Enhancing face mask detection performance with comprehensive dataset and YOLOv8

Trong Thua Huynh, Hoang Thanh Nguyen
Information Security Technology Lab, Faculty of Information Technology, Posts and Telecommunications Institute of Technology, Ho Chi Minh City, Vietnam

Article Info

Article history:

Received Aug 5, 2024
Revised Mar 26, 2025
Accepted Jun 8, 2025

Keywords:

Dataset
Deep learning
Face mask detection
Object detection
YOLOv8

ABSTRACT

In the context of the COVID-19 pandemic and the risk of similar infectious diseases, monitoring and promoting public health measures like wearing face masks have become crucial in controlling virus transmission. Deep learning-based mask recognition systems play an important role, but their effectiveness depends on the quality and diversity of training datasets. This study proposes the diverse and robust dataset for face mask detection (DRFMD), designed to address limitations of existing datasets and enhance mask recognition models' performance. DRFMD integrates data from sources such as AIZOO, face mask detector by Karan-Malik (KFMD), masked faces (MAFA), MOXA3K, properly wearing masked face detection dataset (PWMFD), and the Zalo AI challenge 2022, comprising 14,727 images with 29,846 instances, divided into training, validation, and testing sets. The dataset's scale and diversity ensure higher accuracy and better generalization for mask recognition models. Experiments with variations of the YOLOv8 model (n, s, m, l, x), an advanced object detection algorithm, on the DRFMD dataset, demonstrate superior performance through metrics like precision, recall, and mAP@50. Additionally, comparisons with previous dataset like FMMD show that models trained on DRFMD maintain strong generalization capabilities and higher performance. This study significantly contributes to improving accuracy of public health monitoring systems, aiding in the prevention of hazards from infectious diseases and air pollution.

Corresponding Author:

Trong Thua Huynh
Information Security Technology Lab, Faculty of Information Technology
Posts and Telecommunications Institute of Technology
11 Nguyen Dinh Chieu Street, District 1, Ho Chi Minh City, Vietnam
Email: thuaht@ptit.edu.vn

1. INTRODUCTION

In an era marked by global health challenges, technology plays a crucial role in public safety, particularly through face mask recognition technology. This innovation uses advanced algorithms and deep learning to accurately identify individuals wearing masks, ensuring compliance with public health guidelines. The urgent need for effective preventive measures during the COVID-19 pandemic highlighted the limitations of manual enforcement, prompting the development of automated solutions like face mask recognition. This technology reliably monitors compliance in crowded public spaces such as airports and shopping centers, reducing virus transmission risk and enhancing public health safety [1], [2]. Beyond monitoring, it aids data collection and analysis, offering insights into compliance rates and the success of health campaigns. Additionally, its integration with other surveillance systems forms a comprehensive public safety network, boosting preparedness for future health crises.

Deep learning approaches used for mask recognition focus on identifying and categorizing individuals according to their adherence to mask-wearing protocols. Among these methods, convolutional neural networks (CNNs) stand out due to their exceptional capabilities in image analysis, making them particularly effective for mask detection applications [3]. Fan *et al*. [4] proposed a lightweight mask detection suite based on deep learning, combining the residual context attention module (RCAM) and Gaussian heatmap regression (SGHR) to improve feature extraction capabilities. Using MobileNet as the backbone, this model is suitable for embedded systems and achieves high performance. RCAM helps focus on important mask-related areas, while SGHR learns the distinguishing features between masked and unmasked faces. Tests on the AIZOO and Moxa3K datasets showed the model achieved higher mean average precision (mAP) than YOLOv3-tiny by 1.7% and 10.47%, respectively. Joodi *et al.* [5] introduces a novel deep learning model for mask detection, structured in two stages: face detection using the Haar cascade detector and classification with a CNN model built from scratch. Experiments conducted using the benchmark masked faces (MAFA) dataset achieved relatively high accuracy across different learning rates while maintaining low computational complexity. Similarly, the proposed model in [6] enables real-time mask-wearing recognition based on the MobileNetV2 architecture, applicable to embedded devices such as the NVIDIA Jetson Nano. Experimental results indicate a very high accuracy rate in both training and testing. The model is also designed to be lightweight and efficient, supporting multi-mask detection, which is beneficial in crowded environments where multiple individuals need to be monitored simultaneously.

Several studies have explored deep learning models for mask detection. Research by Khoramdel *et al.* [7], three models (SSD, YOLOv4-tiny, and YOLOv4-tiny-3l) were tested on 1,531 images, with YOLOv4-tiny achieving the highest mAP (85.31%) and 50.66 FPS, suitable for real-time use. Al-Dmour *et al.* [8] proposed a CNN-based system to recognize covered faces, achieving high accuracy in distinguishing masked from unmasked faces. Additionally, the DB-YOLO mask detection algorithm [9], integrated into an Android app, demonstrated high precision and a detection speed of 33 FPS using a lightweight architecture based on YOLOv5, optimized for mobile devices. Aburaed *et al.* [10] compared YOLOv5 and YOLOv6 for detecting impact craters on mars and the moon. The results indicate that YOLOv6 outperformed YOLOv5 in speed and accuracy with Adam optimizer.

The YOLOv8 algorithm [11] represents an advanced object detection framework, renowned for its excellent accuracy, real-time processing capabilities, and robust performance. As an evolution of the YOLO series, YOLOv8 maintains the fundamental principle of performing object detection in a single pass through a neural network [12], [13], making it highly efficient and suitable for real-time ap-plications. YOLOv8 employs a deep convolutional neural network (DCNN) with multiple convolutional layers, down-sampling, and up-sampling operations. This architecture allows for capturing features at various scales and preserving crucial spatial information for precise object identification and localization. YOLOv8 processes the input image by segmenting it into a grid, where each cell is responsible for predicting bounding boxes along with their corresponding class probabilities. This structured grid-based technique allows YOLOv8 to effectively identify multiple objects within an image, accommodating variations in size and aspect ratio [10]. Moreover, YOLOv8 integrates advanced techniques such as batch normalization, dropout, and complex activation functions, improving accuracy and recall rates compared to previous versions. These enhancements reduce errors and improve reliability in detecting objects across different scenarios.

Dewi *et al.* [14] introduced and utilized a dataset we refer to as face and medical mask dataset (FMMD), which is a combination of the face mask dataset (FMD) [15] and the medical mask dataset (MMD) [16], in training and evaluating mask recognition. However, FMMD still has some limitations, such as small scale, lack of data source diversity, and insufficient label quality and detail, with only 1,067 images and 5,796 instances. To address these limitations and enhance the effectiveness of mask recognition models, we propose a new, more diverse, and robust dataset called diverse and robust dataset for face mask detection (DRFMD). Additionally, this work explores and analyzes the human in the loop (HITL)-MMD [17], an open-access dataset designed to contribute to the global fight against COVID-19. HITL-MMD provides a rich and diverse data source, supplementing and improving upon existing research.

The main contributions of this work are: i) constructing a large-scale and diverse dataset with 10,304 images and 20,603 instances to improve recognition performance compared to previous datasets. ii) implementing a deep learning-based object detection model that can automatically identify and locate faces without masks, faces with masks, and improperly worn masks in images. iii) analyzing and comparing YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x models to identify and evaluate the benefits and limitations related to using YOLOv8 in mask recognition systems. Besides the introduction, we will present the dataset construction methodology in section 2. Section 3 illustrates training results, evaluation and discussion. Finally, conclusion will be provided in section 4.

2. METHOD

In this section, we briefly introduce two datasets used in previous studies, namely FMMD [14] and HITL-MMD [17]. These datasets play a crucial role in research on face recognition with and without masks, providing a reliable reference source for computer vision-based recognition systems. Next, we present a detailed explanation of the methodology used to construct the proposed dataset, called DRFMD. This section will cover the data collection process, the criteria for image selection, preprocessing steps, and key features that make DRFMD a diverse and robust dataset. These characteristics ensure its effectiveness in supporting models for face mask detection.

2.1. Face mask dataset and medical mask dataset

Dewi *et al.* [14] introduced and used the FMMD dataset, which is a combination of FMD [15] and MMD [16], for training and evaluating mask recognition with an input resolution (image size) of 416. The FMD is a publicly available dataset of MAFA, which includes 853 images, stored in PASCAL VOC format. The MMD includes 682 images, with over 3,000 MAFA wearing medical masks. The combination of these two datasets resulted in a distinct and more extensive dataset, with a total of 1,415 collected images undergoing a rigorous selection process. Low-quality or duplicate images from the original dataset were removed to ensure the quality and consistency of the final dataset. We will use this dataset to evaluate and compare it with our proposed dataset (DRFMD) in the next section.

2.2. Human in the loop medical mask dataset

HITL [17] provides an open-access dataset designed to support global efforts in combating COVID-19. This dataset comprises 6,000 publicly available images, carefully curated to ensure diversity by including individuals from various ethnic backgrounds, age groups, and geographic regions. Furthermore, it incorporates 20 distinct types of accessories and categorizes facial images into three groups: wearing masks correctly, not wearing masks, and improper mask usage. The dataset was compiled and annotated by refugee workers affiliated with HITL in Bulgaria. To promote accessibility and broader usage, this MMD has been released into the public domain under the CC0 1.0 license. In this study, we use the HITL-MMD dataset to evaluate the model trained on the previous FMMD dataset and our proposed dataset. For the evaluation, we used the LabelImg tool [18] to annotate data for 1,311 images with a total of 1,598 instances, where the labels without mask (0), with mask (1), and wear mask incorrect (2) are 462, 1,030, and 106 respectively.

2.3. Diverse and robust dataset for face mask detection

The dataset we propose in this study is called DRFMD, which is collected from various sources with partial or complete data from AIZOO [19], face mask detector by Karan-Malik (KFMD) [20], MOXA3k [21], MAFA [22], and properly wearing masked face detection dataset (PWMFD) [23]. The dataset construction method is described in Figure 1. The DRFMD dataset is built by aggregating various data sources to enrich the final dataset. Each input dataset is partially or fully used, then reviewed and refined as necessary, and finally converted to YOLO labeling standards. Additionally, we collected and labeled some data from the Zalo AI challenge 2022 dataset [24], and adjusted some labels of the aforementioned datasets.

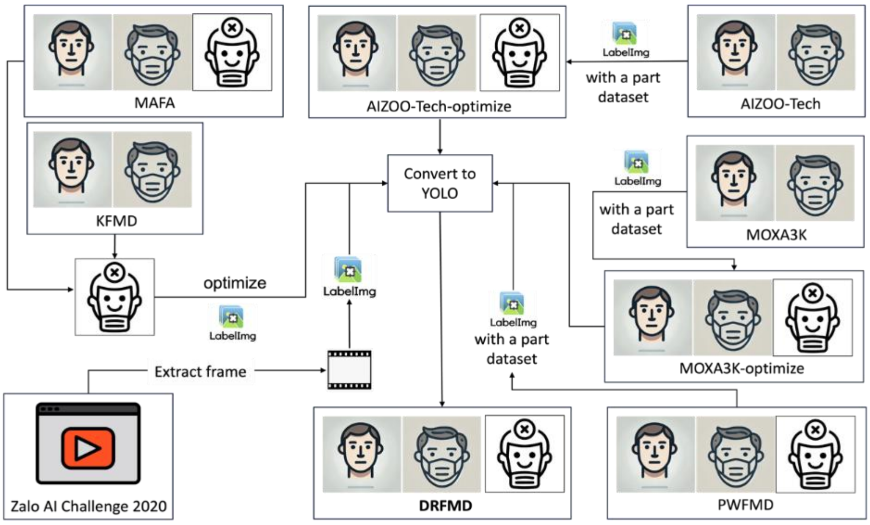


Figure 1. The process of creating the DRFMD dataset

The DRFMD dataset includes 14,727 images with 29,846 instances, comprising 10,304 instances for the training set, 1,474 instances for the validation set, and 2,949 instances for the test set. We noticed that FMD often lack the number of instances of improperly worn masks. Therefore, we focused on extracting all images with improperly worn masks from the MAFA and KFMD datasets. The KFMD dataset [20], which originally only had labels for wearing masks and not wearing masks, was annotated with improperly worn mask labels using the LabelImg tool. We then adjusted the labels to ensure the accuracy of each label. For the AIZOO-Tech and MOXA3K datasets, which only had labels for wearing masks and not wearing masks, we added improperly worn mask labels and extracted part of the data. The PWFMD dataset is quite large, with 9,205 images of varying sizes, from small to large, so we extracted a subset of 4,703 images. The parameters of these data sources are detailed in Table 1, with all images in the datasets being real images. Most datasets are labeled for all image sizes (small, medium, large), except for the MAFA dataset, which is labeled only for medium and large sizes.

Table 1. Summary of some parameters of open datasets for face detection to build the DRFMD [25]

Dataset name	Key features	Number of images	Classification	Mask number	Head position	Scence
AIZOO-Tech	The dataset is created by modifying incorrect annotations from the WIDER Face and MAFA datasets	Train: 6,130 Valid: 1,839	Two	12,620 faces without masks; 4,034 MAFA	Diversity	Medium
KFMD [20]	Images created by Karan	1,508	Two	753 faces without mask; 755 face with masked;	Diversity	Medium
MAFA	The images are from the Internet, annotated with six attributes for each face region, similar to the more occluded face datasets	30,811	Many types of masks	35,806 MAFA	Diversity	Complex
MOXA3K	The images come from Kaggle datasets featuring data from Russia, Italy, China, and India during the ongoing pandemic	3,000	Two	9,161 faces without masks; 3015 MAFA	Diversity	Complex
PWMFD	More than half of the images are collected from WIDER Face, MAFA, and RWMFD. The “With Mask” class requires covering both the face and nose	9,205	Three	10,471 faces without masks; 7,695 correct MAFA; 366 incorrect MAFA	Frontal to Profile	Medium

The liveness detection dataset - Zalo AI challenge 2022 is used to train and test models for face liveness detection, aiming to distinguish between real and fake faces. It includes multiple videos under various lighting conditions and contexts. For this dataset, we extracted frames from multiple videos and manually labeled them using LabelImg [18], a software tool designed for visually annotating and identifying objects within images. Finally, all images and labels from the aforementioned datasets were converted to YOLO standard labeling format. Consequently, our DRFMD dataset is more diverse in terms of the number of images, mask types, distribution of mask-wearing, non-mask-wearing, and improper mask-wearing, head positions, surrounding contexts, and image dimensions. The DRFMD dataset consists of a total of 14,727 images, which are categorized into three subsets: i) a training set containing 10,304 samples, ii) a validation set comprising 1,474 samples, and iii) a testset with 2,949 samples. Further details about the dataset parameters can be found in Table 2.

Table 2. Synthetic parameters of the DRFMD dataset

Dataset	Train	Valid	Test	Total
AIZOO	2,686	383	782	3,851
KFMD [18]	495	64	126	685
MAFA	1,006	141	289	1,436
MOXA3k	999	144	296	1,439
PWMFD	3,283	482	938	4,703
Zalo AI Chalenge 2022	1,835	260	518	2,613
DRFMD	10,304	1,474	2,949	14,727

According to Table 3 (DRFMD dataset column), the data labels were reannotated to YOLO standards with three categories: without mask (0) with 14,157 labels, with mask (1) with 11,590 labels, and wear mask incorrect (2) with 2,866 labels. Clearly, when comparing all parameters (images, instances without mask, instances with mask, instances wear mask incorrect) across different datasets (train, valid, test), the DRFMD dataset demonstrates a significant advantage in quantity. Specifically, it contains approximately 10 times more images and around 4 times more instances than the FMMD dataset. Additionally, in DRFMD, the distribution of Instances with mask and Instances without mask is more balanced compared to the FMMD dataset.

Table 3. Key parameters of the DRFMD dataset vs FMMD dataset

Parameters	DRFMD dataset				FMMD dataset		
	Train	Valid	Test	Total	Train	Valid	Test
Images	10,304	1,474	2,949	14,727	1,067	456	507
Instances	20,603	3,052	6,191	29,846	5,796	2,156	2,663
Without Mask (0)	9,683	1,392	3,082	14,157	1,030	352	449
With Mask (1)	8,926	1,370	2,527	11,590	4,589	1,728	2,122
Wear Mask Incorrect (2)	1,994	290	582	2866	177	76	92

3. RESULTS AND DISCUSSION

3.1. Training result

To evaluate the effectiveness of object detection models, average precision (AP) is commonly used, incorporating key metrics such as intersection over union (IoU), precision, and recall. As mentioned in [26], these metrics are mathematically defined in (1) to (3). IoU measures the overlap between the predicted bounding box (pred) and the actual ground truth box (gt). Precision assesses the accuracy of the model's outputs, whereas Recall evaluates its ability to detect all gt instances.

$$IoU = \frac{Area_{intersect}}{Area_{match}} = \frac{Area_{pred} \cap Area_{gt}}{Area_{pred} \cup Area_{gt}} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} = \frac{TP}{N} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

where TP represents true positive, FP represents false positive, FN represents false negative, and N represents the total number of recovered objects, including true positives and false positives.

In object detection, multiple object classes need to be identified. The mAP index is used to compute the AP for each class and then derive the overall average. This metric provides a comprehensive evaluation of the model's performance across all categories, considering the varying difficulty levels in detecting different objects. According to Dewi *et al.* [14], mAP is defined in (4), where the variable p(o) represents detection accuracy.

$$mAP = \int_0^1 p(o)do \quad (4)$$

Data augmentation is a widely adopted technique in deep learning, aimed at enhancing the variability of a training dataset by applying different transformations to the original data. Throughout the training process, various augmentation methods, including padding, cropping, and horizontal flipping, among others, are utilized. These techniques play a crucial role in the development of large-scale neural networks due to their effectiveness in improving model generalization. In our experiment, we trained the model for 100 epochs with a weight decay of 0.0005, an initial learning rate of 0.01, a final learning rate of 0.01, a batch size of 16, an input image size of 640, and an IoU threshold of 0.7. Furthermore, we applied a Mosaic configuration of 1.0 for the first 90 epochs, set close_mosaic to 10, and used mixup at 0.243. Additional data augmentation parameters included hsv_h at 0.0138, hsv_s at 0.664, hsv_v at 0.464, translate at 0.1, scale at 0.898, and shear at 0.602.

All the above configurations will be used to train all five YOLOv8 models on both the FMMD and DRFMD datasets. We then evaluate the results for all trained models on the testset of the three datasets FMMD, DRFMD, and HITL-MMD. The training environment for the models was carried out on a Dell R730 server, which includes two Nvidia Tesla P40 GPU accelerators with 24 GB of RAM each, two Xeon E2680v4 central processors, and 128 GB of DDR4 2400 bus memory. The YOLOv8 training process was performed on two GPUs to achieve real-time detection capabilities.

The training results of the YOLOv8 models (n, s, m, l, x) on the FMMD dataset (456 images, 2156 instances) and the DRFMD dataset (1,474 images, 3,052 instances) are shown in Table 4. In this table, we only present the overall training results for all images. Detailed results for the cases without mask (0), with mask (1), and wear mask incorrect (2) are published in [27]. The training performance is evaluated on the metrics precision (P), recall (R), and mAP@50 (represents the mAP value when the IoU is 0.5). The specific results are: i) YOLOv8n (nano) achieved P=0.838, R=0.796, and mAP@50=0.849. Despite being the smallest version, this model shows relatively high accuracy and recognition capability; ii) YOLOv8s (small) significantly improved with P=0.845, R=0.820, and mAP@50=0.872, indicating enhanced detection of harder objects; iii) YOLOv8m (medium) further enhanced performance with P=0.862, R=0.823, and mAP@50=0.888, demonstrating stronger object recognition capabilities; iv) YOLOv8l (large) maintained high accuracy with P=0.859, R=0.844, and mAP@50=0.889, indicating stable performance in object detection and classification; v) YOLOv8x (extra-large) achieved the highest performance with P=0.859, R=0.838, and mAP@50=0.895, showcasing superior object recognition and classification capabilities.

Table 4. Training performance for 5 models YOLOv8 on FMMD and DRFMD datasets

Model	FMMD dataset (456 images, 2,156 instances)			DRFMD dataset (1,474 images, 3,052 instances)		
	P	R	mAP@50	P	R	mAP@50
YOLOv8n	0.904	0.777	0.868	0.838	0.796	0.849
YOLOv8s	0.945	0.863	0.924	0.845	0.820	0.872
YOLOv8m	0.966	0.879	0.948	0.862	0.823	0.888
YOLOv8l	0.979	0.898	0.956	0.859	0.844	0.889
YOLOv8x	0.955	0.924	0.963	0.859	0.838	0.895

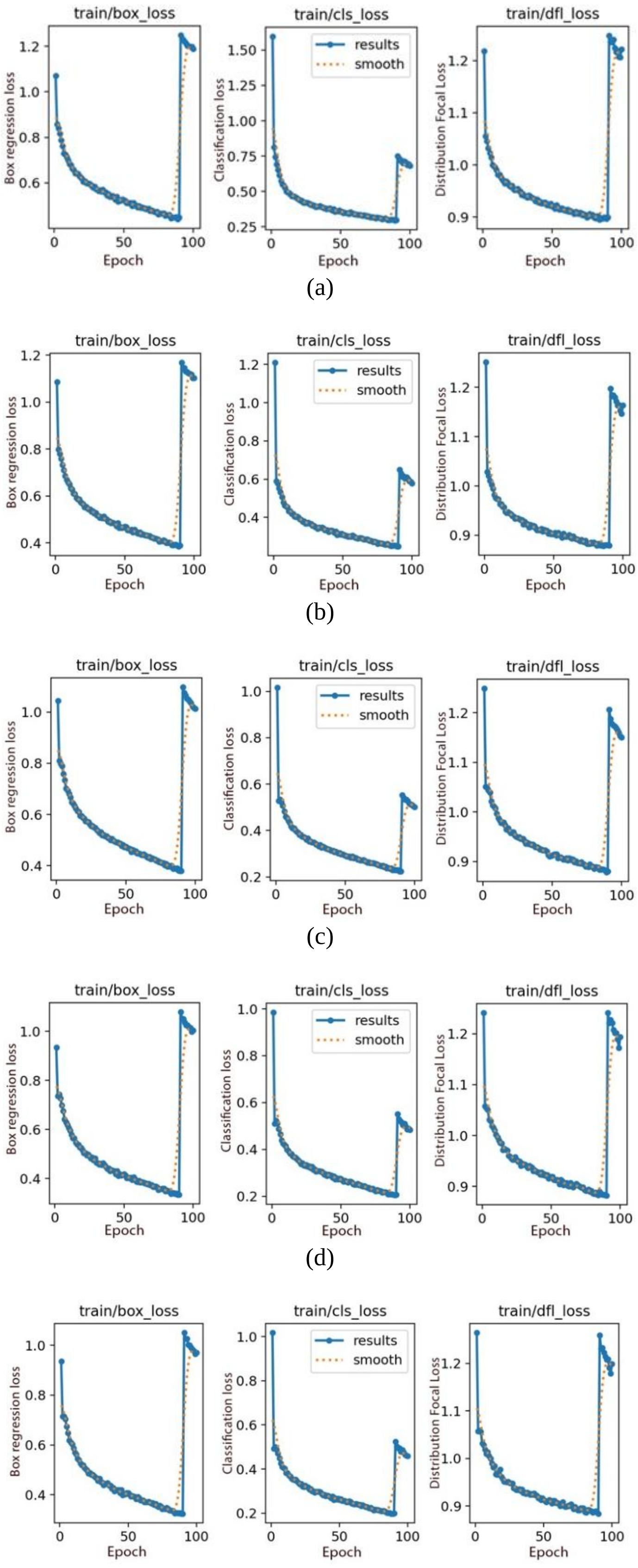
Figure 2 illustrates the training charts of YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x over 100 epochs on the DRFMD dataset. The loss values for YOLOv8n as shown in Figure 2(a) are box_loss=1.188, cls_loss=0.68317, and dfl_loss=1.2216. Similarly, YOLOv8s as shown in Figure 2(b) has box_loss=1.1024, cls_loss=0.5783, and dfl_loss=1.1639, while YOLOv8m as shown in Figure 2(c) records box_loss=1.0167, cls_loss=0.504, and dfl_loss=1.1531. For YOLOv8l as shown in Figure 2(d), the values are box_loss=1.0036, cls_loss=0.48459, and dfl_loss=1.1938, whereas YOLOv8x as shown in Figure 2(e) exhibits box_loss=0.97032, cls_loss=0.45936, and dfl_loss=1.1992. These results suggest that larger models (YOLOv8m, YOLOv8l, YOLOv8x) outperform smaller ones (YOLOv8n, YOLOv8s) across all types of loss functions. As models become more complex, box_loss and cls_loss decrease significantly, highlighting their improved object localization and classification capabilities. However, dfl_loss remains relatively stable across different models, with only minor fluctuations, indicating that their ability to learn distribution weights does not vary significantly.

3.2. Evaluation results and discussion

This evaluation section is conducted based on the training results using both the FMMD and DRFMD datasets as presented in section 3.1. For convenience in discussing the results, we distinguish as follows: i) the training results with the FMMD dataset for all five YOLOv8 models are referred to as YOLOv8 models with FMMD dataset (YwFMMD); and ii) the training results with the proposed DRFMD dataset for all five YOLOv8 models are referred to as YOLOv8 models with DRFMD dataset (YwDRFMD). In the next sections, to avoid confusion with too many numbers, we filtered out the detailed data including without mask (0), with mask (1), and wear mask incorrectly (2). The details are provided in [27].

3.2.1. Evaluate the YwFMMD models via various testsets

In this section, we use three testsets: FMMD, HITL-MMD, and DRFMD, as described in section 2, to evaluate YwFMMD. Table 5 presents the evaluation results of the YOLOv8 models (n, s, m, l, x) on the FMMD dataset with i) the testset extracted from the FMMD dataset itself (507 images, 2,663 instances), ii) the testset extracted from the DRFMD dataset (2,949 images, 6,191 instances), and iii) the testset extracted from the HITL-MMD dataset (1,311 images, 2,964 instances). Results in column (I) show that the models achieve very high performance on the dataset they were trained on (the FMMD dataset itself), with precision, recall, and mAP metrics all above average. This indicates that the models are well-trained and capable of good recognition on the seen dataset. Results in columns (II) and (III) show that the performance of the YOLOv8 models with the FMMD dataset decreases, indicating that the generalization capability of the FMMD dataset is not diverse enough, leading to underfitting.



(a)

(b)

(c)

(d)

(e)

Figure 2. Training performance using (a) YOLOv8n, (b) YOLOv8s, (c) YOLOv8m, (d) YOLOv8l, and (e) YOLOv8x with DRFMD dataset

Table 5. Testing result of YwFMMD models on FMMD, DRFMD, and HITL-MMD testsets

Model	FMMD Testset (I)			DRFMD Testset (II)			HITL-MMD Testset (III)		
	P	R	mAP@50	P	R	mAP@50	P	R	mAP@50
YOLOv8n	0.904	0.818	0.876	0.702	0.577	0.604	0.772	0.596	0.661
YOLOv8s	0.927	0.873	0.920	0.712	0.595	0.615	0.747	0.626	0.684
YOLOv8m	0.968	0.883	0.937	0.720	0.632	0.655	0.715	0.653	0.682
YOLOv8l	0.973	0.922	0.967	0.719	0.634	0.658	0.710	0.635	0.683
YOLOv8x	0.966	0.917	0.948	0.691	0.633	0.639	0.729	0.650	0.675

Based on Table 5, we see that when training and testing the YwFMMD model on the FMMD dataset itself, the results are good in most versions of the YwFMMD model. However, when evaluating on the new datasets DRFMD and HITL-MMD, the Precision, Recall, and mAP@50 metrics all decline significantly. Specifically, for YOLOv8m, the precision when training and testing with the FMMD, DRFMD, and HITL-MMD datasets are as follows: 0.966, 0.968 (↑0.02), 0.720 (↓0.246), and 0.715 (↓0.251). For recall, the results are 0.879, 0.883 (↑0.04), 0.632 (↓0.247), and 0.653 (↓0.226). For mAP@50, the results are 0.948, 0.937 (↓0.043), 0.655 (↓0.293), and 0.682 (↓0.266). This indicates that training on the FMMD dataset results in an overfitted model.

In terms of dataset size, DRFMD (14,727 images and 29,846 instances) is much larger than FMMD (2,030 images and 10,615 instances), as shown in Table 3. In terms of instance distribution, DRFMD is relatively more balanced compared to FMMD. Notably, the proportion of improperly worn masks in DRFMD is 9.6% (2,866 instances) compared to 3.2% (345 instances) in FMMD. Due to this, when testing with the testsets of DRFMD and HITL-MMD, the results significantly decline, demonstrating that the FMMD dataset is quite limited in the number of instances, especially for improperly worn masks.

3.2.2. Evaluate the YwDRFMD models via various testsets

In this section, we use three testsets: FMMD, HITL-MMD, and DRFMD as described in section 2 to evaluate YwDRFMD to see the contribution of the proposed DRFMD dataset. Table 6 presents the evaluation results of YOLOv8 models (n, s, m, l, x) on the DRFMD dataset with i) the testset extracted from the DRFMD dataset itself (2,949 images, 6,191 instances), ii) the testset extracted from the FMMD dataset (507 images, 2,663 instances), and iii) the testset extracted from the HITL-MMD dataset (1,311 images, 2,964 instances). Results in column (I) show that the YOLOv8 models trained on the DRFMD dataset perform well on the testset of this dataset itself (DRFMD). Easy to see that, the precision and recall metrics are high for all versions of YOLOv8. Especially, the larger models like YOLOv8l and YOLOv8x have the highest mAP@50 (0.867 and 0.560 respectively). This demonstrates that the DRFMD dataset provides a solid foundation for training recognition models with high accuracy and recognition capability. Results in columns (II) and (III) show that the YwDRFMD model, when tested on other datasets (FMMD and HITL-MMD), has reduced performance compared to when tested on the DRFMD dataset.

However, the precision and recall metrics remain relatively high, ranging from 0.771 to 0.830 and 0.778 to 0.795 for precision, and from 0.686 to 0.758 and 0.646 to 0.716 for recall on the FMMD and HITL-MMD testsets, respectively. Notably, mAP@50 maintains acceptable values, with 0.737 to 0.800 for FMMD and 0.705 to 0.779 for HITL-MMD. These results suggest that the model trained on DRFMD demonstrates strong generalization capability, allowing it to perform effectively on other datasets despite not achieving the highest performance. Overall, the YOLOv8 model trained on the DRFMD dataset (YwDRFMD) demonstrates good recognition and classification capabilities on this dataset itself, while also showing the ability to generalize and apply to other datasets with good performance. This affirms the diversity and robustness of the DRFMD dataset, making a significant contribution to improving the performance of object recognition models. Based on Table 6, we observe that when training and testing the YwDRFMD model on the DRFMD dataset itself, it yields good results across most versions of the YwDRFMD model. We also evaluated the model on other datasets such as FMMD and HITL-MMD, where the parameters Precision, Recall, and mAP@50 showed a slight decrease. Specifically, for YOLOv8m, the precision when trained and tested with the DRFMD, FMMD, and HITL-MMD datasets are 0.862, 0.858 (↓0.004), 0.807 (↓0.055), and 0.795 (↓0.067), respectively; for recall, the results are 0.823, 0.800 (↓0.023), 0.750 (↓0.073), and 0.716 (↓0.107), respectively; for mAP@50, the results are 0.888, 0.856 (↓0.032), 0.783 (↓0.105), and 0.779 (↓0.109), respectively. This indicates that training on the DRFMD dataset results in a model with high generalization capabilities.

Including the HITL-MMD dataset to evaluate the accuracy of the models (YwFMMD and YwDRFMD) is aimed at testing the generalization and objectivity of these models. Specifically, HITL-MMD includes cases of improperly worn masks, partially covered faces, or other obstructions, providing a more comprehensive assessment of the model's capabilities. While datasets like FMMD and DRFMD offer a solid foundation for model training and testing, the diversity and complexity of HITL-MMD highlight aspects that other datasets might miss. The absence of HITL-MMD could result in the model being limited to the seen data scope and reduce its practical application. Figure 3 shows the evaluation results between the YwFMMD and YwDRFMD models on the HITL-MMD testset. It is clear that the YwDRFMD model, trained on the DRFMD dataset, performs significantly better than the YwFMMD model, which was trained on the FMMD dataset. This demonstrates that the YOLOv8 models trained on the DRFMD dataset have better generalization capabilities. This confirms the diversity and robustness of the proposed DRFMD dataset.

Table 6. Testing result of YwDRFMD models on DRFMD, FMMD, and HITL-MMD testsets

Model	DRFMD Testset (I)			FMMD Testset (II)			HITL-MMD Testset (III)		
	P	R	mAP@50	P	R	mAP@50	P	R	mAP@50
YOLOv8n	0.840	0.756	0.814	0.810	0.686	0.737	0.778	0.646	0.705
YOLOv8s	0.839	0.797	0.844	0.814	0.715	0.772	0.778	0.698	0.731
YOLOv8m	0.858	0.800	0.856	0.807	0.750	0.783	0.795	0.716	0.779
YOLOv8l	0.866	0.811	0.867	0.771	0.758	0.776	0.787	0.712	0.769
YOLOv8x	0.860	0.827	0.874	0.830	0.753	0.800	0.786	0.715	0.770

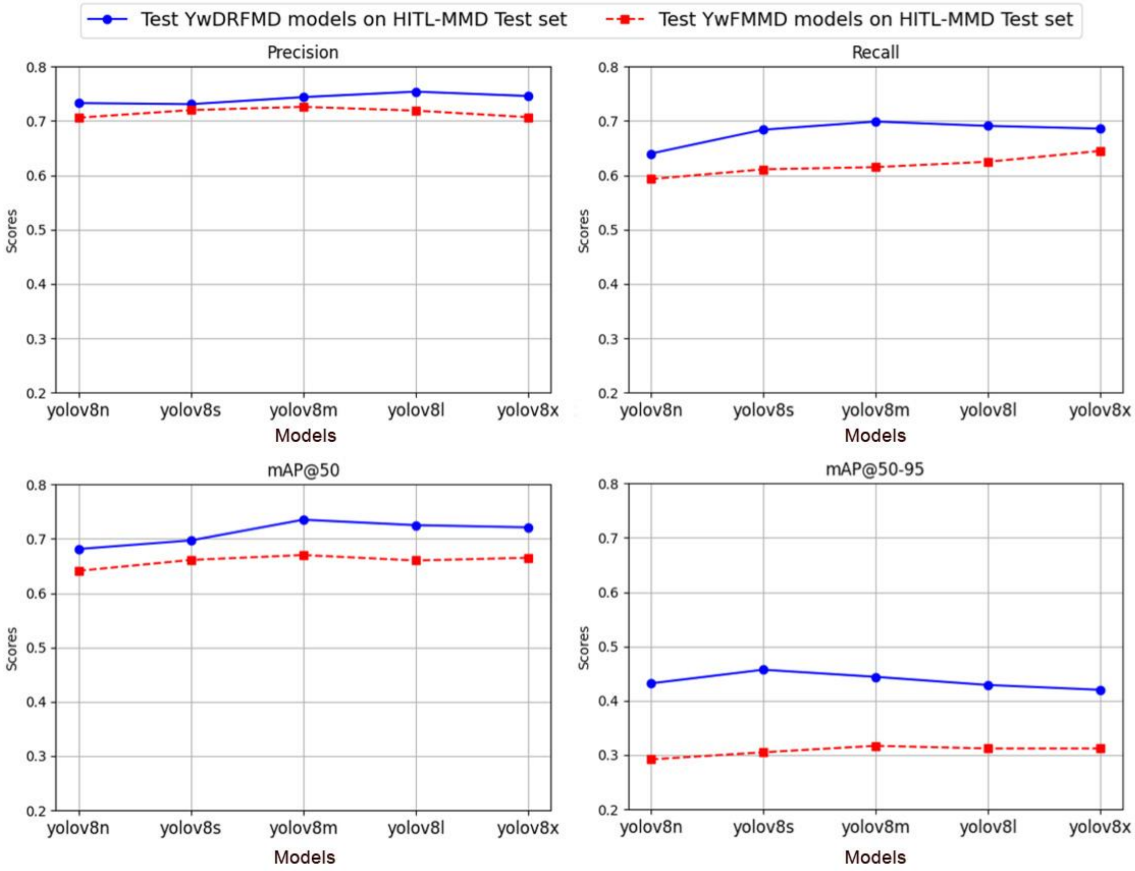


Figure 3. Comparing the performance of YwDRFMD vs YwFMMD models on the same testset HITL-MMD

3.2.3. Discussion on the effectiveness of the YwDRFMD compared to other studies

In this discussion, we select two studies [5], [6], to compare with our proposed approach. Since the datasets used in these studies are different, comparing accuracy metrics or evaluation measures such as precision, recall, F1 score, or mAP@ is not appropriate. Therefore, even though the models proposed in [5], [6] achieve significantly higher accuracy than our solution, the datasets used in these proposals have notable differences.

According to Joodi *et al.* [5], the MAFA dataset primarily focuses on faces with masks, which may limit the diversity of scenarios and contexts in which faces are captured. This could affect the model’s ability to generalize to real-world situations where lighting conditions, angles, and backgrounds vary significantly. Although the MAFA dataset contains a substantial number of images (35,806 MAFA), selecting 5,902 images for face analysis may not encompass all possible variations in masks, such as different types, colors, and styles. This could limit the model's robustness in detecting masks under diverse conditions. Additionally, the dataset selection for the study focuses on frontal faces, which may not fully represent the challenges of detecting masks on faces captured at different angles or in motion. This could be a limitation when applying the model in dynamic environments. With a specific focus on MAFA, there is a risk that models trained on the MAFA dataset may overfit to its characteristics, potentially reducing their effectiveness on other datasets or in real-world applications where faces are either unmasked or only partially covered. Meanwhile, the comprehensive dataset proposed in our study addresses these issues while maintaining high performance when evaluated on different datasets, as shown in the previous section (3.2.2).

Research by Hassan *et al.* [6], the dataset consists of 2,165 images of MAFA and 1,930 images of unmasked faces. While this may be sufficient for initial model training, it may not be large enough to capture the full variability of real-world scenarios, potentially affecting the model's generalization capability. The dataset is categorized into only two classes masked and unmasked faces. This binary classification does not account for partially worn masks or improperly used masks, which are common in real-world situations and may lead to misclassification. The images in the dataset were cropped to focus solely on the faces. While this simplifies the model's task, it may not reflect real-world conditions where faces are not always perfectly aligned or fully visible, potentially impacting the model's performance in practical applications. Additionally, with a relatively small dataset, the reported accuracy of 99% during training and 100% during testing raises concerns about overfitting, where the model learns the training data too well but fails to perform effectively on unseen data. This issue is particularly concerning if the dataset does not include a wide range of variations in facial appearances and mask types. Meanwhile, the DRFMD dataset we propose offers greater diversity in terms of demographics, lighting conditions, and multiple image angles. This is particularly useful for effectively detecting various cases of both proper and improper mask-wearing, ensuring a more robust and generalizable model for real-world applications.

4. CONCLUSION

This study focuses on proposing the DRFMD dataset and applying YOLOv8 models to improve mask recognition performance on faces across various input image types. The results obtained from training and testing on the DRFMD dataset show that YOLOv8 can accurately detect and classify cases of wearing masks, not wearing masks, and wearing masks improperly with high accuracy. Experiments demonstrate that the YOLOv8 model trained on DRFMD outperforms YOLOv8 models trained on other datasets like FMMD, proving its broad applicability in public health monitoring and disease prevention. This proposed dataset is compiled from reputable sources such as AIZOO, KFMD, MAFA, MOXA3K, and the Zalo AI challenge, ensuring greater diversity and generalization capability for the model. Additionally, using data augmentation techniques such as padding, cropping, and horizontal flipping has enhanced the model's performance, enabling it to better handle diverse real-world situations. This research significantly contributes to improving the effectiveness and accuracy of mask recognition systems, especially in the context of current public health issues, and opens up new directions for developing rich datasets and advanced deep learning techniques. Furthermore, based on the proposed DRFMD dataset, our future research aims to enhance the YOLOv8 model and subsequent YOLO versions to reduce training time and further improve accuracy.

ACKNOWLEDGMENTS

The authors extend their appreciation to the Posts and Telecommunications Institute of Technology (PTIT) in Vietnam for for financial supporting and facilitating this research.

FUNDING INFORMATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Trong Thua Huynh	✓	✓		✓	✓	✓			✓	✓		✓	✓	✓
Hoang Thanh Nguyen	✓	✓	✓	✓		✓	✓	✓	✓		✓			

C :	**C**onceptualization	I :	**I**nvestigation	Vi :	**Vi**sualization
M :	**M**ethodology	R :	**R**esources	Su :	**Su**pervision
So :	**So**ftware	D :	**D**ata Curation	P :	**P**roject administration
Va :	**Va**lidation	O :	Writing - **O**riginal Draft	Fu :	**Fu**nding acquisition
Fo :	**Fo**rmal analysis	E :	Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

Not applicable.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Kaggle at https://doi.org/10.34740/kaggle/dsv/12142420.

REFERENCES

[1] Y. Himeur, S. Al-Maadeed, I. Varlamis, N. Al-Maadeed, K. Abualsaud, and A. Mohamed, “Face mask detection in smart cities using deep and transfer learning: lessons learned from the covid-19 pandemic,” *Systems*, vol. 11, no. 2, 2023, doi: 10.3390/systems11020107.

[2] A. Sharma, R. Gautam, and J. Singh, “Deep learning for face mask detection: a survey,” *Multimedia Tools and Applications*, vol. 82, no. 22, pp. 34321–34361, 2023, doi: 10.1007/s11042-023-14686-6.

[3] W. Chen, L. Gao, X. Li, and W. Shen, “Lightweight convolutional neural network with knowledge distillation for cervical cells classification,” *Biomedical Signal Processing and Control*, vol. 71, 2022, doi: 10.1016/j.bspc.2021.103177.

[4] X. Fan, M. Jiang, and H. Yan, “A deep learning based light-weight face mask detector with residual context attention and gaussian heatmap to fight against covid-19,” *IEEE Access*, vol. 9, pp. 96964–96974, 2021, doi: 10.1109/ACCESS.2021.3095191.

[5] M. A. Joodi, M. H. Saleh, and D. J. Kadhim, “Increasing validation accuracy of a face mask detection by new deep learning model-based classification,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 1, pp. 304–314, 2023, doi: 10.11591/ijeecs.v29.i1.pp304-314.

[6] N. F. A. Hassan, A. A. Abed, and T. Y. Abdalla, “Face mask detection using deep learning on NVIDIA Jetson Nano,” *International Journal of Electrical and Computer Engineering*, vol. 12, no. 5, pp. 5427–5434, 2022, doi: 10.11591/ijece.v12i5.pp5427-5434.

[7] J. Khoramdel, S. Hatami, and M. Sadedel, “Wearing face mask detection using deep learning during COVID-19 pandemic,” *Scientia Iranica*, vol. 30, no. 3, pp. 1058–1067, 2023, doi: 10.24200/sci.2023.59093.6057.

[8] H. Al-Dmour, A. Tareef, A. M. Alkalbani, A. Hammouri, and B. Alrahmani, “Masked face detection and recognition system based on deep learning algorithms,” *Journal of Advances in Information Technology*, vol. 14, no. 2, pp. 224–232, 2023, doi: 10.12720/jait.14.2.224-232.

[9] B. Qin *et al.*, “Lightweight DB-YOLO facemask intelligent detection and android application based on bidirectional weighted feature fusion,” *Electronics*, vol. 12, no. 24, 2023, doi: 10.3390/electronics12244936.

[10] N. Aburaed, M. Alsaad, S. Al Mansoori, and H. Al-Ahmad, “A study on the autonomous detection of impact craters,” in *Artificial Neural Networks in Pattern Recognition*, Cham, Switzerland: Springer, 2023, pp. 181–194. doi: 10.1007/978-3-031-20650-4_15.

[11] G. Jocher, A. Chaurasia, and J. Qiu, “Ultralytics YOLOv8,” *Ultralytics*. Accessed: Jul. 18, 2024. [Online]. Available: https://docs.ultralytics.com/models/yolov8.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: unified, real-time object detection,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[13] J. Redmon and A. Farhadi, “YOLOv3: an incremental improvement,” *arXiv-Computer Science*, pp. 1–6, 2018.

[14] C. Dewi, D. Manongga, Hendry, E. Mailoa, and K. D. Hartomo, “Deep learning and YOLOv8 utilized in an accurate face mask detection system,” *Big Data and Cognitive Computing*, vol. 8, no. 1, 2024, doi: 10.3390/bdcc8010009.

[15] A. Maranhão, “Face mask detection,” *Kaggle,* 2020. Accessed: Jun. 09, 2024. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/face-mask-detection

[16] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, “A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic,” *Measurement*, vol. 167, Jan. 2021, doi: 10.1016/j.measurement.2020.108288.

[17] "Medical mask dataset," *Human in The Loop*. Accessed: Jun. 10, 2024. [Online]. Available: https://humansintheloop.org/resources/datasets/medical-mask-dataset/.

[18] T. Lin, "LabelImg," *GitHub*. Accessed: Apr. 29, 2024. [Online]. Available: https://github.com/csq20081052/labelImg

[19] AIZOOTech, "Face mask detection," *GitHub*. Accessed: May 15, 2024. [Online]. Available: https://github.com/AIZOOTech/FaceMaskDetection/tree/master.

[20] K. Malik, "Face mask detector," *GitHub*. Accessed: Jun. 04, 2024. [Online]. Available: https://github.com/Karan-Malik/FaceMaskDetector/tree/master.

[21] B. Roy, S. Nandy, D. Ghosh, D. Dutta, P. Biswas, and T. Das, "MOXA: A deep learning based unmanned approach for real-time monitoring of people wearing medical masks," *Transactions of the Indian National Academy of Engineering*, vol. 5, no. 3, pp. 509–518, 2020, doi: 10.1007/s41403-020-00157-z.

[22] S. Ge, J. Li, Q. Ye, and Z. Luo, "Detecting masked faces in the wild with LLE-CNNs," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017, pp. 426–434. doi: 10.1109/CVPR.2017.53.

[23] X. Jiang, T. Gao, Z. Zhu, and Y. Zhao, "Real-time face mask detection method based on YOLOv3," *Electronics*, vol. 10, no. 7, 2021, doi: 10.3390/electronics10070837.

[24] S. Thai, H. Bui, H. Vu, K. Nguyen, T. Huynh, and K. Hoang, "Enhancing face anti-spoofing with swin transformer-driven multi-stage pipeline," in *Proceedings of the 12th International Symposium on Information and Communication Technology*, New York, United States: ACM, 2023, pp. 40–47. doi: 10.1145/3628797.3628948.

[25] B. Wang, J. Zheng, and C. L. P. Chen, "A survey on masked facial detection methods and datasets for fighting against covid-19," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 3, pp. 323–343, 2022, doi: 10.1109/TAI.2021.3139058.

[26] H. Kang and C. Chen, "Fast implementation of real-time fruit detection in apple orchards using deep learning," *Computers and Electronics in Agriculture*, vol. 168, Jan. 2020, doi: 10.1016/j.compag.2019.105108.

[27] T. T. Huynh and H. T. Nguyen, "Training and testing results on YOLOv8 (n, s, m, l, x) with FMMD and DRFMD datasets," GitHub. Accessed: Jul. 31, 2024. [Online]. Available: https://github.com/hthanhsg/Enhancing-face-mask-detection-performance-with-comprehensive-dataset-and-YOLOv8.

BIOGRAPHIES OF AUTHORS

Trong Thua Huynh is currently the head in the Department of Information Security, Faculty of Information Technology 2, at the Posts and Telecommunications Institute of Technology (PTIT), Vietnam. He received a bachelor's degree in Information Technology from Ho Chi Minh City University of Natural Sciences, a master's degree in Computer Engineering from Kyung Hee University, Korea, and a Ph.D. degree in Computer Science from the Ho Chi Minh City University of Technology, Vietnam National University at Ho Chi Minh City. His key areas of research include cybersecurity, AI and big data, and intelligent information systems. He can be contacted at email: thuaht@ptit.edu.vn.

Hoang Thanh Nguyen is currently the lecturer in Ho Chi Minh City, Vietnam. He received a master's degree in Information Systems from the Institute of Post and Telecommunications Technology. His research areas are information security and machine learning. He can be contacted at email: thanhnh@ptit.edu.vn.