and Zhao, R., "Face detection and segmentation based on improved mask r-cnn", Discrete dynamics in nature and society, (2020).
[20]. Mask Dataset. [Online]. Available: https://www.kaggle.com/andrewmvd/face-mask-detection
[21]. Chang, Y. Y., Li, P. C., Chang, R. F., Yao, C. D., Chen, Y. Y., Chang, W. Y., and Yen, H. H., "Deep learning-based endoscopic anatomy classification: an accelerated approach for data preparation and model validation", Surgical Endoscopy, 2021, 1-11.