Facial Expression Classification System Using Stacked CNN

Aditya Wikan Mahastama[1], Edwin Mahendra[2], Antonius Rachmat Chrismanto[3]*,
Maria Nila Anggia Rini[4], Andhika Galuh Prabawati[5]
Faculty of Information Technology, Universitas Kristen Duta Wacana, Yogyakarta, Indonesia

***Abstract*—Automatic emotion recognition technology through facial expressions has broad potential, ranging from human-computer interaction to stress detection and blood pressure assessment. Facial expressions exhibit patterns and characteristics that can be identified and analyzed by image processing and machine learning methods. These methods provide a basis for the development of emotion recognition systems. This research develops a facial emotion recognition model using Convolutional Neural Network (CNN) architecture, a popular architecture in image classification, segmentation, and object detection. CNNs offer automatic feature extraction and complex pattern recognition advantages on image data. This research uses three types of datasets, FER2013, CK+, and IMED, to optimize the deep learning approach. The developed model achieved an overall accuracy of 71% on the three datasets combined, with an average precision, recall, and F1-Score of 71%. The results show that CNN architecture performed well in facial emotion classification, supporting potential practical applications in various fields.**

***Keywords*—FER; CNN; deep learning; image classification**

I. Introduction

Artificial intelligence and machine learning technologies have created new possibilities in interpreting human emotions through facial expressions. Facial expressions are one of the natural ways humans communicate emotions. Automatic emotion recognition technology, mainly through facial expressions, has broad potential in various applications, ranging from human-computer interaction to stress detection and blood pressure assessment [1]. Facial expressions, as one of the non-verbal communication mediums, exhibit patterns and characteristics that can be identified and analyzed by image processing and machine learning methods. This provides a basis for developing systems to interpret facial expressions to recognize human emotions [2].

Convolutional Neural Network (CNN) is a deep learning network introduced in the 1960s. CNN is also applied in computer vision and is generally used in image classification, segmentation, object detection, and video processing [3]. Supporting the implementation of FER, deep learning-based technology with Convolutional Neural Networks (CNN) architecture was used by previous researchers as a potential solution to overcome problems in face and expression recognition and classification. Convolutional Neural Networks (CNN) are increasingly used in FER because they automatically extract features from images [4].

Facial Emotion Recognition refers to the ability to identify and recognize emotions expressed through the human face, belonging to an important research topic in computer vision and artificial intelligence. Facial emotions play an essential role in human communication, helping to understand the intentions and feelings of others, with two-thirds of human communication conveyed through nonverbal components, of which facial expressions play a major role. FER has two main approaches: the traditional approach and the Convolutional Neural Network (CNN) based approach. The conventional approach involves face component detection, feature extraction, and expression classification. The initial stage of the CNN architecture involves taking an image as input, followed by a convolution process to extract essential features, such as edges, texture, and shape. Afterward, a subsampling or pooling process is applied to reduce the dimensionality of the feature map while retaining important information. After multiple convolution and pooling layers, the feature map is flattened into a one-dimensional vector and connected to the fully connected layer, which finally uses a Softmax function to convert the output score into a probability distribution over the existing classes [5].

Convolutional Neural Networks (CNN) are one of the Deep Learning models that consist of automatic feature extractors and trainable classifiers. CNNs are designed to understand high-dimensional complex data with a specialized architecture that integrates convolution and subsampling layers. Although many CNN architectures have been developed for various tasks, such as object and handwriting recognition, the basic principle of CNN is to achieve the best performance in pattern recognition [6]. CNN also has a large representation capacity, where it learns the best features at each layer of the visual hierarchy. This makes CNN effective in various computer vision problems, such as object and handwriting recognition. One of the main advantages of CNN is its weight-sharing concept, which reduces the number of parameters that need to be trained and improves generalization [7].

This research aims to address the challenges of facial emotion recognition (FER) by optimizing CNN architectures to enhance accuracy and efficiency. The motivation lies in creating a system that can accurately interpret emotions, particularly for real-world applications such as mental health monitoring and interactive AI systems. The proposed method reduces the need for manual feature extraction and offers scalability across diverse datasets, making it applicable in various fields, including stress detection and user experience enhancement. A CNN-based approach is proposed to address the FER problem, with empirical evaluation showing that the performance improves when the outputs of different structured CNNs are combined and averaged, compared to using a single CNN architecture.

*Corresponding author

This article is written as follows: first, the introduction includes the research background, the research objectives to be addressed, and the general method proposed. Section II contains related works that reference previous studies, their relations, advantages, and limitations, and also theoretical foundation of the methods. Section III is the methodology, followed by the results and discussion in Section IV, which comprehensively presents the research findings and its analysis. Finally, this article gives the conclusion and suggestions for further research in Section V.

II. Related Works

Liu, Zhang, and Pan developed a Facial Emotion Recognition model using Convolutional Neural Networks (CNN) with the FER2013 dataset. The architecture design, layer depth, and number of neurons greatly influence the model's effectiveness. Large-scale CNNs face overfitting challenges and require high computational power. The model consists of three subnets with 8 to 10 layers, including an input layer, and three convolution layers with 3x3 filters at 64, 128, and 256 filters, respectively, followed by a max-pooling layer. Furthermore, there are three fully connected layers with 4096, 4096, and 7 neurons, ending with a SoftmaxLoss layer for classification. The third subnet performed best with 65.03% validation accuracy, especially in the surprise emotion category. However, the accuracy of the training data was not specified, and the model struggled with class imbalance, particularly in recognizing emotions like fear and sadness, where accuracy ranged between 58-60% [8].

This advanced research focuses on identifying human facial expressions in Indonesia using Convolutional Neural Networks (CNN) with the Indonesian Mixed Emotion Dataset (IMED). The IMED dataset consists of RGB images grouped into five categories, with 80% for training and 20% for testing. Data pre-processing includes face area cropping, image conversion to grayscale, and image dimension adjustment to 48x48 pixels. The network architecture consists of four-layer blocks, including a convolutional layer, activation layer (ReLU), normalization layer, pooling layer, and dropout layer. The initial stage uses a 3x3 kernel with 32 filters, resulting in a 48x48x32 feature map batch normalization and ReLU activation. Max-pooling with a 2x2 kernel is applied to reduce the spatial dimension. Subsequent blocks add filters to capture more complex information. In the final stage, the feature map is flattened into a 1D vector and processed through a fully connected layer for classification. This study reached a validation accuracy of 93.63% [9].

Research on the use of Convolutional Neural Networks (CNN) for facial expression recognition uses three datasets: FER2013, Cohn-Kanade (CK+), and Karolinska Directed Emotional Faces (KDEF). The CK+ dataset comprises 981 images, while KDEF has 490 images adjusted to 48x48 pixels. The model was developed with multiple convolution and fully connected layers and trained with various optimization scenarios and some epochs. As a result, the accuracy for the FER2013 dataset was 52% at 200 epochs and dropped to 49% at 500 epochs. For KDEF, the accuracy was 81% at 200 epochs and decreased to 77% at 500. The CK+ dataset showed 77% accuracy at 200 epochs and decreased to 71% at 500. When the three datasets were combined, the accuracy was 57% at 200 epochs and decreased to 54% at 500 epochs, indicating that variations in training and datasets affect the model performance [10]. This indicates limitations in the model's ability to generalize across different datasets and suggests potential overfitting, especially as performance declines with extended training.

A comparative study on FER with various machine learning techniques has been done to improve efficiency and accuracy. The methods tested include SVM, LR, ANN, RF, KNN, NB, and CNN, but the CNN architecture needs to be described. The evaluation was performed on ORL and Yale databases using accuracy, confusion matrix, and ROC performance measures. Results show that CNN and other deep learning models such as AlexNet, DenseNet, and LeNet achieve 100% performance on the ORL database, while traditional models such as SVM achieve 98.19%. However, deep learning techniques showed less satisfactory results on the smaller Yale dataset, emphasizing the importance of large datasets for optimal deep learning performance [11].

Research by Adrian et al. optimizes hyperparameters on CNN for facial emotion recognition using the FER2013 dataset, which is divided into training, validation, and test data with a ratio of 80-10-10. The proposed architecture includes five convolutional layers with 256, 512, 384, and 192 filter configurations and a dropout technique to reduce overfitting. Image pre-processing includes data augmentation. Results show significant improvement in accuracy, with a validation accuracy of 63.22% after 20 epochs and a test accuracy of 72.16% after 750 epochs. The model shows competitive performance with lower computational overhead than complex models such as VGG and ResNet; it is ranked ninth on the FER2013 benchmark on PapersWithCode [12]. Although the model achieved good performance on the test set, real-time application testing is not mentioned. It is unclear how well the model would perform when integrated into a real-time system, where speed and efficiency are crucial, especially on devices with limited computational resources.

Khaireddin et al. discussed facial emotion recognition using CNN architecture with a pre-trained VGGNet model on the FER2013 dataset divided by an 80:10:10 ratio for training, validation, and testing. The research emphasizes the effectiveness of CNN in automatic feature extraction and computational efficiency. The model uses four convolution layers and optimization techniques to improve performance and change the learning rate. The best accuracy achieved was 73.28% without additional training data, with Reduce Learning on Plateau callback optimization, and ranked eighth on the FER2013 accuracy benchmark in PapersWithCode [13].

Various previous studies show that Convolutional Neural Networks (CNN) technology has become the dominant method in facial expression recognition. Variations in the architecture, datasets, and training techniques used provide important insights into the effectiveness and adaptability of CNN models. These studies show that the depth and complexity of the CNN architecture, the number of neurons, and the selection of the right dataset greatly affect the model's performance, but also risk overfitting if not properly managed. There is great potential for

model development for facial expression classification, especially in the context of the Indonesian population, which may require models that are adaptive and relevant to local characteristics. In addition, to improve the model's generalizability, an effort will be made to combine the dataset of Western faces with non-Indonesian faces so that it is expected to produce a more diverse and adaptive model for the emotion category used.

III. METHODOLOGY

This research develops a Convolutional Neural Network (CNN)-based Facial Emotion Recognition (FER) model as shown in Fig. 1. In general, the process started with data collection and filtering of emotional data, followed by exploratory analysis to understand the characteristics of the data. Defective and duplicate images are removed to maintain the quality of the dataset. Images are then converted to grayscale, faces are detected using Haar Cascade and resized for uniformity. The processed data was divided into training and testing sets. The model was trained by adding a new convolution layer with a 1024 filter and using callbacks such as early stopping and ReduceLROnPlateau to optimize performance. Finally, the model was tested to evaluate its accuracy and performance in recognizing emotions from facial images.

A. Data Collection

This study uses secondary data from three primary datasets: CK+, FER2013, and IMED. The FER2013 dataset includes 35,882 facial images with a resolution of 48x48 pixels in grayscale format, grouped into seven emotion categories. There is no specific information about the gender or age of the subjects in the images. Fig. 2 displays each example image in each class of the FER2013 dataset [14] [15]. The figure shows images converted into grayscale format with a resolution of 48x48 pixels. When viewed, images in the FER2013 dataset are covered by watermarks and hand poses that cover the face, which is expected to interfere with the learning process of the model and affect the accuracy. In addition, some emotion classes may need to be better balanced, with some emotions, such as disgust, having a much smaller number of samples than other emotions, such as happy or sad.

The CK+ dataset consists of 981 facial expressions from 210 individuals aged 18-50, with a gender distribution of 31% male and 69% female and a racial distribution of 81% Euro-American, 13% Afro-American, and 6% other. The dataset has seven emotion categories, and the images are in grayscale format with 48x48 pixel resolution. Fig. 3 shows each example image in each class of the CK+ dataset [10]. The number of images in each emotion class was observed, showing that this dataset has the least number compared to the other datasets used in this study. The pixel size of each image is 48x48 pixels and is in grayscale. The facial expressions of each emotion category are very clearly identified. In addition, the images in this dataset were taken under good and controlled lighting conditions, and all subjects were oriented directly toward the camera, which facilitated the analysis of facial expressions.

The IMED dataset contains 9,183 images of six male and nine female Indonesian subjects aged 17-32 from various Java, Batak, Sunda, Minang, and Manado ethnicities. The dataset comprises seven emotion categories and is organized in grayscale format with a resolution of 720x480 pixels. This dataset was obtained with special permission through the website https://imed.cs.ui.ac.id/. Fig. 4 shows each example image in each class of the IMED dataset [16]. It can be seen that the images in the IMED dataset have a resolution of 720 pixels wide and 480 pixels high, with an RGB image mode indicating full-color representation. The image quality in the IMED dataset has good and uniform lighting, and all the datasets are clean, with no noise, and are not covered by watermarks or hand poses covering the face area. However, the images are framed in black frames, indicating that they are still raw. Therefore, it is necessary to pre-process the data to produce uniform images with the other datasets, which ideally have a size of 48x48 pixels and are in grayscale mode.

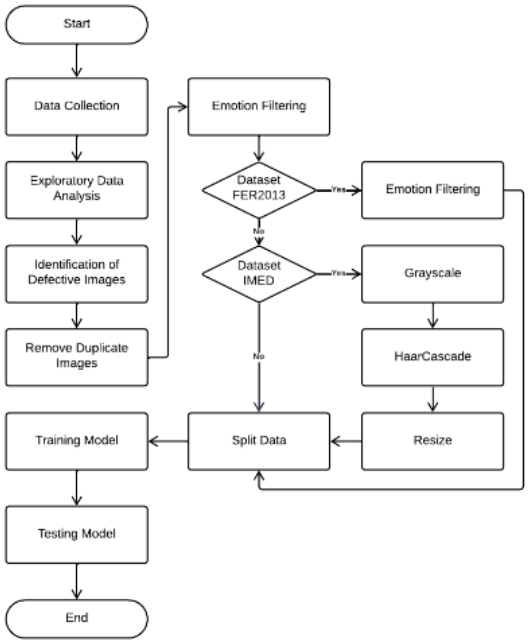


Fig. 1. Face recognition flowchart.

Fig. 2. FER2013 dataset images example.

Fig. 3. CK+ dataset images example.

Fig. 4. IMED dataset images example.

Table I shows the calculation details of each emotion class in the dataset.

TABLE I. CLASS DISTRIBUTION IN EACH DATASET

Emotion Class	FER2013	IMED	CK+
Neutral	6.193	518	54
Angry	4.953	1.623	135
Disgust	5.470	1.413	177
Fear	5.121	1.466	75
Happy	8.989	1.319	207
Sad	6.077	1.793	84
Surprise	4.002	1.051	249
Total	35.882	9.183	981

B. Data Preparation

In this study, the data pre-processing steps were carried out with a special focus on the IMED dataset, which is still in raw form, unlike the other datasets, which have been converted to grayscale and adjusted to 48x48 pixel dimensions. Details on the processing of the IMED dataset are shown in Fig. 5.

- Face detection and cropping of the image using the Haar Cascade Classifier method from the OpenCV library. Face detection is performed to ensure that only parts of the face are analyzed. This method compares parts of the image with pre-trained facial features like eyes, nose, or mouth. Once the face is detected, the area is cropped for face isolation [17]
- Grayscale conversion. In the IMED dataset, the image is converted to grayscale mode. This conversion reduces the complexity of the data but still retains the important features of the face. This process uses the cv2.cvtColor function with the cv2.COLOR_BGR2GRAY parameter.
- Resize Image: after the image is converted to a grayscale format, the image with variable sizes is resized to 48x48 pixels to maintain consistency with other datasets and ease of model processing.

Fig. 5. IMED dataset pre-processing steps.

In the FER2013 dataset, the data provided in CSV format requires conversion from a string of pixel values to an array of images. This process involved using np. from string to convert the string into a one-dimensional integer array, which was then reshaped into an image with a resolution of 48x48 pixels. This reshaping technique converts the data into a pixel matrix that the CNN model can process. A visualization of the processing of the FER2013 dataset is shown in Fig. 6.

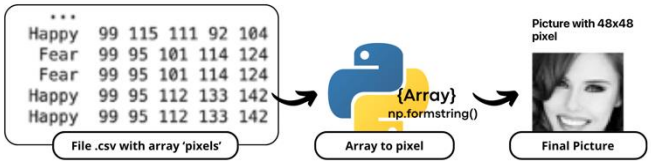


Fig. 6. FER2013 dataset pre-processing steps.

Afterward, emotion selection was conducted to simplify the emotion categories from 7 to 4. The four emotions are happy, angry, neutral, and sad. After emotion selection, the data was divided with a ratio of 80:18:2 for training, validation, and test data. The test data was then expanded by adding images taken from photos of real people, which was done to ensure a more genuine representation. The downsampling process balances the amount of test data between each emotion label. Two experiments were conducted with dataset splitting. The first experiment is conducted with the number of emotions in each dataset intact, with no reductions. Table II shows the training, validation, and test data used in the first experiment.

TABLE II. NUMBER OF TRAINING, VALIDATION, AND TESTING DATA

Data Split	Angry	Happy	Neutral
Training	5.160	8.208	5.240
Validation	1.132	1.856	1.183
Testing	148	148	148
Total	6.440	10.212	6.571

C. Model Architecture

This research uses the Convolutional Neural Network architecture. The CNN architecture used is adapted with the addition of one layer of convolutional blocks based on references from the PapersWithCode website, specifically from research entitled "Convolutional Neural Network Hyperparameters Optimization for Facial Emotion Recognition" [12], with details observable in Fig. 7.

The model starts with an input layer, where a grayscale and normalized face image of 48x48 pixels is used as input. Next, a series of convolution blocks extract visual features from the image. Each convolution block consists of a 2D Convolution layer with 3x3 filters; the first block has 256 filters, the second block 512, the third block 384, the fourth block 192, and the fifth block 512. The output of each convolution layer is normalized using BatchNormalization, and its dimension is reduced through MaxPooling2D with a 2x2 window. A dropout of 0.5 is applied after pooling to prevent overfitting.

Fully Connected Layers in this model include several layers: a Flatten Layer, which converts the multidimensional output of the last convolutional block into a one-dimensional vector, followed by a Dense Layer with 256 neurons and ReLU activation. A dropout of 0.5 is applied to the Dense Layer output to prevent overfitting. Finally, the Output Dense Layer uses four neurons corresponding to the number of emotion classes, equipped with a softmax activation function to generate class probabilities.

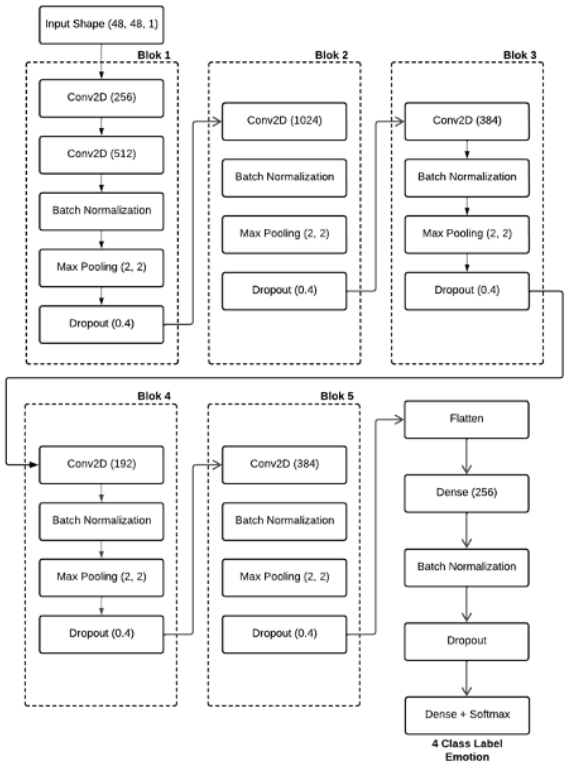


Fig. 7. Implemented CNN architecture.

D. Testing Scenario

Before going into the training and testing phase of the model, nine test scenarios were conducted to evaluate the model's performance under various conditions. These scenarios were designed to understand how datasets and model architecture variations affect the accuracy and loss of training and validation data. Table III summarizes the nine test scenarios conducted.

TABLE III. TESTING SCENARIO

Scenario	Description
1	Train using IMED dataset only
2	Train using CK+ dataset only.
3	Train using the FER2013 dataset only.
4	Train using combined IMED and FER2013 datasets
5	Train using combined IMED and CK+ datasets
6	Train using combined FER2013 and CK+ datasets
7	The first train without additional convolution layers with no minimum count using the FER2013 dataset
8	The second train without additional convolution layers with 1000 per label using the FER2013 dataset
9	The third train with the addition of a convolution block with a 1024 filter using the FER2023 dataset

Every scenario is designed to explore different aspects of model training. Scenarios 1 to 3 aim to measure the model's performance trained with individual datasets such as IMED, CK+, and FER2013. Scenarios 4 to 6 evaluate combinations of datasets to see the impact of integrating data from different sources. Scenarios 7 and 8 investigate the effects of variations in the amount of data and the use of additional convolution layers on the FER2013 dataset. Finally, scenario 9 involves adding convolution layers with large filters, batch normalization, max pooling, and dropout to capture more complex features and reduce overfitting. These scenarios helped identify the best configuration to implement the facial emotion recognition model.

E. Callback Mechanism

A callback method is used to optimize the process in model training using Keras TensorFlow. An early stopping callback is applied to automatically stop training if there is no improvement in validation accuracy [18] [19]. This allows the model to stop by seven earlier than the maximum number of epochs if the model is no longer improving, thus saving computational time and resources. Model saving is done through Model Checkpoint, which only saves the model if there is an increase in validation accuracy; this ensures that the saved file is optimal based on validation accuracy. Furthermore, the ReduceLROnPlateau callback method reduces the learning rate of the model by 80% from its previous value if there is no improvement in validation accuracy for three consecutive epochs, considering the minimal change in validation accuracy of 0.0001 to be regarded as an improvement [20]. Changes smaller than this value will not be considered.

The model is compiled with a loss function or sparse categorical cross-entropy loss function suitable for classification tasks where the target class label is an integer [21][22]. The optimizer used is Adam, which is used with early stopping callbacks, and ReduceLROnPlateau to optimize the training process with an initial learning rate of 0.001 [23]. The metric measured is accuracy, which indicates the percentage of correct predictions. Training is performed with the training dataset and validation data to evaluate model performance on data not used during training. During training, the model will go through a maximum of 50 epochs, which implies that the data will be processed 50 times to optimize the model weights based on the loss function and the accuracy measurement.

IV. RESULT AND DISCUSSION

After performing duplicate and corrupted image detection on each dataset, 3369 duplicate images were found on the FER2013 dataset and 273 duplicate images on the IMED dataset. No corrupted images were found in any of the analyzed datasets. In the image pre-processing stage, nine images from the IMED dataset are not detected by the Haar Cascade Classifier, as shown in Fig. 8. The duplicate images that the Haar Cascade Classifier does not detect will be removed to ensure the consistency and quality of the data used in training the model.

Fig. 8. Example of images that Haar Cascade Classifier does not detect.

After the pre-processing stage is complete, the next step is to perform model training using a Convolutional Neural Network (CNN). This training is done to optimize the model in recognizing facial emotions from images. Several experiments according to the design of the test scenario were conducted to evaluate the model's performance on various combinations of datasets, both individually and combined, and to explore the effect of the model architecture on its performance.

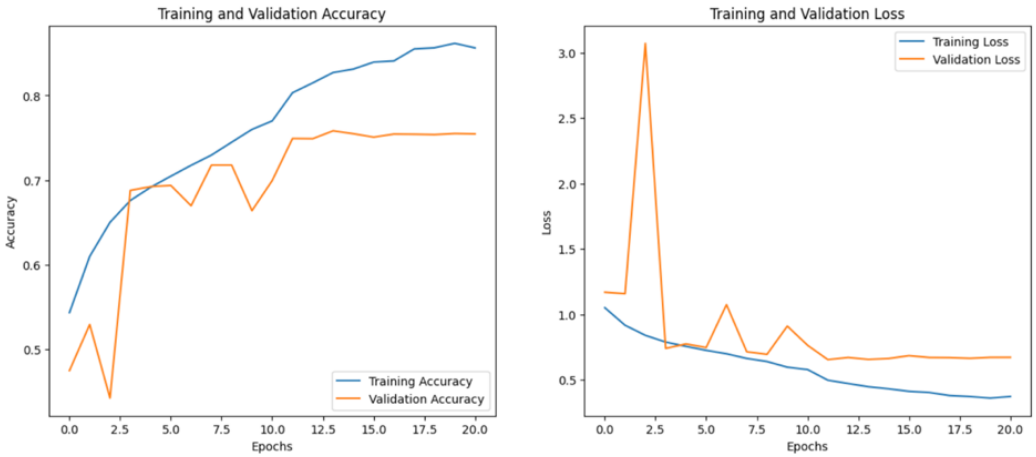


Fig. 9. Scenario 7: Train-validation accuracy and loss plot (training process was stopped early at the 21st epoch).

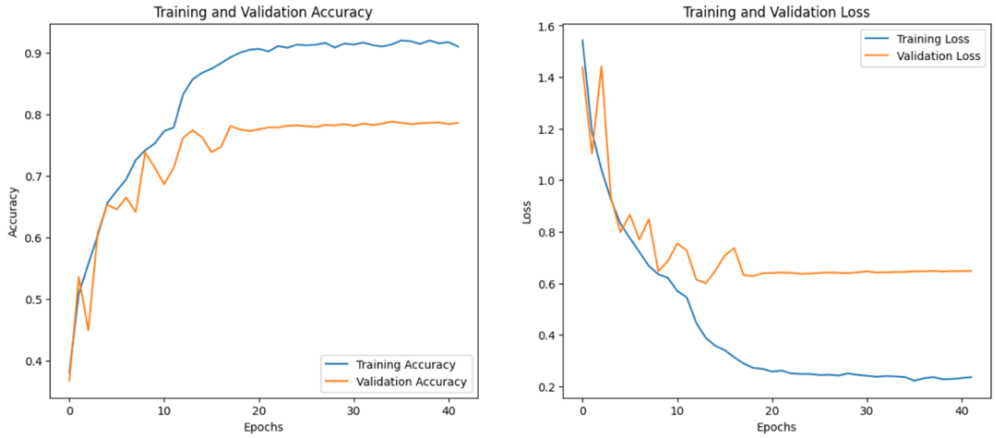


Fig. 10. Scenario 8: Train-validation accuracy and loss plot (each epoch took an average of 24 seconds).

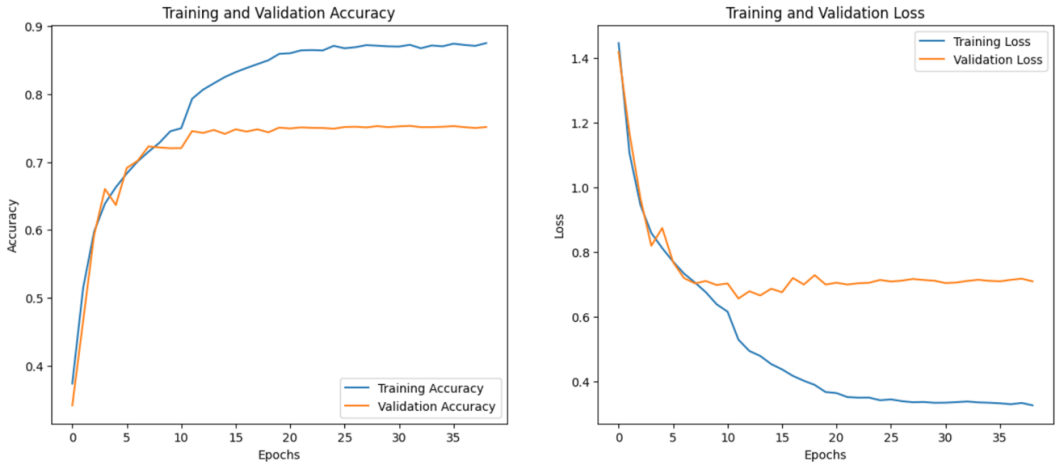


Fig. 11. Scenario 9: Train-validation accuracy and loss plot (validation loss shows a more stable decrease).

In the first experiment of training the combined model of three datasets, the highest validation accuracy was obtained at the 14th epoch with an accuracy value of 0.7582. At the 14th epoch, the model also recorded a validation loss of 0.6572, one

of the lowest loss values during the training session. For each epoch, the training took 49 seconds. The lowest validation loss value was recorded at the 12th epoch with a value of 0.6556, while the highest was recorded at the 3rd epoch with a very high value of 3.0713. This shows significant fluctuations in the model's performance at the beginning of training. During the training session, the model decreased the learning rate three times, precisely at the 11th, 17th, and 20th epochs. This decrease is done automatically by the ReduceLROnPlateau callback, which aims to improve validation accuracy when stagnation in performance improvement is detected. In addition, the more epochs performed, the higher the validation loss. The training process was stopped early at the 21st epoch, as shown in Fig. 9.

The training process of the second combined dataset lasted for 42 epochs before being stopped by the early stopping mechanism. Each epoch had 227 batches of data used for training. Furthermore, the early stopping method was set to stop training if there was no improvement in validation accuracy for 7 consecutive epochs, restoring the model weights to the state when the best performance was achieved. The model training started with a high initial loss of 15.42%, an accuracy of 38.03% on the training data, and about 36.69% on the validation data. Significant improvement occurred from the start of training, with training accuracy increasing to 75.24% and validation accuracy reaching 73.76% by the 9th epoch. The biggest improvement occurred at the 24th epoch, where the validation accuracy soared from 66.47% in the previous epoch to 78.07%. Each epoch took an average of 24 seconds. The accuracy results are shown in Fig. 10.

Furthermore, for the final experiment, adding a convolution layer with a filter of 1024, batch normalization, max pooling, and dropout resulted in a slight increase in model accuracy on the validation data. With this new configuration, the model achieved a validation accuracy of 79.59% after 39 epochs, with early stopping at the 39th epoch. This is compared to previous experiments that showed an increase in accuracy from 75.82% to 78.81% after data and hyperparameter adjustments. Adding a new convolutional layer with a large filter gives the model additional capacity to capture more complex features, thus improving classification ability. In addition, the validation loss of the new model shows a more stable decrease, as shown in Fig. 11.

TABLE IV. TRAINING RESULTS OF ALL SCENARIOS

Scenario	Train Accuracy	Train Loss	Validation Accuracy	Validation Loss
1	0.9855	0.1297	0.9809	0.0877
2	0.9766	0.0800	0.9792	0.1037
3	0.8272	0.4393	0.7226	0.7922
4	0.8637	0.3551	0.7575	0.7100
5	0.9716	0.0792	0.9859	0.1060
6	0.3018	0.8822	0.7178	0.8671
7	0.8271	0.4488	0.7582	0.6572
8	0.9135	0.2367	0.7881	0.6440
9	0.9339	0.1814	0.79558	0.6206

Table IV summarizes the model's best training results on four different datasets, namely IMED, CK+, FER2013, and a combined dataset, showing the variation in model performance in terms of accuracy and loss for both training and validation data at various points of the last epoch when training is stopped. This study shows that the model's accuracy decreases when the FER2013 dataset is mixed with other datasets, such as CK+ or IMED. The combination of FER2013 and CK+ resulted in a lower validation accuracy of 71.78%, with a high validation loss of 86.71% at the 39th epoch. Although the CK+ dataset is of high quality, the presence of FER2013 with high variability reduces the model's overall performance. The combination of IMED and FER2013 showed a validation accuracy of 75.75% and a validation loss of 71%. Although there is a performance improvement compared to using FER2013 alone, the results are still lower than those of using the IMED dataset individually. The CK+ and IMED datasets are more homogeneous and controlled, with more consistent expression variation and higher image quality. When the more varied and uncontrolled FER2013 dataset is mixed with the more homogeneous dataset, the model needs help generalizing the relevant patterns, thus lowering the overall performance. In the first and second experiments, no additional convolution layer was used.

Considering the results in Table IV for the three combined datasets in scenarios 7, 8, and 9, the best model performance is achieved in scenario 9. Therefore, the final model selected for further analysis is scenario 9. A confusion matrix study is conducted to better understand the model's performance in facial emotion classification. The confusion matrix provides an overview of how well the model classifies each emotion category and identifies areas that require further improvement [5] [24]. This analysis is important to evaluate the accuracy and weaknesses of the model in emotion recognition and to find out the most frequent types of errors.

TABLE V. TESTING RESULT CONFUSION MATRIX

	Predicted				
	Class	**Angry**	**Happy**	**Neutral**	**Sad**
True	Angry	32	3	10	2
	Happy	0	41	3	3
	Neutral	6	5	32	4
	Sad	9	3	7	28

The confusion matrix in Table V shows the model's performance in classifying facial emotions into four categories: angry, happy, sad, and neutral. For the original label angry, the model successfully classified 32 samples correctly, but ten samples were misclassified as neutral, three as happy, and two as sad. The happy original label performed better, with 41 samples correctly classified, although three samples were misclassified as neutral and three as sad. For the neutral label, 32 samples were correctly classified, while six were incorrectly classified as angry, five as happy, and four as Sad. Finally, the model classified 28 samples correctly for the original label sad, but there were errors with nine samples classified as angry, three as happy, and seven as neutral. This analysis shows that the model is quite accurate in classifying happy and neutral emotions but still needs improvement in classifying angry and

sad feelings, especially in distinguishing between angry and neutral.

TABLE VI. CLASS CLASSIFICATION RESULT

	Precision	Recall	F1-Score	Support
Angry	0.68	0.68	0.68	47
Happy	0.79	0.87	0.83	47
Neutral	0.65	0.68	0.65	47
Sad	0.76	0.60	0.67	47
Accuracy			0.71	188
Macro average	0.71	0.71	0.71	188
Weighted average	0.71	0.71	0.71	188

The results of the emotion classification metrics in Table VI show the model's precision, recall, and F1-Score performance for the angry, happy, sad, and neutral categories. For anger, the model achieved a precision and recall of 0.68. The happy category shows the best performance by the highest precision of 0.79, recall of 0.87, and F1-score of 0.83. The neutral category gets a precision of 0.65 and a recall of 0.68, while the sad category gets a precision of 0.76 and a recall of 0.60 with an F1-Score of 0.67. The model's overall accuracy was 0.71, with a macro-average and weighted average of 0.71 each.

Furthermore, the model is analyzed using the ROC Curve. The ROC curve illustrates the performance of the classification model at all possible classification thresholds by plotting two parameters: True Positive Rate (TPR) and False Positive Rate (FPR). AUC is the probability that the model will rank random positive samples higher than random negative samples. The ROC curve AUC is presented in the form of a curve graph. AUC values range from 0 to 1, where 1 indicates perfect prediction, and 0.5 indicates performance no better than a random guess [25].

Fig. 12 shows the ROC Curve in the range of 0-1 (higher is better) for scenario 9. The X axis is the false positive rate in the range of 0-1 and the Y axis is the true positive rate in the range of 0-1. The ROC shows the best performance for the happy class (AUC 0.97) and good for the angry and sad class (AUC 0.90). The neutral class has the lowest AUC of 0.87.

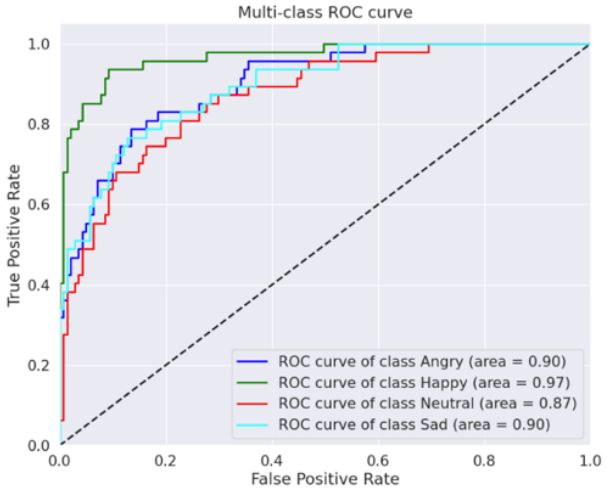
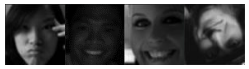
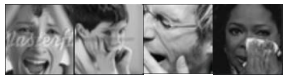


Fig. 12. ROC Curve on Scenario 9 classes.

The dataset characteristics affect the model's effectiveness, with variability in lighting and image quality making it difficult to extract important features. Table VII shows that images with unclear expressions or partially obscured faces degrade the model's performance, and poorly lit or non-face images hinder accurate classification.

TABLE VII. NON-CONFORMING IMAGES

Remarks	Images
Images with cartoon face	
Poorly lit images	
Face area covered by hands or watermarks	
Non-face images	

V. CONCLUSION

This research successfully developed a Facial Emotion Recognition model using Convolutional Neural Network (CNN) architecture by combining CK+, FER2013, and IMED datasets. In the 9th scenario, the model incorporating the three datasets with an additional convolutional layer of 1024 filters achieved the highest validation accuracy of 79.59%. The main factors for improved accuracy are better feature extraction capabilities and the application of batch normalization, max pooling, and dropout, which reduce overfitting and improve validation stability. The model achieved an average accuracy, precision, recall, and F1-Score of 71.5% on the test data.

Using the IMED and CK+ datasets individually provides high validation accuracy (97-98%), while the more varied FER2013 dataset provides a validation accuracy of 72.64%. Training parameter adjustments such as learning rate reduction and early stopping also improved model performance. This research shows that combining datasets and improved CNN architecture significantly affects the performance of facial emotion recognition.

ACKNOWLEDGMENT

We would like to thank: Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) Universitas Kristen Duta Wacana, Indonesia for the grant support No. 108/D.01/LPPM/2024.

REFERENCES

[1] T. Kumar Arora et al., "Optimal Facial Feature Based Emotional Recognition Using Deep Learning Algorithm," Comput Intell Neurosci, vol. 2022, 2022, doi: 10.1155/2022/8379202.

[2] S. Zhao, Y. Gao, X. Jiang, H. Yao, T. S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia, Association for Computing Machinery, Nov. 2014, pp. 47–56. doi: 10.1145/2647868.2654930.

[3] Purwono, A. Ma'arif, W. Rahmaniar, H. I. K. Fathurrahman, A. Z. K. Frisky, and Q. M. U. Haq, "Understanding of Convolutional Neural Network (CNN): A Review," International Journal of Robotics and

Control Systems, vol. 2, no. 4, pp. 739–748, 2022, doi: 10.31763/ijrcs.v2i4.888.

[4] L. Bejjagam and R. Chakradhara, "Facial Emotion Recognition using Convolutional Neural Network with Multiclass Classification and Bayesian Optimization for Hyper Parameter Tuning," 2022. [Online]. Available: www.bth.se

[5] B. C. Ko, "A brief review of facial emotion recognition based on visual information," Sensors (Switzerland), vol. 18, no. 2, Feb. 2018, doi: 10.3390/s18020401.

[6] J. U. Rahman, F. Makhdoom, and D. Lu, "ASU-CNN: An Efficient Deep Architecture for Image Classification and Feature Visualizations."

[7] L. Alzubaidi et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," J Big Data, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00444-8.

[8] K. Liu, M. Zhang, and Z. Pan, "Facial Expression Recognition with CNN Ensemble," in Proceedings - 2016 International Conference on Cyberworlds, CW 2016, Institute of Electrical and Electronics Engineers Inc., Nov. 2016, pp. 163–166. doi: 10.1109/CW.2016.34.

[9] A. Bayu Jala, "Implementasi Algoritma Convolutional Neural Network (CNN) Untuk Klasifikasi Ekspresi Wajah Manusia Di Indonesia," Bandung, 2020. Accessed: Oct. 09, 2023. [Online]. Available: https://repository.telkomuniversity.ac.id/pustaka/165458/implementasi-algoritma-convolutional-neural-network-cnn-untuk-klasifikasi-ekspresi-wajah-manusia-di-indonesia.html

[10] F. Azizi Nur, "Deteksi Emosi Menggunakan Citra Ekspresi Wajah Secara Otomati," 2021.

[11] A. A. Komlavi, K. Chaibou, and H. Naroua, "Comparative study of machine learning algorithms for face recognition", doi: 10.46298/arima.9291ï.

[12] A. Vulpe-Grigorasi and O. Grigore, "Convolutional Neural Network Hyperparameters optimization for Facial Emotion Recognition," in 12th International Symposium on Advanced Topics in Electrical Engineering, ATEE 2021, Institute of Electrical and Electronics Engineers Inc., Mar. 2021. doi: 10.1109/ATEE52255.2021.9425073.

[13] Y. Khaireddin and Z. Chen, "Facial Emotion Recognition: State of the Art Performance on FER2013."

[14] L. Pham, T. H. Vu, and T. A. Tran, "Facial expression recognition using residual masking network," in Proceedings - International Conference on Pattern Recognition, Institute of Electrical and Electronics Engineers Inc., 2020, pp. 4513–4519. doi: 10.1109/ICPR48806.2021.9411919.

[15] L. Zahara, P. Musa, E. Prasetyo Wibowo, I. Karim, and S. Bahri Musa, "The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi," in 2020 5th International Conference on Informatics and Computing, ICIC 2020, Institute of Electrical and Electronics Engineers Inc., Nov. 2020. doi: 10.1109/ICIC50835.2020.9288560.

[16] D. Y. Liliana, T. Basaruddin, and I. I. D. Oriza, "The Indonesian Mixed Emotion Dataset (IMED): A facial expression dataset for mixed emotion recognition," in ACM International Conference Proceeding Series, Association for Computing Machinery, Nov. 2018, pp. 56–60. doi: 10.1145/3293663.3293671.

[17] S. Yulina, "100-109 Dokumen diterima pada 21 Januari," 2021. [Online]. Available: https://jurnal.pcr.ac.id/index.php/jkt/

[18] F. Fischer, A. Birk, P. Somers, K. Frenner, C. Tarín, and A. Herkommer, "FeaSel-Net: A Recursive Feature Selection Callback in Neural Networks," Mach Learn Knowl Extr, vol. 4, no. 4, pp. 968–993, Dec. 2022, doi: 10.3390/make4040049.

[19] S. Samidin and A. Fadjeri, "Klasifikasi Gambar Batu-Kertas-Gunting Menggunakan Convolutional Neural Network dengan Fungsi Callback untuk Mencegah Overfitting," Jurnal Penelitian Inovatif, vol. 4, no. 2, pp. 785–794, Jun. 2024, doi: 10.54082/jupin.413.

[20] Y. Wu and L. Liu, "Selecting and Composing Learning Rate Policies for Deep Neural Networks," ACM Trans Intell Syst Technol, vol. 14, no. 2, Feb. 2023, doi: 10.1145/3570508.

[21] S. Chatterjee and A. Keprate, "Predicting Remaining Fatigue Life of Topside Piping Using Deep Learning," in 2021 International Conference on Applied Artificial Intelligence, ICAPAI 2021, Institute of Electrical and Electronics Engineers Inc., May 2021. doi: 10.1109/ICAPAI49758.2021.9462055.

[22] B. N. Chaithanya, T. J. Swasthika Jain, A. Usha Ruby, and A. Parveen, "An approach to categorize chest X-ray images using sparse categorical cross entropy," Indonesian Journal of Electrical Engineering and Computer Science, vol. 24, no. 3, pp. 1700–1710, Dec. 2021, doi: 10.11591/ijeecs.v24.i3.pp1700-1710.

[23] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Dec. 2014, [Online]. Available: http://arxiv.org/abs/1412.6980

[24] Z. Karimi, "Confusion Matrix," 2021. [Online]. Available: https://www.researchgate.net/publication/355096788

[25] M. Tripathi, "FACIAL EMOTION RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK," ICTACT JOURNAL ON IMAGE AND VIDEO PROCESSING, p. 1, 2021, doi: 10.21917/ijivp.2021.0359.