Emotion Recognition System
Using Autoencoder + CNN + Attention

Mikhail Aresa Latumahina*[1], Handri Santoso[2], Ito Wasito[3], Haryono[4]
[1,2,3,4]Pradita University; Scientia Business Park, Jl. Gading Serpong Boulevard No.1, Curug Sangereng, Kec. Klp. Dua, Kabupaten Tangerang, Banten 15810, 02155689999
[1]Magister Teknologi Informasi, Pradita University, Banten
e-mail: ***[1]mikhail.latumahina@gmail.com**, [2]handri.santoso@pradita.ac.id, [3]ito.wasito@pradita.ac.id, [4]haryono@pradita.ac.id

Abstrak

Pada era Digital Transformation banyak bisnis yang menggunakan teknologi berupa Deep Learning yang digunakan untuk mengubah cara menjalankan bisnis, salah satu metode yang digunakan adalah Emotion Recognition. Emotion Recognition sendiri merupakan bagian dari Computer Vision, dan task-task computer Vision sudah biasa dilakukan dengan menggunakan algoritma CNN. Akurasi merupakan hal yang penting didalam Emotion Recognition dimana banyak penelitian menggunakan berbagai metode baik Transfer ataupun Hybrid learning untuk mencoba meningkatkan aspek tersebut, sehingga penelitian ini bermaksud untuk merancang Autoencoder + CNN + Attention yang dapat digunakan untuk Emotion recognition, yang dibuat dengan menggabungkan Encoder, CNN, dan Attention Mechanism. model ini dilatih dengan menggunakan FER2013 dan dibandingkan dengan model CNN + Attention yang dilatih dengan cara yang sama. Walau Autoencoder + CNN + Attention tersebut berhasil mendapatkan 64% Accuracy di Evaluate Test_Model dibandingkan dengan CNN + Attention yang mendapatkan 55%, Harus diketahui bahwa penyesuaian masih harus perlu diperlakukan karena akurasi 43% testing terhadap data luar seperti tuning, penyesuaian layer, dan augmentasi data FER2013.

Kata kunci—: *Autoencoder, CNN, Computer Vision, Emotion Recognition, FER2013*

Abstract

In the Digital Transformation era, many businesses use technology in the form of Deep Learning which is used to change the way business is run, one of the methods used is Emotion Recognition. Emotion Recognition itself is part of Computer Vision, and computer vision tasks are usually done using the CNN algorithm. Accuracy is important in Emotion Recognition where many studies use various methods, both Transfer and Hybrid learning to try to improve this aspect, so this research intends to design a Autoencoder + CNN + Attention that can be used for Emotion recognition, which is made by combining Encoder, CNN, and Attention Mechanisms. this model is circumspect by using FER2013 and compared to the CNN + Attention model which is shutting down in the same way. Even though the Autoencoder + CNN + Attention managed to get 64% Accuracy in Evaluate Test_Model compared to CNN + Attention which got 55%, it should be noted that adjustments still have to be treated because of the 43% sensitivity of testing on external data such as tuning, layer adjustments, and FER2013 data augmentation.

Keywords— *Autoencoder, CNN, Computer Vision, Emotion Recognition, FER2013.*

1. INTRODUCTION

Digital Transformation is a term for which the process of integrating digital technology into the processes of an organization or business with the aim of creating a new business model, increasing revenue, reaching a larger scope of consumers, and providing a competitive advantage from an increasingly competitive business. others. One of the advantages brought by the utilization of Digital Transformation in a company is personalizing content for consumers, one of which is by using Artificial Intelligence or AI to detect a person's emotions towards a product or experience offered by an Institution. The AI method in question is Emotion Recognition or Emotion Recognition [1].

Emotion Recognition or Emotion Detection is a process of identifying one's emotions with a manual interpretation, or an automated process that uses AI assistance [2]. Emotion Recognition has been applied in several study case fields such as the use in measuring Virtual Reality Immersion, Robotics Interaction, psychology, and course Digital Marketing. There are several Emotion Recognition methods such as identifying a human emotion through Voice Analysis, Text Analysis using Natural Language Processing or NLP, and Face Recognition using CNN.

CNN works by identifying a person's emotions through the recognition of facial expressions produced by humans [3]. In 2020 Gede Putra Kusuma and Jonathan Andreas Pangestu Lim conducted Emotion Recognition research using a custom CNN model, namely VGGnet+GAP (Global Average Pooling) to increase the accuracy of the Emotion Recognition system. In 2020 Amil Khanzada, Charles Bai, and Ferhat Turker Celepcikay conducted Emotion Recognition research using Transfer Learning and Five-Layers technique to increase the overall accuracy for Emotion Recognition system [4]. And in 2021 Badrulhisham & Mangshor conducted research on Emotion Reignition only by using the CNN algorithm [5]. The three studies that have been exemplified emphasize how important the accuracy of the Emotion Recognition system is, and some suggestions that can be applied in continuing research are to try different hyperparameters and different models to improve the accuracy of the Emotion Recognition system. While progress has been made in refining Emotion Recognition, there's a realm of undiscovered methods. Combining Autoencoders, CNNs, and Attention could potentially boost accuracy

Based on these problems in this research, an Autoencoder + CNN + Attention will be designed by combining Encoder and CNN with the addition of an Attention Mechanism. The model will be trained using the FER2013 dataset and will be compared to the CNN + Attention Model which will be trained using a similar dataset and setup.

2. METHODOLOGY

2.1 *Literature Review*

The following section aims to briefly explore the theoretical framework that serves as the basis for our study.

2.1.1 *Deep Learning*

Deep Learning is a subset of the realm of machine learning which refers to the concept of Neural Networks that have more than one layer of neurons [6]. Deep Learning uses these layers of neurons to learn abstractions and representations that describe the intended data [7]. The representation layer in deep learning is a Neural Network system because the concept of deep learning has methods based on the concept of neural networks, especially artificial neural networks (ANN), the training process can be carried out supervised, semi-supervised, and unsupervised [8].

2.1.2 *Human Emotion*

According to Michael Cabanac, the definition of emotion is usually interpreted based on certain emotions [9]. The emotions in question can be classified as the basic emotions coined by Paul Ekman (1970) where emotions are discrete and not connected which include Happy, Sad, Anger, Fear, Surprise, and Disgust [10]. In contrast to Elkman, Robert Plutchik (1980) sparked the Plutchik Wheel of Emotion which describes a spectrum of emotions consisting of 8 primary emotions that are formed opposite each other between emotions that have different contrasts [11].

2.1.3 *Computer Vision*

Computer Vision is part of AI that allows computers to receive information from a digital image, video, or other visual input and act based on the input it provides. Computer Vision is created by combining image processing and pattern recognition processes [12], with the main objective of creating a model that can extract data and information from visual data [13]. The performance of the Computer Vision method will depend on the computer system, and the purpose of using Computer Vision [14].

2.1.4 *Emotion Recognition*

Emotion Recognition or Emotion Recognition is a process of identifying one's emotions with a manual interpretation, or an automated process that uses AI assistance [15]. In the context of this study, Facial Emotion Recognition, or FER, is where technology is used to analyze a person's facial expressions through photos or videos to determine the emotions a person is feeling [16].

2.1.5 *FER2013*

FER2013 or Facial Emotion Recognition 2013 is a dataset provided by Kaggle which was introduced by Pierre-Luc Carrier and Aaron Courvill, where in this dataset 35,887 photos measuring 48x48 emotions which can be categorized into 7 emotions such as Angry, Disgust, Fear, Happy, Sadness, Surprise, Contempt [17].

2.1.6 *Convolutional Neural Network*

Convolutional Neural Networks are part of the Deep Learning algorithms which are based on Artificial Neural Networks, CNNs are commonly used for Image Classification, Object Detection, and Computer Vision Tasks [18]. CNN consists of 4 layers including the Convolutional Layer, Pooling Layer, Activation Function, and Fully Connected Layer [19].

2.1.7 *Autoencoder*

Autoencoder is a Neural Network that can be categorized as Unsupervised Learning which is used for dimension reduction, feature learning, or data compression. The autoencoder consists of two parts, namely the Encoder which is used to process the Mapping data obtained from the input into a smaller form or Lower Dimensional, and the Decoder which is used to reconstruct the input data based on the data obtained from the Encoder [20].

Autoencoders can be designed using different types of neural networks such as Feed Forward Networks, Convolutional Networks, and Recurrent Networks, based on the objectives and methods that will be addressed to the model.

2.1.8 *Attention*

In Deep Learning Attention is a mechanism that helps a model to focus on certain parts, and gives a value or Weight Value to that focused part. In the context of Image Classification, the model can provide higher values for certain parts that are more important in the classification tasks performed by the model, this can increase the accuracy of the model's predictions [21].

2.2 *Related Works*

Gede Putra Kusuma and Jonathan Andreas Pangestu Lim (2020) researched FER using the Custom Sequential CNN model with the addition of the GAP or Global Average Pooling GAP method to increase the accuracy of the emotion recognition model. The model gets an accuracy of 69.40% [5].

Yijun Gan (2018) conducted FER research by making comparisons between several CNN Architecture models adapted to FER research where some of these architectures include VGG, ResNet, GoogleNet, and AlexNet, where AlexNet gets the best accuracy with 63.47% [22].

Vedat Tumen (2017) researched FER using Sequential CNN which was developed for FER, and trained using FER2013. The model obtained an accuracy of 57.1% in the FER2013 Test Dataset [23].

Khanzanda et al. (2018) researched FER using several techniques, namely vanilla CNN with 62% accuracy, Five-Layer model obtaining 75%, and Transfer Learning which combines several Vanilla Layers and CNN Architectures in the form of VGG, ResNet, and SeNet which obtains an average accuracy of 72.3% [4].

2.3 *Research Methods*

Research methods consist of dataset, designed model, and evaluation.

2.3.1 *Dataset*

FER2013 is the dataset to be used. FER itself has an extension of the Facial Emotion Recognition dataset 2013 which consists of 7 basic emotions with a size of 48x48, the data is provided in Kaggle datasets on behalf of FER2013 which is Free to Use. Some example data can be viewed below in figure 1 and the spread of training data and test data from FER2013 can be viewed in the table 1 below.

Figure 1. Examples From the FER2013 Dataset

Table 1. FER2013 Total Dataset

Number	**Emotion**	**Total Data**	
		Training Data	**Test Data**
1	Angry	3995	958
2	Sad	4830	1247
3	Disgust	436	111
4	Fear	4097	1024
5	Happy	7215	1774
6	Neutral	4965	1233
7	Surprise	3171	831

2.3.2 *Designed Model*

The Emotion Recognition system that is created will be divided into 2 models, namely Autoencoder + CNN + Attention, and CNN with Attention Mechanism, each model will be trained and used with the same dataset parameters, and the same epoch, namely 50 epochs. Figure 2 below illustrates the architectural process and model comparison.

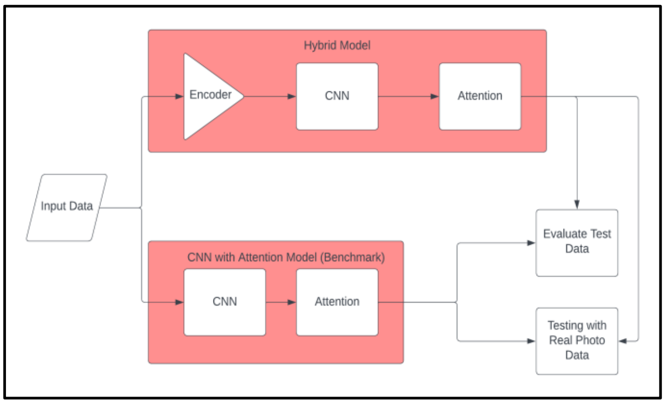


Figure 2. Illustration of Designed Model

2.3.3 *Evaluation*

Both models will be tested using the Model.Evaluate where the test dataset will be used to evaluate and predict manually from random photos that will be given to related models. The evaluation of the model will be carried out based on a comparison of the accuracy between the two models made in this study, and the model that has the highest accuracy will be determined.

3. RESULT AND DISCUSSIONS

3.1 *Model Training*

Because of hardware limitation out of the 34.000 data from the FER2013 Dataset only about 22.000 of the image data can be used for training the model, in which Table 2 represents the results of training both of the model.

Table 2. Training Results for Models

Training Results		
Model	**Accuracy**	**Loss**
Autoencoder+CNN+Attention	64%	1.24
CNN+Attention	55%	1.19

3.2 *Results*

The model is tested with 3 different methods, one is by using Model.Evaluate using the training data that has been provided by FER2013, detecting a sample of 100 randomly chosen images from the unused images of the training data, and trying to detect their emotions, and the other one is by using a sample image representing each emotion from the dataset. The test yielded results such as the figures and table below show.

By using Table 2 as a reference, we can conclude our position among 4 of the related works that have been stated above. And the models that have been proposed in this research standings could be viewed in Table 3 below which shows performance comparison.

Table 3. Comparison Model Performance

Model	**Accuracy**	**Reference**
ResNet50	49%	[21]

Sequential CNN	51%	[22]
CNN+Attention	**55%**	**(Our Comparison Model)**
VGG	59,64%	[21]
CNN Vanilla	62%	[4]
AlexNet	63,47%	[21]
GoogleNet	63,91%	[21]
Autoencoder+CNN+Attention	**64%**	**Our Model**
CNN + GAP	69,40%	[3]
Transfer Learning	72,30%	[4]
Five-Layer	75%	[4]

Comparison of the performance of models trained using FER2013 and evaluating them with the FER Dataset Test, where it can be seen that the Autoencoder + CNN+Attention model has higher accuracy compared to the comparison model, namely CNN+Attention and also several models such as Sequential CNN, VGG, and AlexNet. However, the Autoencoder + CNN + Attention model lost in the score evaluation to several models including CNN + GAP with an accuracy of 69.40% Transfer Learning which utilizes the combination of several CNN subsets with Pre-Trained Architecture such as VGG, FerNet50, and SeNet50 with a significant accuracy of 72.5%, then there are Five Layers which has 75% accuracy.

In addition to comparisons with Test Evaluate, comparisons between the Autoencoder + CNN + Attention and CNN + Attention models are carried out using a Custom Function to select 100 random data from the Test dataset and identify them with Table 4 displaying the results out of 15 iterations.

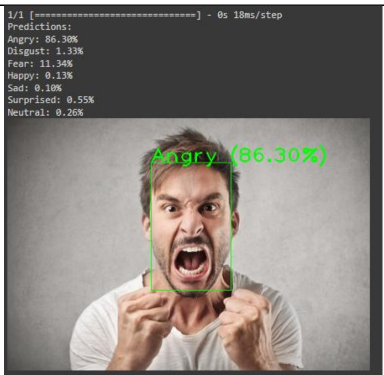
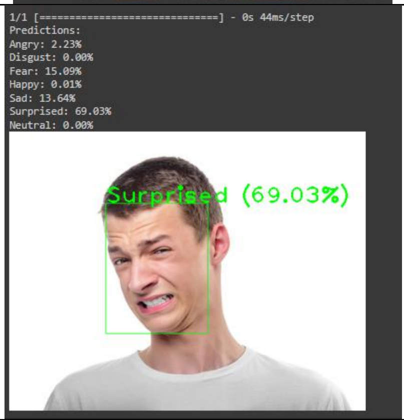
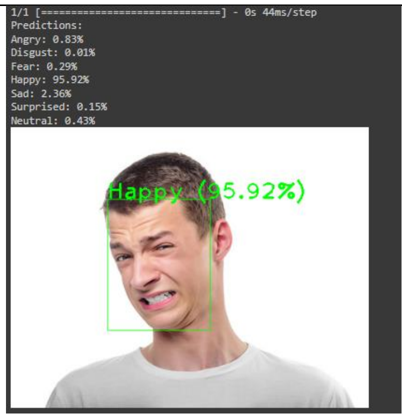
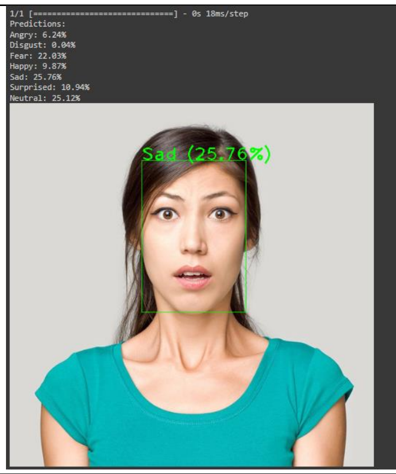
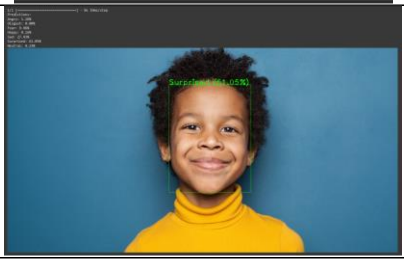
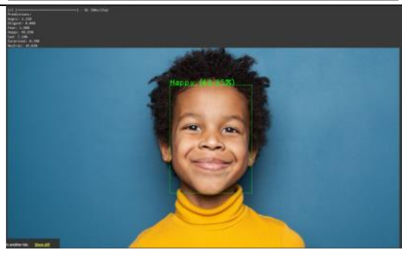
Table 4. Results of 15 Iterations of detection random 100 Picture Samples

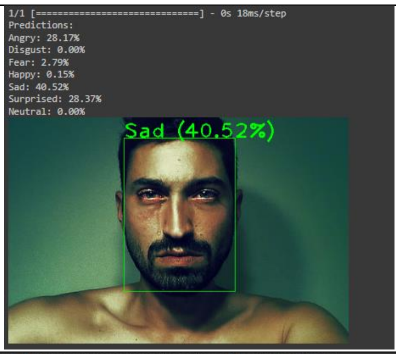
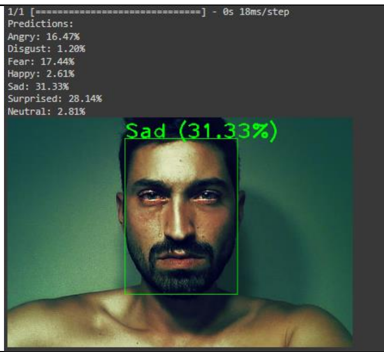
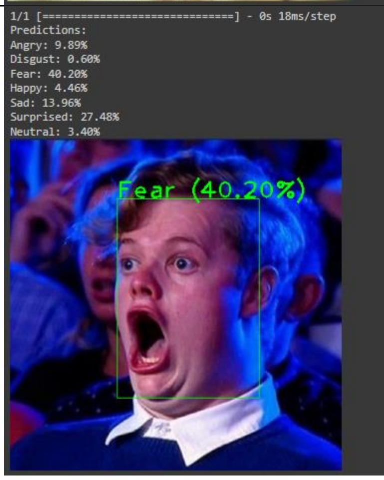
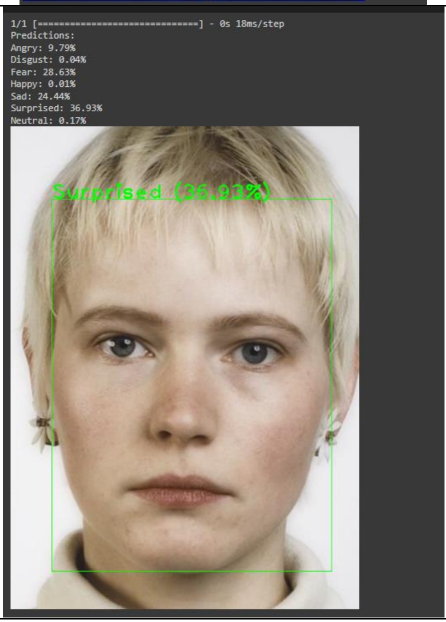
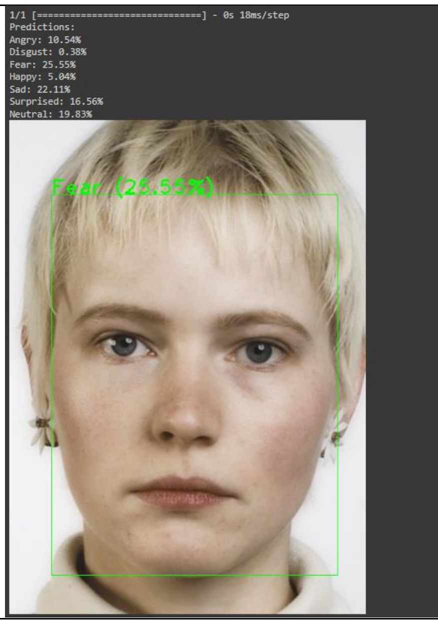
Iterasi	**Autoencoder+CNN+Attention**	**CNN+Attention**
1	65%	56%
2	58%	46%
4	57%	53%
5	61%	55%
6	60%	53%
7	67%	54%
8	66%	61%
9	64%	61%
10	69%	53%
11	66%	60%
12	62%	52%
13	77%	63%
14	65%	56%
15	71%	60%
Average	65%	56%

And the last test of using some random real data using image that is obtained from different sources and the results shown in table 5 below.

Table 5. Results of detecting emotion from Both Model.

Prediction

Emotion Name	*Autoencoder + CNN + Attention*	*CNN+Attention*
Angry		
Disgust		
Fear		
Happy		

<table>
<tr>
<td>Sad</td>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>Surprised</td>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>Neutral</td>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>ACC</td>
<td>43%</td>
<td>43%</td>
</tr>
</table>

Recognition and with the use of the FER2013 Dataset as a training base and getting a value of 43% can be an indication initial that this Hybrid model has potential, and can be developed by making some changes and adjustments to improve the performance of the model in future research. However, there are also several things to consider that affect the low accuracy value such as excessive Spatial Reduction, inadequate author computation, and the performance of the FER2013 dataset in real-data detection caused by Bias, Misslabeling Issues, and Error or Corrupt Images.

Based on all the experiments that have been carried out have been summarized in the table below for a summary of the experiment as table 6 provide.

Table 6. Summary of the Results

	Autoencoder + CNN + Attention	**CNN + Attention**
Test_Dataset Evaluate	64%	55%
100 Random Test Dataset Test Average	65%	56%
Dataset Luar Evaluate	43%	43%
Size Model	54MB	2.8GB

4. CONCLUSION

Attention Mechanism can be used to design an Emotion Recognition system using the FER2013 Dataset and when compared to the CNN + Attention model it has upper hand in Evaluation with Test_Dataset with 64%, 100 random data test set with 65%, and using an encoder has a much smaller model size compared to CNN + Attention with a model size of 54MB.

On the other hand, although it can be said that the two models performed poorly in testing using Outside Data, both models have the same level of accuracy with a value of 43% by guessing 3 emotions from the 7 provided Outside Data images. This can be caused by several things such as excessive spatial reduction or poor performance of the FER2013 dataset against real data.

5. SUGGESTION

In order to improve the research on Emotion Recognition, several suggestions can be made. Firstly, researchers should consider using powerful devices capable of efficiently training models using large datasets, often composed of numerous images. Secondly, it is recommended to impose spatial reduction restrictions, limiting the image dimensions to 24x24, which can aid in managing computational resources and potentially enhance model performance. Thirdly, utilizing a different dataset like AffectNet, which offers greater data variation compared to FER2013, could lead to more comprehensive results. Additionally, exploring alternative base models such as ResNet or EfficientNet, which were developed after VGG16, may provide different insights and outcomes. Moreover, incorporating Generative Adversarial Networks (GANs) into the research can generate unique training data and offer diverse representational features for the model. Lastly, conducting further investigations into the combination of different models, such as Autoencoder + CNN + Attention, has the potential to yield more accurate, efficient, and lightweight solutions. These suggestions aim to address the limitations of the current research and pave the way for advancements in Emotion Recognition.

REFERENCES

[1] Vial, G., 2019, Understanding digital transformation: A review and a research agenda, The Journal of Strategic Information Systems, 28(2), 118–144, https://doi.org/10.1016/j.jsis.2019.01.003.

[2] Tang, D., 2021, What is Digital Transformation? EDPACS, 64(1), 9–13, https://doi.org/10.1080/07366981.2020.1847813.

[3] Kusuma, G. P., Jonathan, J., & Lim, A. P., 2020, Emotion recognition on FER-2013 face images using fine-tuned VGG-16, *Advances in Science, Technology and Engineering Systems Journal*, 5(6), 315–322, https://doi.org/10.25046/aj050638.

[4] Khanzada, A., Bai, C., & Celepcikay, F. T., 2020, Facial expression recognition with deep learning, arXiv.org, https://arxiv.org/abs/2004.11823, [Accessed on 4/12/2023].

[5] Badrulhisham, N. A., & Mangshor, N. N., 2021, Emotion recognition using Convolutional Neural Network (CNN), *Journal of Physics: Conference Series*, 1962(1), 012040, https://doi.org/10.1088/1742-6596/1962/1/012040.

[6] Paluszek, M., & Thomas, S. J., 2020, *Practical matlab deep learning: A projectbased approach*, Anglin, S. Apress, New York.

[7] Sarker, I. H., 2021, Deep learning: A comprehensive overview on techniques, taxonomy, applications and Research Directions, *SN Computer Science*, https://doi.org/10.20944/preprints202108.0060.v1.

[8] Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. A., & Asari, V. K., 2019, A stateof-the-art survey on Deep Learning Theory and Architectures. *Electronics*, 8(3), 292, https://doi.org/10.3390/electronics8030292.

[9] Cabanac, M., 2002, What is emotion? *Behavioural Processes*, 60(2), 69–83, https://doi.org/10.1016/s0376-6357(02)00078-5.

[10] Ekman, P., 2005, Basic emotions, *Handbook of Cognition and Emotion*, 45–60, https://doi.org/10.1002/0470013494.ch3.

[11] Plutchik, R., 1980, A general psychoevolutionary theory of emotion, *Theories of Emotion*, 3–33, https://doi.org/10.1016/b978-0-12558701-3.50007-7.

[12] Long, J., Shelhamer, E., & Darrell, T., 2015, Fully convolutional networks for semantic segmentation, *2015 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), https://doi.org/10.1109/cvpr.2015.7298965.

[13] Babatunde, O. H., Armstrong, L., Leng, J., & Diepeveen, D., 2015, A survey of computer-based vision systems for automatic identification of plant species, *Journal of Agricultural Informatics*, 6(1), 61–71, https://doi.org/10.17700/jai.2015.6.1.152.

[14] Wiley, V., & Lucas, T., 2018, Computer Vision and image processing: A paper&review, *International Journal of Artificial Intelligence Research*, 2(1), 22. https://doi.org/10.29099/ijair.v2i1.42.

[15] Sailunaz, K., Dhaliwal, M., Rokne, J., & Alhajj, R., 2018, Emotion detection from text and speech: A survey, *Social Network Analysis and Mining*, 8(1), 28, https://doi.org/10.1007/s13278-018-0505-2.

[16] Zedrick, T., Horvath, A., & Vemou, K., 2021, facial emotion recognition. European Data Protection Supervisor, https://edps.europa.eu/dataprotection/ourwork/publications/techdispatch/techdispatch-12021-facial-emotionrecognition_en, [Accessed on 1/29/2023].

[17] Zahara, L., Musa, P., Prasetyo Wibowo, E., Karim, I., & Bahri Musa, S., 2020, The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the Convolutional Neural Network (CNN) algorithm-based Raspberry Pi, *2020 Fifth International Conference on Informatics and Computing* (ICIC), https://doi.org/10.1109/icic50835.2020.9288560.

[18] Valueva, M. V., Nagornov, N. N., Lyakhov, P. A., Valuev, G. V., & Chervyakov, N. I., 2020, Application of the residue number system to reduce hardware costs of the convolutional neural network implementation, *Mathematics and Computers in Simulation*, 177, 232–243, https://doi.org/10.1016/j.matcom.2020.04.031.

[19] Indolia, S., Goswami, A. K., Mishra, S. P., & Asopa, P., 2018, Conceptual understanding of convolutional neural network- a deep learning approach, *Procedia Computer Science,* 132, 679–688, https://doi.org/10.1016/j.procs.2018.05.069.

[20] Michelucci, U., 2022, An introduction to autoencoders, arXiv.org, https://arxiv.org/abs/2201.03898, [Accessed on 4/4/2023].

[21] Brauwers, G., & Frasincar, F., 2022, A general survey on attention mechanisms in deep learning, arXiv.org, https://arxiv.org/abs/2203.14263, [Accessed on 5/24/2023].

[22] Gan, Y., 2018, Facial expression recognition using convolutional neural network, *Proceedings of the 2nd International Conference on Vision, Image and Signal Processing*, 1–5, https://doi.org/10.1145/3271553.3271584.

[23] Tumen, V., Soylemez, O. F., & Ergen, B., 2017, Facial emotion recognition on a dataset using convolutional neural network, *2017 International Artificial Intelligence and Data Processing Symposium* (IDAP), https://doi.org/10.1109/idap.2017.8090281.