International Journal of
Advanced Science Computing and Engineering

Journal Homepage www.ijasce.org/index.php/ijasce

Building Enthusiasm Level Detection Model on Online Learning Using YOLOv11 with Hyperparameter Optimization

Hilmi Aziz [a,*], Siti Yulianti [a], Rianto [a]

[a] *Informatics, Faculty of Engineering, Siliwangi University, Tawang, Tasikmalaya, Indonesia*

*Corresponding author: *hilmiaziz4212@gmail.com*

***Abstract*—This study aims to develop a model for detecting enthusiasm levels in online learning using the YOLOv11 algorithm, enhanced through hyperparameter optimization. Facial expressions serve as crucial indicators in determining enthusiasm, as they reflect the level of attention and interest a learner has toward the material. By increasing the number of interest level categories, the model is expected to provide a more detailed and accurate assessment of student engagement. The dataset used in this research is sourced from FER2013, which initially consists of seven emotion classes. These emotions are reorganized and classified into five enthusiasm levels to represent different levels of interest in learning better. Each level contains 1,000 images, resulting in a dataset of 5,000 images. This dataset was refined from previous studies to enhance its relevance and improve detection performance, making it more suitable for real-world applications. To achieve optimal performance, key hyperparameters, including the number of epochs, batch size, and image size, were fine-tuned. Before optimization, the model demonstrated an average precision (mAP 50-95) of 95.2% with an inference time of 1.7 milliseconds. After hyperparameter tuning, the model's performance improved significantly, reaching an average precision (mAP 50-95) of 97%. However, this enhancement came with a slight increase in inference time to 3.1 milliseconds. The results highlight that fine-tuning model parameters can enhance detection accuracy while maintaining efficient processing speed, making it highly applicable in educational settings for assessing learner engagement.**

***Keywords*—Detection; enthusiasm; hyperparameter optimization; YOLOv11.**

Manuscript received 3 Dec. 2024; revised 9 Mar. 2025; accepted 6 Apr. 2025. Date of publication 30 Apr. 2025.

I. Introduction

Enthusiasm is described as a positive emotional state arising from feelings of enjoyment and interest [1]. Emotional expressions are direct expressions of internal states, implying that they are involuntary, uncontrollable, and essentially honest. [2]. Knowledge of the level of enthusiasm/interest can be obtained from emotions that function as automatic human responses that appear on the face [3]. Enthusiasm in online learning is evident in learners' interest in engaging with the material presented on the screen, such as by facing the camera. Online education makes quite a lot of students experience learning difficulties, especially since students do not feel the presence of social interaction, but students still have to try to hold their attention to the teacher [4]. The solution is for teachers to be able to know the level of interest of students during learning, to carry out effective and enjoyable learning.

Nowadays, deep learning has shown its ability to recognize and learn complex patterns in detecting various objects, both living and non-living. Deep Learning is a subset of machine learning that involves algorithms that use a deep, hierarchically structured set of non-linear transformation functions to model high-level abstractions of data [5]. Many deep learning algorithms have been used in the expression detection process, among which the convolutional neural network (CNN) algorithm is quite popular. CNN algorithms have proven successful in detecting emotions from humas's expressions with the highest validation accuracy up to 98.65% [6]. Another deep learning algorithm that is widely used in the detection of various objects is YOLO. YOLO algorithm has proven to be very good in detecting multiple types of objects, such as human activities, very quickly [7]. In addition, the newest version of YOLO, YOLOv11, is also used in early Diagnoses of Acute Lymphoblastic Leukemia [8]. YOLOv11 also achieves the fastest inference time on fruit detection with only 2.4 ms, although the best performance was achieved by YOLOv9 gelan-base and YOLOv9 gelan-e with a score of 93.5% in the same research [9].

There is one of the efforts to obtain optimal performance in the Convolutional Neural Network (CNN) like YOLO model by hyperparameter optimization involving epoch adjustment, batch size, and learning rate as has been done in the study 3D printer error detection research using the YOLOv8 algorithm to find out the best configuration for the model to find improvements and different results from each configuration [10]. Based on previous related research, there is an opportunity to create a faster enthusiasm level detection model using YOLOv11 with hyperparameter optimization to achieve more accurate performance results. This research aims to develop an enthusiasm detection model that recognizes the level of enthusiasm in online learning. This will help teachers monitor and acknowledge their students' interests more quickly, enabling them to respond and adjust to students' needs more easily.

II. Materials and Method

A. Related Works

Several studies have built human expression detection models using CNN and YOLO algorithms. Research [11] using the FER2013 dataset in human emotion detection with the CNN algorithm achieved a fairly good accuracy rate at 73,8%. Meanwhile, with the same dataset, FER2013, the research [12] grouped the dataset into two enthusiasm categories in building an enthusiasm level detection model with YOLOv8, achieving very good accuracy at 95.3% with an inference time of 62 ms. This research demonstrates that the YOLOv8 algorithm's excellent performance in human expression detection, combined with its real-time detection features, renders it a superior detection model.

Related research was also conducted in [13] the classification of interest levels of kindergarten children using CNN with three classes of interest levels from a dataset of 243 images. The model achieved its highest accuracy of 81.6%. The accuracy of this CNN model is lower when compared to the YOLOv8 model. The latest YOLO algorithm was also used in a study by [8] early diagnosis of acute lymphoblastic leukemia by comparing the performance of YOLOv8 and YOLOv11 with a dataset of 3,256 images. The results show that YOLOv11s is superior with an accuracy of 98.8%.

YOLOv11 brings improvements to the architecture and detection capabilities. It combines a convolutional backbone and Feature Pyramid Network (FPN) to support better multi-scale detection. YOLOv11 also proved to be faster than the previous generation in a study comparing the performance of YOLOv8, YOLOv9, YOLOv10, and YOLOv11 in fruit detection [9]. Although the best performance was achieved by YOLOv9 gelan-base and YOLOv9 gelan-e with a score of 93.5%, the fastest inference time was achieved by YOLOv11n with only 2.4 ms.

B. Methodology

This research methodology is designed to build an enthusiasm level detection model using the Yolov11 algorithm with hyperparameter optimization. The steps of this research are outlined in Figure 1.

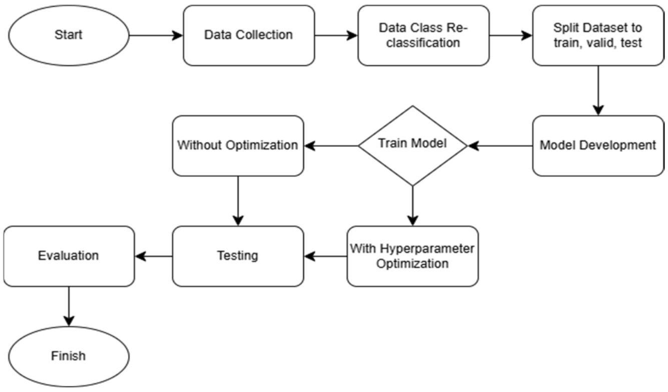


Fig. 1 Research Stages

1) Step 1-Data Collection: The dataset used in this research is FER 2013, which can be obtained from the Kaggle site. This dataset comprises 35,887 digital image data points, categorized into seven classes of human facial expressions: Anger, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

2) Step 2-Data Class Re-classification: The dataset containing seven emotion classes was re-classified into five enthusiasm level classes named "Highly Interested", "Interested", "Quite Interested", "Less Interested", and "Not Interested" with each new dataset class containing 1000 image data [9]. Data class re-classification was carried out with the help of a lecturer and fellow students, and is in line with related research that classifies the level of interest of kindergarten children to avoid subjective preferences of a person. The result from the re-classification of dataset classes is shown in Figure 2.

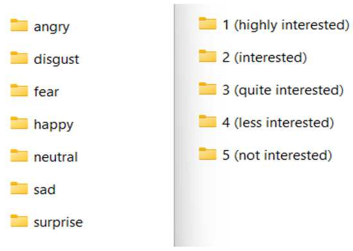
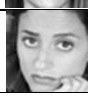
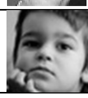


Fig. 2 Re-classification of Dataset Class

Those new classes are described in Table 1 along with sample data of each class.

TABLE I
THE DATASET: NEW CLASSES OF ENTHUSIASM (INTEREST LEVEL)

Class Name	Samples		Description
Highly Interested			Showed a fond expression with a gaze facing the screen
Interested			Showed a neutral expression with the gaze facing the screen
Quite Interested			Showed a disliked expression, but gaze still facing the screen
Less Interested			Showed an expression of dislike with the face facing forward, but the eyes looking the other way
Not Interested			Showed a displeased expression with a face that was not even facing the screen

After the dataset was re-classified, the new dataset was labeled with a class order that is adjusted to the order of class names in Table 1. The labeling process is done by creating a txt extension file for each image in the dataset, which contains class information and bounding boxes [14].

3) Step 3-Split Dataset: The dataset totaling 5000 image data is divided into train data, valid data, and test data with a ratio of 70:15:15 [15]. The training data is used to learn the pattern of the object to be detected. Validation data is used during model training to evaluate the performance of the model on unseen data during each training epoch. This test data provides an objective assessment of the model's ability to detect objects on new and previously unassessed data.

4) *Step 4-Model Development:* The model that will be used in this research is YOLOv11. The YOLOv11 architecture consists of three main components: backbone, neck, and head. The backbone, which typically consists of a convolutional neural network, serves as the central feature extractor, transforming raw image data into a multi-scale feature map. The neck then processes this feature map with layers designed to combine and enhance feature representations across multiple scales. Finally, the head generates the final prediction for object location and classification based on the processed feature map. Based on this foundation, YOLO11 introduces architectural enhancements and parameter optimizations, improving detection performance and accuracy over previous versions [16]. The architecture of this enthusiasm detection model with YOLOv11 can be seen in Figure 3.

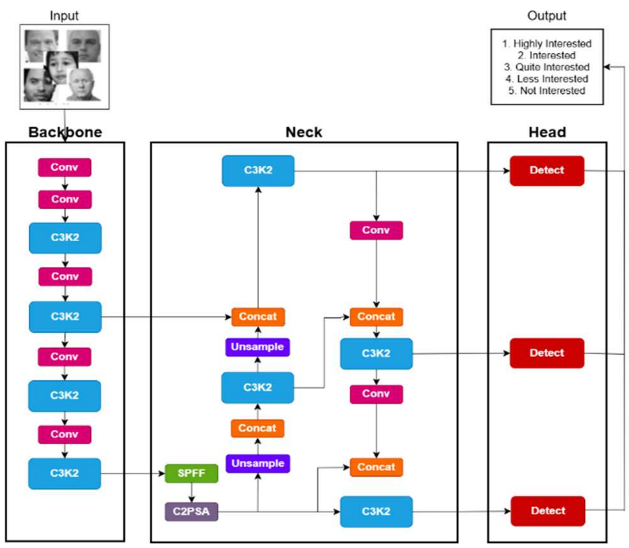


Fig. 3 The Architecture of the Detection Model with YOLOv11

The variation of YOLOv11 used in this research is the nano variation. This variation mode was chosen because it is a model that has the fastest inference time, as evidenced in the study, reaching a time of 2.4 ms.

5) Step 5-Train Model: The model to be trained in this research is divided into two parts. The first model is the model before hyperparameter optimization. The model uses the same hyperparameter configuration settings as the YOLOv11 hyperparameter configuration in the research [9] as shown in Table 2.

TABLE II
HYPERPARAMETER CONFIGURATION FOR 1ST MODEL

Hyperparameter	Value
Initial Learning Rate (lr0)	0.01
Final Learning Rate (lrf)	0.01
Momentum	0.937
Weight Decay	0.0005
Warmup Epochs	3.0
Box Loss Gain (box)	7.5
Class Loss Gain (cls)	0.5
Definition Loss Gain (dfl)	1.5

The second model is a model with hyperparameter optimization. The model is the best-performing model of several models trained with different hyperparameter configurations, following what was done in research [10]. The hyperparameter configuration to be used is shown in Table 3:

TABLE III
HYPERPARAMETER CONFIGURATION FOR OPTIMIZATION)

Hyper-parameter	Explanation	Influence	Value
Image size	The dataset image size determines how much information the model can obtain.	The larger the image size, the more information the model can obtain, but it also increases computational cost.	16, 32
Batch Size	The number of samples processed before the model updates its weights.	A larger batch size makes the training process more stable and efficient, but requires more memory.	48x48, 360x360
Epoch	The number of times the entire dataset is passed through the training algorithm.	A higher number of epochs can generally improve model accuracy, but too many epochs may lead to overfitting.	50, 100, 200

6) Step 6-Testing: Model testing is done using 750 test data, in contrast to the training process, which uses valid data as test data to get an assessment of its performance when detecting images it has never seen.

7) Step 7-Evaluation: Model evaluation is a critical stage after testing to check the performance and object detection capabilities of the trained model. This model evaluation can be done using Mean Average Precision (mAP), a metric that measures the accuracy of the model in detecting and recognizing objects at various levels of precision. Calculating mAP can be calculated as described in equations (1), (2), (3), and (4) [17].

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{1}$$

Precision is the ratio of the value of true positive predictions to the total results with positive predictions.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2}$$

Recall is the ratio of true positive predicted values to all true positive data that are true positive.

$$Average\ Precision\ (\mathrm{AP}) = \sum_n (\mathrm{R_n} - \mathrm{R_{n-1}} \times \mathrm{P_n}) \qquad (3)$$

AP is a measure that describes the Precision-Recall curve (precision plotted against recall) in a single number, or the area below it.

$$Mean\ Average\ Precision\ (\mathrm{mAP}) = \frac{1}{n} \sum_{i=1}^{n} AP_i \qquad (4)$$

mAP is the average of APs for all classes in the dataset.

Model evaluation is also done with the Confusion Matrix table, which evaluates the performance of classification models in machine learning by comparing the model's predictions with the actual data to help understand where the model went wrong.

III. Results and Discussion

This section explains how the detection model compares before and after hyperparameter optimization. There are 2 accuracies of this model, namely training accuracy (using valid data during the model training process) and testing accuracy (using test data during the model testing process). The training accuracy of 1st model received an average score of precision value (mAP50-95) with a score of 95.2% and an inference time of 1.7 ms. The Confusion Matrix of the training 1st model is shown in Figure 4.

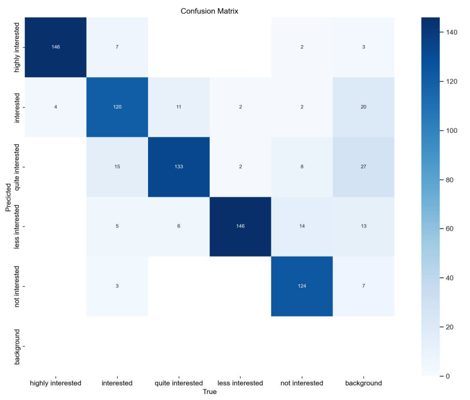


Fig. 4 Confusion Matrix of Training Accuracy for 1st Model

Meanwhile, the testing accuracy of 1st model received an average precision value (mAP50-95) with a score of 94.7% with an inference time of 3.2 ms. The performance of the 1st model in testing is slightly lower than in the training process, with a difference of 0.5%. The Confusion Matrix of 1st model for testing results is shown in Figure 5.

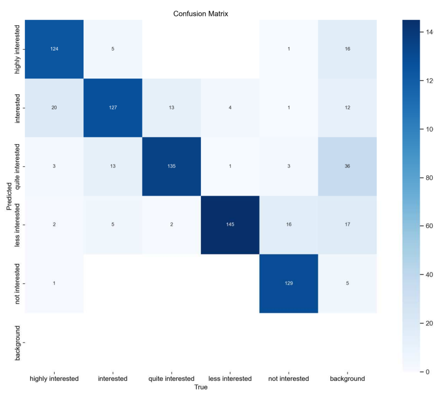


Fig. 5 Confusion Matrix of Testing Accuracy for 1st Model

Furthermore, the training accuracy and testing accuracy for the first model with hyperparameter optimization are shown in Table 4. Then the confusion matrix of the second model with the best performance for both training and testing accuracy is shown in Figures 6 and 7.

TABLE IV
TRAINING ACCURACY AND TESTING ACCURACY OF MODEL B)

Hyperparameter			Training Accuracy		Testing Accuracy	
Image size	**Batch Size**	**Epoch**	**mAP50-95**	**Inference Time**	**mAP50-95**	**Inference Time**
48x48	16	50	75%	0.5 ms	73,3%	0.5 ms
		100	80,2%	0.4 ms	75,3%	0.5 ms
		200	65,3%	1 ms	63%	0.4 ms
	32	50	76,2%	0.3 ms	72,1%	0.4 ms
		100	78,6%	0.3ms	74,4%	0.3 ms
		200	81,6%	0.7 ms	80,4%	0.2 ms
360x360	16	50	92,6%	1.5 ms	91,6%	2.7 ms
		100	95,2%	1.5 ms	94,7%	2.7 ms
		200	95,7%	3.5 ms	95,7%	2.7 ms
	32	50	93,5%	1.5 ms	93,2%	2.7 ms
		100	95,6%	1.5 ms	95%	2.8 ms
		200	**97%**	3.1 ms	96, 7%	2.7 ms

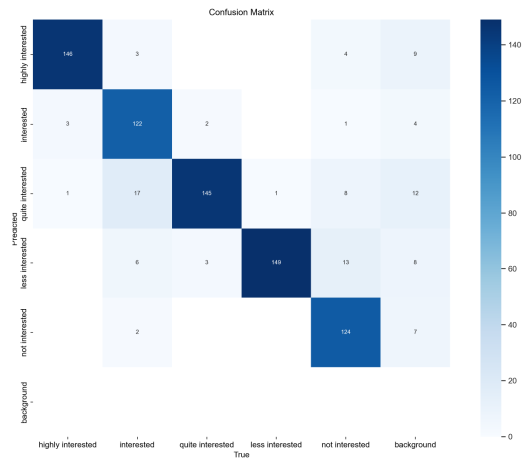


Fig. 1 Confusion Matrix of Training Accuracy for 2nd Model

The best performance for both training and testing accuracy was achieved when model b) used an image size of 360x360 pixels, batch size of 32, and 200 epochs with an average precision score (mAP 50-95) of 97% with an inference time of 3.1 ms for training accuracy, and 96.7% with an inference time of 2.7 ms for testing accuracy.

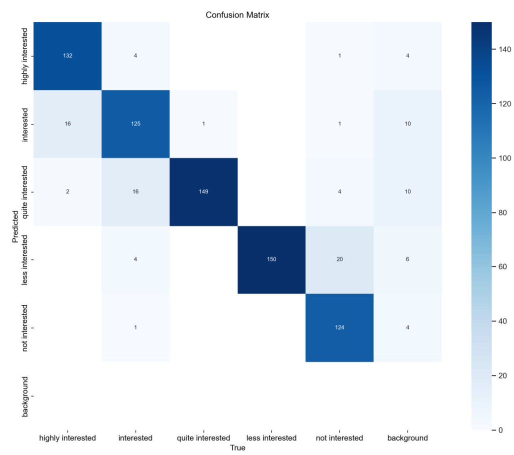


Fig. 2 Confusion Matrix of Testing Accuracy for 2nd Model

The configuration of the three hyperparameters significantly affects the performance of the model. A larger image size has a significant impact on the model's performance. The performance of the model with a 360x360 pixel dataset far outperforms the model with a 48x48 pixel dataset. Batch size is also quite influential on model performance, although increasing batch size from 16 to 32 only slightly improves performance with the same image size and epoch. The larger epoch also affects the performance of the model, although not too much, as seen in Figures 8 and 9.

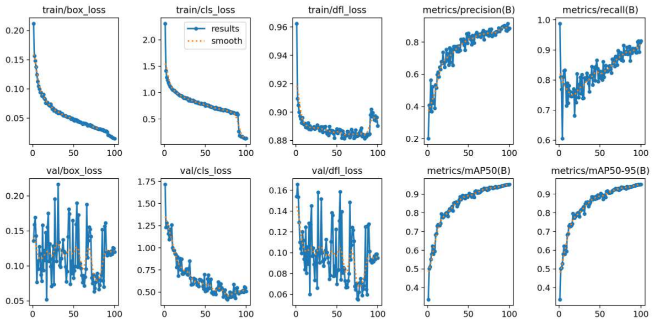


Fig. 3 Performance Graph for 1st model Before Optimization

Figure 8 shows the performance of 1st model before optimization. As the epochs increase, the loss or model error graph becomes smaller. The average precision of the model also increases as the epoch increases. Figure 9 shows the slight difference between the optimized 2nd model and the 1st model before optimization. The most striking difference is in the validation loss graph, which is more stable and decreases as the epoch increases.

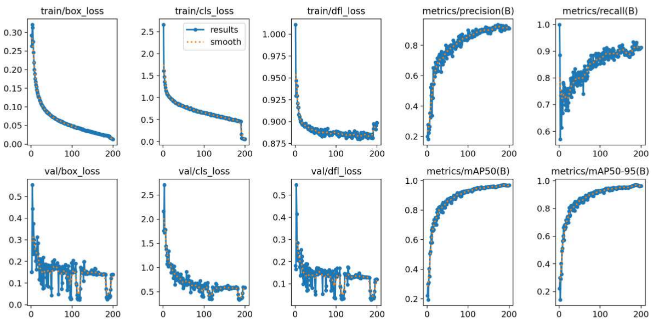


Fig. 4 Performance Graph for 2nd Model with Optimization

IV. Conclusion

The YOLOv11n detection model proved to have an excellent performance level in both accuracy and inference speed at all hyperparameter settings, both before and after optimization. Before optimization, the model achieved an average precision (mAP50-95) of 95.2% and an inference time of 1.7 ms. After optimization with a configuration of 360 pixels, a batch size of 32, and 200 epochs, the model performance increased to 97% average precision and inference time at 3.1 ms.

The weakness of the developed model lies in the automatic annotation process, where the entire image in the dataset is enclosed within a bounding box. As a result, the model may learns excessive patterns from the dataset. Further research is expected to perform manual annotation by creating bounding boxes only around the facial area in the dataset.

References

[1] I. Burić and A. Moè, "What makes teachers enthusiastic: The interplay of positive affect, self-efficacy and job satisfaction," *Teach. Teach. Educ.*, vol. 89, p. 103008, Mar. 2020, doi: 10.1016/j.tate.2019.103008.

[2] M. E. Kret *et al.*, "Emotional expressions in human and non-human great apes," *Neurosci. Biobehav. Rev.*, vol. 115, pp. 378–395, Aug. 2020, doi: 10.1016/j.neubiorev.2020.01.027.

[3] F. F. Rahmawati, D. Setiawan, and M. Roysa, "Penyebab kesulitan belajar siswa pada pembelajaran daring," *J. Lesson Learn. Stud.*, vol. 4, no. 3, pp. 302–308, Dec. 2021, doi: 10.23887/jlls.v4i3.32506.

[4] C. Conrad *et al.*, "How student perceptions about online learning difficulty influenced their satisfaction during Canada's Covid-19 response," *Br. J. Educ. Technol.*, vol. 53, no. 3, pp. 534–557, Feb. 2022, doi: 10.1111/bjet.13206.

[5] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electron. Mark.*, vol. 31, no. 3, pp. 685–695, Apr. 2021, doi:10.1007/s12525-021-00475-2.

[6] A. Jaiswal, A. K. Raju, and S. Deb, "Facial emotion detection using deep learning," in *Proc. Int. Conf. Emerg. Technol. (INCET)*, Jun. 2020, doi: 10.1109/incet49848.2020.9154121.

[7] N. P. Motwani and S. S, "Human activities detection using deep learning technique—YOLOv8," *ITM Web Conf.*, vol. 56, p. 03003, 2023, doi: 10.1051/itmconf/20235603003.

[8] A. Awad and S. A. Aly, "Early diagnosis of acute lymphoblastic leukemia using YOLOv8 and YOLOv11 deep learning models," in *Proc. 12th Int. Japan-Africa Conf. Electron., Commun., Comput. (JAC-ECC)*, Dec. 2024, pp. 11–16, doi: 10.1109/jac-ecc64419.2024.11061246.

[9] R. Sapkota, Z. Meng, M. Churuvija, X. Du, Z. Ma, and M. Karkee, "Comprehensive performance evaluation of YOLO11, YOLOv10, YOLOv9 and YOLOv8 on detecting and counting fruitlet in complex orchard environments," *Qeios*, Oct. 2024, doi: 10.32388/e9y7xi.

[10] N. B. A. Karna *et al.*, "Toward accurate fused deposition modeling 3D printer fault detection using improved YOLOv8 with hyperparameter optimization," *IEEE Access*, vol. 11, pp. 74251–74262, 2023, doi:10.1109/access.2023.3293056.

[11] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State of the art performance on FER2013," *arXiv preprint arXiv:2105.03588*, 2021. [Online]. Available: https://arxiv.org/abs/2105.03588.

[12] K. Salma and S. Hidayat, "Deteksi antusiasme siswa dengan algoritma YOLOv8 pada proses pembelajaran daring," *J. Indones.: Manaj. Inf. Komun.*, vol. 5, no. 2, pp. 1611–1618, May 2024, doi:10.35870/jimik.v5i2.716.

[13] A. R. Kusumastuti, Y. Kristian, and E. Setyati, "Klasifikasi ketertarikan anak PAUD melalui ekspresi wajah menggunakan metode CNN," *J. Teknol. Inf. Terapan*, vol. 7, no. 2, pp. 92–96, Dec. 2020, doi: 10.25047/jtit.v7i2.176.

[14] X. Tao *et al.*, "Pavement crack detection and identification based on improved YOLOv8," *Int. J. Cogn. Inform. Nat. Intell.*, vol. 18, no. 1, pp. 1–20, Sep. 2024, doi: 10.4018/ijcini.356363.

[15] A. Serikbay *et al.*, "Ensemble pretrained convolutional neural networks for the classification of insulator surface conditions," *Energies*, vol. 17, no. 22, p. 5595, Nov. 2024, doi:10.3390/en17225595.

[16] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024. [Online]. Available: https://arxiv.org/abs/2410.17725.

[17] D. Krstinić *et al.*, "Multi-label classifier performance evaluation with confusion matrix," *Comput. Sci. Inf. Technol.*, Jun. 2020, doi:10.5121/csit.2020.100801.