Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 21, No. 1, November 2021, pp. 199~206
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v21i1.1526

Convolutional Neural Network with Batch Normalization for Classification of Emotional Expressions Based on Facial Images

Bambang Krismono Triwijoyo[1], Ahmat Adil[2], Anthony Anggrawan[3]
[1,2,3]Universitas Bumigora, Indonesia

Article Info

Article history:

Received October 24, 2021
Revised November 1, 2021
Accepted November 10, 2021

Keywords:

Convolutional Neural Network
Batch Normalization
Classification
Emotional Expressions
Facial Images

ABSTRACT

Emotion recognition through facial images is one of the most challenging topics in human psychological interactions with machines. Along with advances in robotics, computer graphics, and computer vision, research on facial expression recognition is an important part of intelligent systems technology for interactive human interfaces where each person may have different emotional expressions, making it difficult to classify facial expressions and requires training data. large, so the deep learning approach is an alternative solution., The purpose of this study is to propose a different Convolutional Neural Network (CNN) model architecture with batch normalization consisting of three layers of multiple convolution layers with a simpler architectural model for the recognition of emotional expressions based on human facial images in the FER2013 dataset from Kaggle. The experimental results show that the training accuracy level reaches 98%, but there is still overfitting where the validation accuracy level is still 62%. The proposed model has better performance than the model without using batch normalization.

Corresponding Author:

Bambang Krismono Triwijoyo,
Department of Computer Science,
Universitas Bumigora.
Email: bkrismono@universitasbumigora.ac.id

1 INTRODUCTION

The study of human emotions is a multidisciplinary study including psychology, cognitive, sociology, and computer science. Research to recognize emotions through facial images is a challenging topic in recent years in the study of human-machine interactions. 55 percent of the effect of the message conveyed in communication is implied through facial expressions, while words or verbal only contribute 7 percent of the message conveyed and the remaining 38 percent is contributed by vocals. Facial expression recognition studies make it possible to understand human interactions not only through actions but also to understand feelings better. Based on this, it shows that recognizing facial expressions is the main capital of human interaction in communicating [1].

At first, the study of facial expressions was an area of research in psychology, sociology, and acting, but with the development of robotics and computer vision at the end of the 20th century, many computer researchers studied facial expressions which are the main part of interactive interfaces in intelligent systems technology [2]. Every human being can have different emotional expressions, this will be very easy to do by humans who can recognize various types of expressions. However, it is a challenging task in the field of computer vision to recognize a person or to classify the facial expressions of various people with varying emotional expressions. Facial expressions are a reflection of the emotional state of people [3]. There are six forms of facial emotional expression: fear, happiness, disgust, sadness, surprise, and anger [2]. Recognition of facial expressions through facial images shows very good results, however, when applied to the real world, many obstacles are caused by environmental lighting, facial skin color, and the position or angle of view of the camera [4]. Various methods are used to overcome this, including the use of implicit features for image classification [5].

Many studies on the recognition of facial emotional expressions have been carried out, among others [1] proposing an algorithm for the classification of human emotional states from still images of faces in grayscale. These images are from the Cohn-Kanade Extended public dataset and represent three emotional states: Aversion, Happiness, and Fear. The classification system is Support Vector Machine and has achieved an average accuracy of 98.52% in the classes mentioned above, but this study only classifies three types of expressions, while in the study the author proposes to recognize seven facial expressions. The convolutional neural network (CNN) and long short-term memory (LSTM) classifiers have been proposed by [5] to classify facial expressions with a large variety of poses. The proposed method was applied to two multi-pose facial expression databases: KDFE and RaFD, and satisfactory results were obtained for most of the facial expressions in various poses and have achieved an accuracy of 87.25%, the difference with our study is in the database used, where our study used the FER2013 database. The Gabor Wavelet Transform (GWT) and Local Binary Pattern (LBP) methods proposed by researchers [6], The dataset was taken from photos of 9 male facial transplant patients with an age range of 22 to 39 years. The experiment compared facial transplant patients with healthy people, and the results found that transplant patients did not reflect multiple emotional expressions and it was difficult to compare the differences in expressions. This study only focused on expression recognition in facial transplant patients, while our study applied to normal people, and succeeded in classifying seven facial emotional expressions.

Facial recognition from streaming face videos has been proposed by researchers [3] using Using entropy and correlation-based analysis, this method was tested using CK+ database. Experimental results show an accuracy of 95.8% and better, respectively, compared to competing approaches. Meanwhile, researchers [7] introduced the Wasserstein generative adversarial network-based method, using the AffectNet and RAF-DB databases. The experimental results show that the proposed method has achieved an accuracy of 83.49%, while the difference with the research that the author proposes is in the method and dataset used, where we use CNN with batch normalization and FER2013 dataset.

Researcher [8] introduced the facial reaction analysis system using Histogram of Oriented Gradient (HOG) and SVM. The experiment was conducted on a sample of 50 students. Each student is shown 100 random pictures. The experimental results show that the system can only detect emotions that are expressed externally on the face through physiological changes in certain parts of the face with an accuracy of 77.5%. The difference between this study and the method the author proposes is also the method and dataset used. Researchers [9] have proposed a facial expression recognition model under natural conditions, by implementing Linear Discriminant Analysis (LDA). The model will read the video when someone gives happy or sad news, and machine learning algorithms are used to distinguish between the emotional expressions of someone's face at the time of breaking the news. The results of the study found that the importance of using a database for training data as well as the use of natural images for the study of recognizing the emotional expression of a person's face. In contrast to research, in this study, the author proposes to use different methods and datasets.

From the previous studies above, all of them used hand-design feature extraction algorithms before the classification stage, while this study proposes the Convolutional Neural Network (CNN) method for recognizing human facial emotional expressions. Machine learning models have been widely used to identify facial objects in images [4]. Deep learning is a type of machine learning where feature extraction is carried out at the network layer, unlike conventional machine learning which requires a unique algorithm for feature extraction, before processing it on the network model. Deep learning computational models can study the data representation at each level of abstraction at each layer that receives raw input data and then automatically finds the representation needed for data pattern recognition or classification [10]. This study chose a deep learning model because the model can extract features from input data automatically and has good ability in image classification [11]. Another advantage of

deep neural networks is that they can generalize input data. One form of deep neural network architecture is Convolutional Neural Networks (CNN) presenting the most promising results for various applications [12] [13], [14].

In our previous study, we proposed a model for recognizing gender from facial images [15]. In this study, a different CNN model architecture with batch normalization is proposed, namely three layers of double convolutional layers with a simpler architectural model for the recognition of emotional expression based on human face images in the FER2013 dataset from Kaggle. After this section, the research methods used will be described which include the architecture of the training variable setting model and the description of the dataset used. In the next sub-section, the results and analysis of the model trials are described and the final section is the conclusion.

2 RESEARCH METHOD

The recognition model of human facial emotion expression in this study uses CNN by implementing batch normalization as used by [24]. Batch normalization (BN) can increase CNN learning much higher and act as a regulator, The drop out technique can be applied to several sophisticated image classification models to reduce network complexity, while the Batch Normalization technique can reduce the number of training steps with good accuracy.

2.1. Classification Model

We propose a deep learning model using CNN with batch normalization for the classification of human facial emotion expressions based on facial images. Figure 1 shows the model scheme that we made. The model consists of eighteen layers, starting from the input layer, the convolutional layer, the batch normalization layer, the pooling layer fully connected layer, and the end is the output layer.

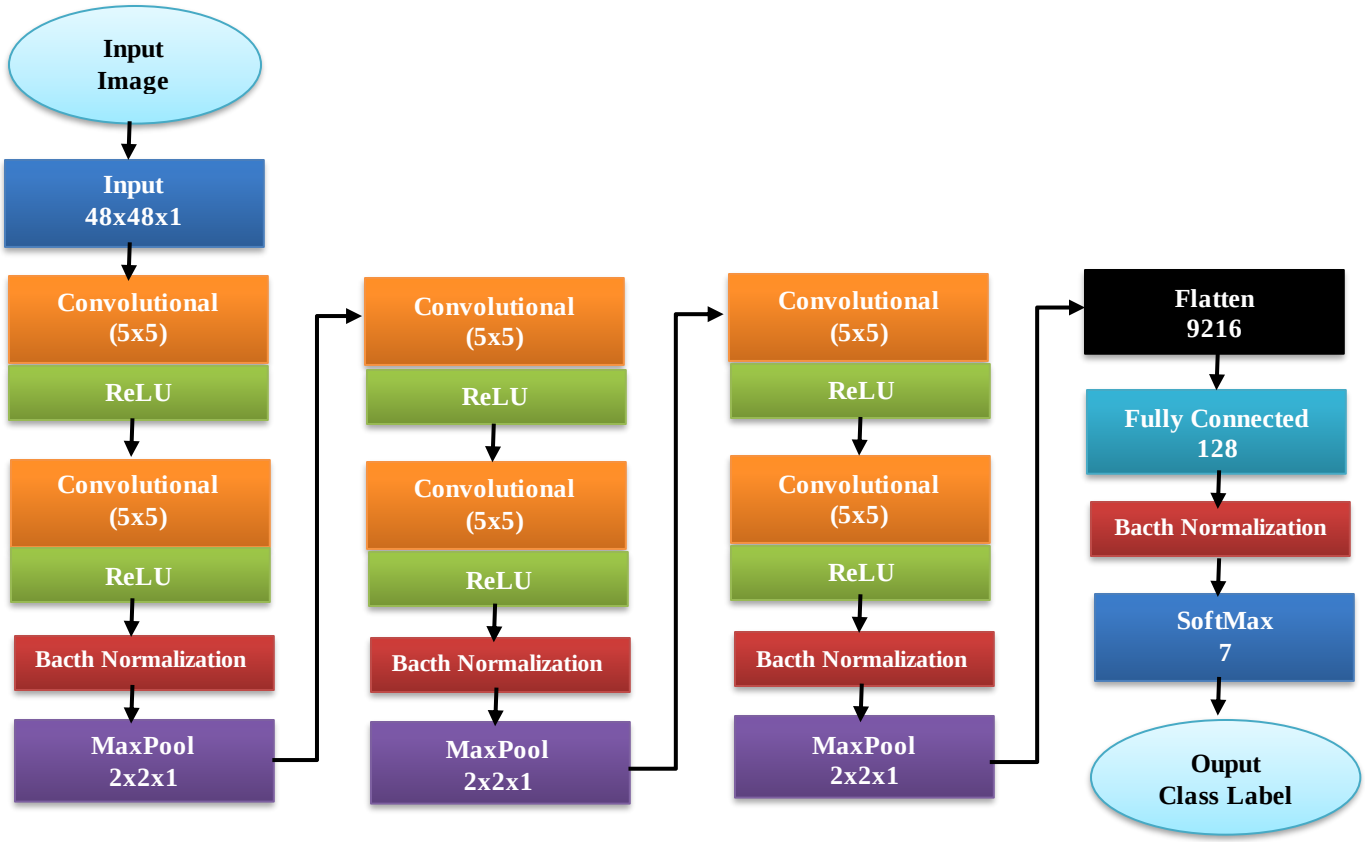

Figure 1. Model architecture proposed

The first layer is the input layer in the form of a face dimension 48x48 grayscale format, then three sets of layers, each consisting of two convolution layers and a ReLU layer, once a batch normalization layer and a pooling layer. The next set of layers is transformed into one dimension layer, fully connected, batch normalization, and the last layer is the softmax output layer. Table 1 shows the configuration of the proposed model.

Table 1. The Model Configuration

No	Layer Type	Batch Size	Kernel Size	Padding	Stride	Dropout	Activation Function	Output	Number Of parameters
1	Input							48x48x1	0
2	Convolutional	64	5x5	2	1		ReLU	48x48x64	1664
3	Convolutional	64	5x5	2	1		ReLU	48x48x64	102464
4	Batch Normalization	64						48x48x64	256
5	Max Pooling	64	2x2		1			24x24x64	0
6	Convolutional	128	5x5	2	1		ReLU	24x24x128	204928
7	Convolutional	128	5x5	2	1		ReLU	24x24x128	409728
8	Batch Normalization	128						24x24x128	512
9	Max Pooling	128	2x2		1			12x12x128	0
10	Convolutional	256	5x5	2	1		ReLU	12x12x256	295168
11	Convolutional	256	5x5	2	1		ReLU	12x12x256	590080
12	Batch Normalization	256						12x12x256	1024
13	Max Pooling	256	2x2		1			6x6x256	0
14	Flatten							9216	0
15	Dense							128	1179776
16	Batch Normalization					0.2	ReLU	128	512
17	Dense							7	903
18	Output						SoftMax	7	0

Input layer x^1 in the form of a 3rd order tensor represented in the following equation:

$$x^1 \in \mathbb{R}^{H^i \times W^j \times D^d} \tag{1}$$

where x^1 is the grayscale input image, with the row size (H) is 48, column size (W) is 48, and there are only one's channels, the dimensions of the image are 48 × 48 × 1, *(i, j)* is the index for each element, where $0 \leq i <$H, $0 \leq j <$W, and $0 \leq d <$ D. w^l, where D is a number of the kernels, each kernel *f*, of H × W, and *f* is $\mathbb{R}^{H \times W \times D}$. The input and output images in the convolutional layer to be of the same size using padding, the result of convolution sized is:

$$(H^l - H + 1) \times (W^l - W + 1) \times D \tag{2}$$

The convolution process is expressed through the following equation:

$$y_{i^{l+1},j^{l+1},d} = \sum_{i=0}^{H} \sum_{j=0}^{W} \sum_{d^l=0}^{d^l} f_{i,j,d^l,d} \times x^l_{i^{l+1}+i,j^{l+1}+j,d^l} + b_d \tag{3}$$

Where (b_d) is Bias and The Rectified Linear Unit (ReLU) layer is the transfer function through the following equation:

$$y_{i,j,d} = max\{0, x^l_{i,j,d}\} \tag{4}$$

Pooling Layer. Using the same notation derived from the convolutional layer, where $x^l \in \mathbb{R}^{H^l \times W^l \times D^l}$, where does not use learning or kernel parameters, where:

$$H^{l+1} = \frac{H^l}{H}, W^{l+1} = \frac{W^l}{W}, D^{l+1} = D^l \tag{5}$$

This study used max pooling, where the maximum pooling operator maps the sub-section to the maximum value of the element through the following equation:

$$max: y_{i^{l+1},j^{l+1},d} = \max_{0 \leq i < H, 0 \leq j < W} x^l_{i^{l+1} \times H + i, j^{l+1} \times W + j, d^l} \tag{6}$$

Where $0 \leq i^{l+1} < H^l, 0 \leq j^{l+1} < W^{l+1}$ dan $0 \leq d < D^{l+1} = D^l$.

At the layer where the CNN model is a fully connected layer. This study uses the softmax after the fully connected layer to generalize the logical function of a k-dimensioned *z* vector into a k-dimensioned $\sigma(z)$ vector with a real number value between [0, 1] using the following equation:

$$\sigma : \mathbb{R}^K \rightarrow [0,1]^K \tag{7}$$

$$\sigma(z) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \tag{8}$$

Where j is a range between 1 to K, σ is softmax, z is a vector the input and output layer, K is the z dimensions, and j output unit index. Batch normalization is applied to the activation network. This study applied affine transformations followed by non-linear elements z as the following equation:

$$z = g(Wu + b) \tag{9}$$

W and b are the learning parameters, and $g(\cdot)$ is the ReLU activation function. This formulation includes all. We add the BN transformation immediately before the activation function, by normalizing the following equation:

$$x = Wu + b \tag{10}$$

so it has a symmetrical distribution and results in activation with the stable distribution. Based on equation (10), bias b can be ignored because the effect will be canceled by the subsequent reduction of the mean. so equation (9) is replaced by:

$$z = g\big(BN(Wu)\big) \tag{11}$$

where BN transformations function to each dimension $x = Wu$, with the parameter pairs being learned $\gamma(k)$, $\beta(k)$ per dimension, with the scaling by γ and shift by β, to replace BN(x). For the convolutional layer, it is also normalized according to the convolutional property so different elements of the same feature map, in different locations, are normalized in the same way by normalizing all activations in a mini-batch, at all locations. B is the set of all values in the map feature in both elements of the mini-batch and spatial location so for mini-batches of size:

$$m' = |B| = m.pq \tag{12}$$

m and feature map size $p \times q$, are measured using parameter effects $\gamma(k)$ and $\beta(k)$ for each feature map.

2.2. Dataset

To test our model using the FER2013 dataset. The 2013 Facial Expression Recognition Database (FER-2013) was introduced in the 2013 ICML Challenge in Learning Representations which can be accessed on the page http://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge. The face is labeled as one of the six basic and neutral expressions. The resulting database in CSV format contains 35,887 grayscale images, 48x48 sized face images with various seven emotions, all labeled mostly in wild settings, Table 2 shows the number of images for six basic expressions and neutral faces in the FER2013 database.

Table 2. Number of images per each expression in FER2013

Label	Expression Type	Number Of Images
0	Anger	4953
1	Disgust	547
2	Fear	5121
3	Happiness	8989
4	Sadness	6077
5	Surprised	4002
6	Neutral	6198

Figure 2 shows an example image of facial emotion expressions from the FER2013 database. This research took 32298 samples for training data and 3589 for validation from the FER2013 source database.

Figure 2. The Sample of FER2013 database

2.3. Performance measurement

The empirical training and validation of model performance were measured using loss and accuracy. The Loss measure using stochastic gradient using momentum as an optimizer to minimize the loss rate formulated as:

$$Loss = \frac{1}{N}\sum_{i=1}^{b}\sum_{j=1}^{b_i}(t_{ij} - o_{ij})^2 \quad (13)$$

where t_{ij} is the actual class of the sample to j^{th} the training data to i^{th}, whereas o^{th} is the predicted class from sample to j^{th} training data to i^{th}, and N is the total number of sample testing or training [26]. While accuracy is the decision of the whole set of samples that positive ratings are positive and negative scores are negative according to the following equation:

$$Accuracy = \frac{(TF+TN)}{N} \quad (14)$$

where true positive (TF) is image class x is classified as image class x, true negative (TN) is image non-class x is classified as image non-class x, and N is the total training or testing samples

3 RESULTS AND ANALYSIS

The trial was carried out twice, in the first experiment using a model with additional batch normalization and the second experiment using a model without batch normalization. For each experiment, model training was carried out until the 20th epoch, the learning rate was 0.001, using categorical cross-entropy and Adam optimizer. The experiment was carried out on a computer with an Intel Core i7-7500U processor, 12 GB RAM, and GPU: NVIDIA GeForce GTX 960, While the software system environment used Windows 10 operating system, Python 3.6 Programming Language with Jupyter notebook editor.

Figure 3 shows the graph of loss and accuracy of each model by adding batch normalization and from the model without normalization. In figure 3, it can be seen that in the model using batch normalization, the loss rate decreases faster to below 20%, and the accuracy rate also increases more quickly above 90% at the 10th epoch, the contrast occurs in the model without batch normalization until the 20th epoch, the loss rate is still above 30% and the training accuracy rate is still below 90%. The experimental results show that up to the 20th epoch the training accuracy level reaches 98%, but there is still overfitting where the validation accuracy level is still 62%. This shows that the facial expression recognition model using batch normalization produces better performance than the model without using batch normalization.

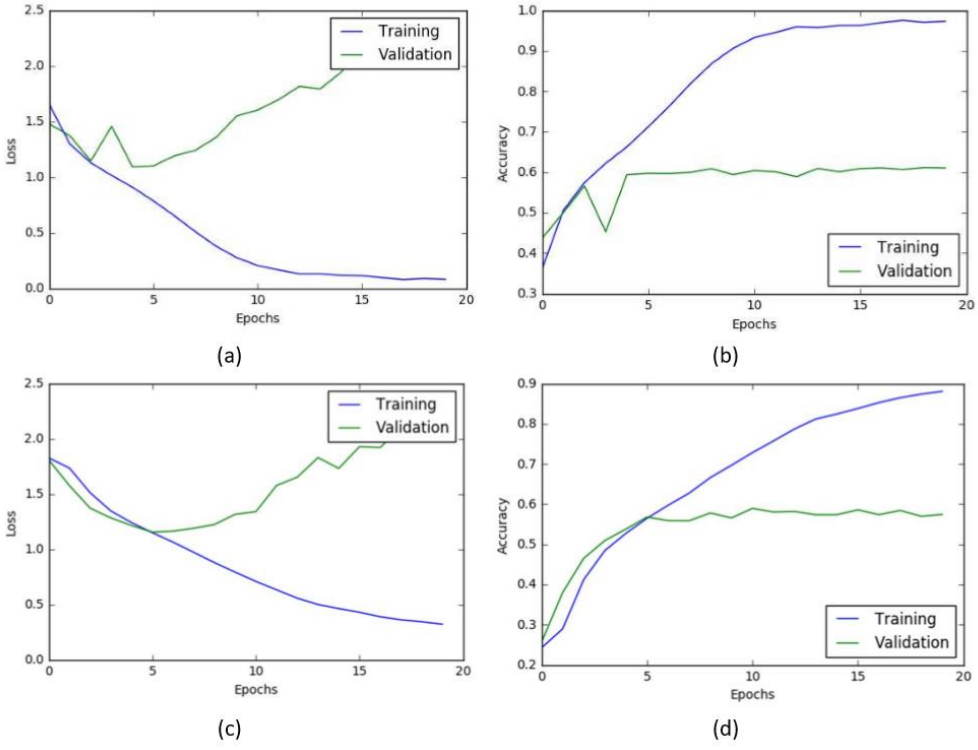

Figure 3. Graph of (a) Loss, (b) Accuracy of the model with BN and (c) Loss, (b) Accuracy of the model without BN.

Table 3 shows the comparison of the accuracy results between the proposed model and the results of previous similar studies, although, with different methods and datasets, it is evident that the proposed model is relatively better, except for the results of [1] . This proves that the use of the CNN model using batch normalization can improve model performance in the training and validation process.

Table 3. The Comparison of Model Performance with Previous Research

Reference	Method	Database	Accuracy
Bortolini et al [1]	Support Vector Machine (SVM)	Cohn-Kanade	98.52%
Agarwal et al[3]	Local Binary Pattern (LBP) and SVM	CK+	95.8%
Hassouneh et al[5]	CNN and LSTM	EEG	87.25%
Bedeloglu et al[6]	Gabor Wavelet Transform (GWT) and Local Binary Pattern (LBP)	photos of 9 male facial transplant patients	n/a
Lu et al[7]	Wasserstein Generative Adversarial Network (WGAN)	AffectNet and RAF-DB	83.49%
Magdin et al[8]	Histogram of Oriented Gradient (HOG) and SVM	NAPS	77.5%
Watson et al[9]	Linear Discriminant Analysis (LDA)	Local Dataset	n/a
Our Proposed Method	CNN with Batch Normalization	FER20013	98%

4 CONCLUSION

This study has proposed a CNN model architecture using batch normalization, the model has three layers of double convolution layers with a simpler architectural model for the recognition of emotional expressions based on human facial images in the FER2013 dataset from Kaggle. The experimental results show the training accuracy level reaches 98%, This result is relatively better than [1], but there is still overfitting where the validation accuracy level is still 62%. The important findings and contributions of this research are the facial expression recognition model using batch normalization produces better performance than the model without using batch normalization, especially in the model training process. The next research is model tuning to eliminate overfitting and improve the performance of the model by adding drop size and transfer function.

REFERENCES

[1] V. P. Jr, C. Bortolini, H. R. Gamba, G. B. Borba, and H. Medeiros, "Facial Expression Classification Using Convolutional Neural Network and Support Vector Machine," in *Conference: WVC 2016 - Workshop de Visão Computacional, At Campo Grande, Mato Grosso, Brazil*, 2016, no. November, pp. 329–333.

[2] N. Meeki, A. Amine, M. A. Boudia, and N. Meeki, "Deep Learning for Non Verbal Sentiment Analysis : Facial Emotional Expressions," in *GeCoDe Laboratory, Department of Computer Science, Tahar Moulay University of Saida.*, 2020, vol. 3014, pp. 1–11.

[3] S. Agarwal, B. Santra, and D. P. Mukherjee, "Anubhav: recognizing emotions through facial expression," *Vis. Comput.*, vol. 34, no. 2, pp. 177–191, 2018.

[4] M. M and M. A, "Facial geometric feature extraction based emotional expression classification using machine learning algorithms," *PLoS One*, vol. 16, no. 2, pp. 1–12, 2021.

[5] A. Hassouneh, A. M. Mutawa, and M. Murugappan, "Development of a Real-Time Emotion Recognition System Using Facial Expressions and EEG based on machine learning and deep neural network methods," *Informatics Med. Unlocked*, vol. 20, p. 100372, 2020.

[6] M. Bedeloglu *et al.*, "Image-based Analysis of Emotional Facial Expressions in Full Face Transplants," *J. Med. Syst.*, vol. 42, no. 3, pp. 1–10, 2018.

[7] Y. Lu, S. Wang, W. Zhao, and Y. Zhao, "WGAN-Based Robust Occluded Facial Expression Recognition," *IEEE Access*, vol. 7, pp. 93594–93610, 2019.

[8] M. Magdin, L. Benko, and Š. Koprda, "A case study of facial emotion classification using affdex," *Sensors*, vol. 19, no. 9, pp. 1–17, 2019.

[9] D. M. Watson, B. B. Brown, and A. Johnston, "A data-driven characterisation of natural facial expressions when giving good and bad news," *PLoS Comput. Biol.*, vol. 16, no. 10, pp. 1–13, 2020.

[10] F. Qin, J. Guo, and W. Sun, "Object-oriented ensemble classification for polarimetric SAR Imagery using restricted Boltzmann machines," *Remote Sens. Lett.*, vol. 8, no. 3, pp. 204–213, 2017.

[11] L. Duran-Lopez, J. P. Dominguez-Morales, A. F. Conde-Martin, S. Vicente-Diaz, and A. Linares-Barranco, "PROMETEO: A CNN-Based Computer-Aided Diagnosis System for WSI Prostate Cancer Detection," *IEEE Access*, vol. 8, pp. 128613–128628, 2020.

[12] G. H. de Rosa and J. P. Papa, "Soft-Tempering Deep Belief Networks Parameters Through Genetic Programming," *J. Artif. Intell. Syst.*, vol. 1, no. 1, pp. 43–59, 2019.

[13] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel Convolutional Neural Network," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2015-Septe, no. July, pp. 1787–1794, 2015.

[14] A. George and S. Marcel, "Learning One Class Representations for Face Presentation Attack Detection Using Multi-Channel Convolutional Neural Networks," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 361–375, 2021.

[15] B. K. Triwijoyo, "Model Fast Tansfer Learning pada Jaringan Syaraf Tiruan Konvolusional untuk Klasifikasi Gender Berdasarkan Citra Wajah," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 18, no. 2, pp. 211–221, 2019.