Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 1, February 2025, pp. 261~267
ISSN: 2302-9285, DOI: 10.11591/eei.v14i1.8707

Hyperparameter tuning for deep learning model used in multimodal emotion recognition data

Fernandi Widardo[1], Andry Chowanda[2]
[1]Department of Computer Science, BINUS Graduate Program–Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia
[2]Department of Computer Science, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received May 17, 2024
Revised Sep 4, 2024
Accepted Sep 28, 2024

Keywords:

Convolutional neural network
Convolutional neural network-bidirectional long short-term memory
Deep learning
Hyperparameter tuning
Multimodal emotion
Recognition

ABSTRACT

This study attempts to address overfitting, a frequent problem with multimodal emotion identification models. This study proposes model optimization using various hyperparameter approaches, such as dropout layer, l2 kernel regularization, batch normalization, and learning rate schedule, and discovers which approach yields the most impact for optimizing the model from overfitting. For the emotion dataset, this research utilizes the interactive emotional dyadic motion capture (IEMOCAP) dataset and uses the motion capture and speech audio data modality. The models used in this experiment are convolutional neural network (CNN) for the motion capture data and CNN-bidirectional long short-term memory (CNN-BiLSTM) for the audio data. This study also applied a smaller model batch size in the experiment to accommodate the limited computing resources. The result of the experiment is that the optimization using hyperparameter tuning raises the validation accuracy to 73.67% and the f1-score to 73% on audio and motion capture data, respectively, from the base model of this research and can competitively compete with another research model result. It is hoped that the optimization experiment results in this study can be useful for future emotion recognition research, especially for those who have encountered overfitting problems.

Corresponding Author:

Fernandi Widardo
Department of Computer Science, BINUS Graduate Program–Master of Computer Science
Bina Nusantara University
St. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
Email: fernandi.widardo@binus.ac.id

1. INTRODUCTION

Emotion is a core aspect of human communication and social aspect. We humans can interpret and respond to emotion cues. Emotional cues can be conveyed through body language or gestures, speech or any vocal audio, facial expression and nowadays a person's emotion cues can also be shown by their text in social media or messages [1]. Then, emotion recognition has become a vastly growing research field in computer science. In recent years, emotion recognition researchers utilize deep learning models. The deep learning models will use modalities to determine what emotion is shown.

Body language/gestures, speech/vocal audio, facial expressions, and text are called modalities in emotion recognition. In the early years, a single modality was used to recognize emotion shown by humans. But with the rapidly growing computing powers in newer technology, multiple modalities now can be used for emotion recognition. According to Liu *et al*. [2] multimodal emotion recognition has more detailed outcome than single modality/unimodal emotion recognition.

The future usages and implementation of emotion recognition are plentiful. Fields ranging from marketing/advertising, healthcare, education, customer service, human computer interaction, security, and surveillance have their own specific purposes for utilizing human emotions. Landowska *et al*. [3] states that products could influence human emotion/feelings whether to buy or not, hence researching human emotion based on products should become an object of interest by designer, investors, producers, and customer as well.

Many deep learning models are used for multimodal emotion recognition such as convolutional neural network (CNN), long short-term memory (LSTM), and recurrent neural network (RNN). The models' accuracies are shown to have improved but there is still a problem that almost every model has, which is overfitting. Overfitting has become one of the most fundamental problems in deep learning models [4]. Overfitting means that the data is not generalized enough and overly learns the training data. It is usually caused by limited or lack of data and the data complexity.

Consequences that are caused by overfitting can result in a model that can poorly perform in real data or not learned data. That is why we need to mitigate this issue by utilizing many forms of model optimization such as hyperparameter tuning or augmenting the data itself. In this paper, we will study and compare optimization methods that are mainly on the hyperparameter spectrum, such as the usage of dropout layer and l2 kernel regularization. This research also will utilize other hyperparameter tuning approaches such as learning rate scheduler and batch normalization. The model that will be used in this paper is CNN and CNN-bidirectional long short-term memory (CNN-BiLSTM) for combined motion capture (face, head rotation, and hand) and speech audio modality respectively. The main purpose of the research will be to try optimizing deep learning model mainly CNN and CNN-BiLSTM using interactive emotional dyadic motion capture (IEMOCAP) dataset from overfitting with hyperparameter tuning.

2. LITERATURE REVIEW AND RELATED WORKS

Multimodal emotion recognition is an emotion recognition model that uses more than one modality e.g., facial expression and audio combination or other combination of modalities. For the model, usually each modality is trained with each of their own models and then the models of each modality are concatenated in a layer before passed into a classification layer. The advantage of multimodal emotion recognition is that the results found are more detailed [2] and more efficient [5] than the unimodal emotion recognition model. Modalities that are commonly used are facial expression, speech-voice, and written text. Other modalities, for example, an aware healthcare system model that uses multimodal psychological signals [6] and electrocardiography (ECG) signals were also used in this research [7]. But in this research, we will discuss the research that uses IEMOCAP and mainly uses motion capture and audio dataset.

Deep learning models such as CNN, LSTM, and RNN are used and kept improving for multimodal emotion recognition. The research conducted uses CNN and LSTM for three modalities (text, audio, and mocap combined) from IEMOCAP dataset that the accuracy resulted is 71.04% with quite high overfitting shown by the accuracy graph that shows the accuracy of validation is not able to keep up with the training accuracy [8]. Then research using the RNN model combined with CNN, three-dimensional convolutional neural network (3DCNN), and LSTM was conducted, and the model's accuracy reached 71.75% for three modalities (text, audio, and mocap combined). Optimization such as data augmentation are also applied and the results are also compared by [9], the best result is for CNN+RNN+3DCNN architecture for audio and video modality with 3 classes of emotion. Dai *et al*. [10] proposed a new model called multimodal end-to-end sparse model (MESM) which replaced the original CNN layers in fully end-to-end model with N cross-modal sparse CNN blocks. The results of this proposed model three modalities (text, audio, and video) are average accuracy of 84.40% and average F1-score of 57.30%. Another proposed method is the usage of EmbraceNet+, which can merge the modalities and produce the prediction [11]. The results EmbraceNet+ for three modalities (face, audio, and text) are 77.60% accuracy with f1-score of 79% and for two modalities (face, audio) are 55.90% accuracy and f1-score of 59%. In this research [12], compares many models such as CNN, LSTM, BiLSTM, and dense, the accuracy can reach up to 98.48% accuracy, but the validation accuracy is only 67.24% for motion capture and audio modality model with 4 class of emotion and using data augmentation on the audio dataset.

In recent years, besides the common deep learning model, transformer models that have started to be used in unimodal emotion recognition, are now being researched for their usage in multimodal emotion recognition. Siriwardhana *et al*. [13] proposed a novel self-supervised transformer that uses feature fusion for text, speech, and video modalities. The results for the IEMOCAP experiment with 4 emotion classes are 86.50% accuracy and 85.7 F1-score. Xie *et al*. [14] proposed cross modality with a combination of FaceNet+RNN for facial modality, GPT for text modality, and wave RNN for speech modality. The accuracy

of the proposed model in combination is 65.00% and F1-score of 64.0 but note that the proposed model is applied on a different dataset, which is the multimodal emotionlines dataset (MELD).

Much of the research tries to optimize their models with many methods such as dataset augmentation and model hyperparameter tuning although some are not explicitly explained or focused on model optimization. Mainly the research focuses on the model architecture and its performance but methods like hyperparameter tuning are also important for the model performance. Hyperparameter tuning could help research on the model used from viewed as "not fitted" or "unsuccessful" to a solid solution by getting the correct combination of input values [15]. In this research, Liao *et al*. [16] also found that hyperparameter tuning results seems not significant in the model-only runtime environment but when applied to mobile devices, the performance become significantly different hence highlighting that other research can benefit from paying attention to various hyperparameter tuning. Hyperparameter tuning implementation can be done in manually or brute forcing, trying to find the correct value combination or using automation with various algorithms for hyperparameter optimization (HPO) [17]. Automated HPO lets us to just wait for the correct hyperparameter tuning combination as the algorithms automatically iterate itself until the most correct value is found but as a trade of human efforts in manually tune the hyperparameter, HPO requires huge amount computational resources and time [18]. The recently popular automated HPO is using meta-heuristic algorithms [19].

In deep learning model, there are optimization methods that can be hyperparameter tuned, such as dropout layer value, L1/L2 kernel regularization value, and learning rate. The dropout layer works by removing nodes from the original neural network based on probability value that can be tuned. This makes the neural network not dependent on the training data that could cause overfitting [20]. L1/L2 kernel regularization method is proposed and commonly used to reduce overfitting in deep learning model [21]. L1 adds a penalty value to the loss function that is proportional to the absolute value of the weights, meanwhile L2 adds a penalty value to the square of the weights [22], this penalty value can be tuned accordingly. Learning rate is the value that determines how big is the step of learning for each iteration the model trains itself on the data. The value of learning rate can be tuned and usually a learning rate scheduler is introduced. The learning rate scheduler usually monitors the loss value and decay the learning rate accordingly. Tuning learning rate method commonly used is manually sets the value but in recent years there are optimizers that can adjust its learning rate [23].

All these proposed approaches on emotion recognition in various datasets resulted in either low accuracy or high accuracy but with high overfitting. In this research, Alzubaidi *et al*. [24] found that the overfitting could come from both the dataset and the deep learning model itself. Hence, there will be the need for optimizing the model to be better fit or more generalized on the validation set.

3. METHOD

3.1. Dataset

The dataset that we are using in this research is IEMOCAP. The dataset is publicly available and contains approximately 12 hours of audiovisual data including video, speech, motion capture of face, and text transcriptions from 5 sessions [25]. This research will use two pieces of data, combined motion capture data and audio data. First, the data is processed to extract the information of the time, name, and the annotated emotions provided. The annotated data comes from the utterances in each recording session. The total utterance after processing is 5,521. 1,627 happy, 1,084 sad, 1,102 angry, and 1,708 neutral utterances.

For the audio dataset, we convert the audio into Mel-spectrograms, with y axis serving the time and x axis for the features. We use the library Librosa to process the feature extraction and Mel-spectrograms. The maximum length of each audio is set at 16 s for each audio as 16 s is the maximum length toleration. For the combined motion capture dataset, this research follows the approach in this research [8]. For each motion capture type (face, head rotation, and hand), we sample the data based on the start and finish time of the recording. Next, the values are split into 200 arrays and averaged. Then after averaging, all the types are concatenated for each utterance. Then the processed audio data and combined motion capture data is split into an 8:2 ratio to be fitted into the model.

3.2. Proposed base model

The high-level explanation of the base model for the hyperparameter tuning methods experiment is a deep learning multimodal emotion recognition model consisting of a combination between CNN and CNN-BiLSTM. The CNN model will be used for processing the combined motion capture modality and the CNN-BiLSTM will be used for audio modality. After both models process their respective modality, their layers will be concatenated before the classification layer. The high-level model is shown in Figure 1.

The model for speech data that uses CNN-BiLSTM consists of a pair of TimeDistributed(Conv1D) and TimeDistributed(MaxPooling1D), with the TimeDistributed(Conv1D) filter values are as follows: 64,

128, 64, 32, and 64. The kernel size is 3 and use relu as activation function. For the TimeDistributed(MaxPooling1D), the pool_size value is 2. After the data is processed in these layer blocks, the input data is flattened before sent to BiLSTM layers. The two BiLSTM layers consist of 1 layer with True return_sequence and 1 with false return_sequence. Both layer units are set at 128. Then the data is sent to the Dense layer with 256 units.

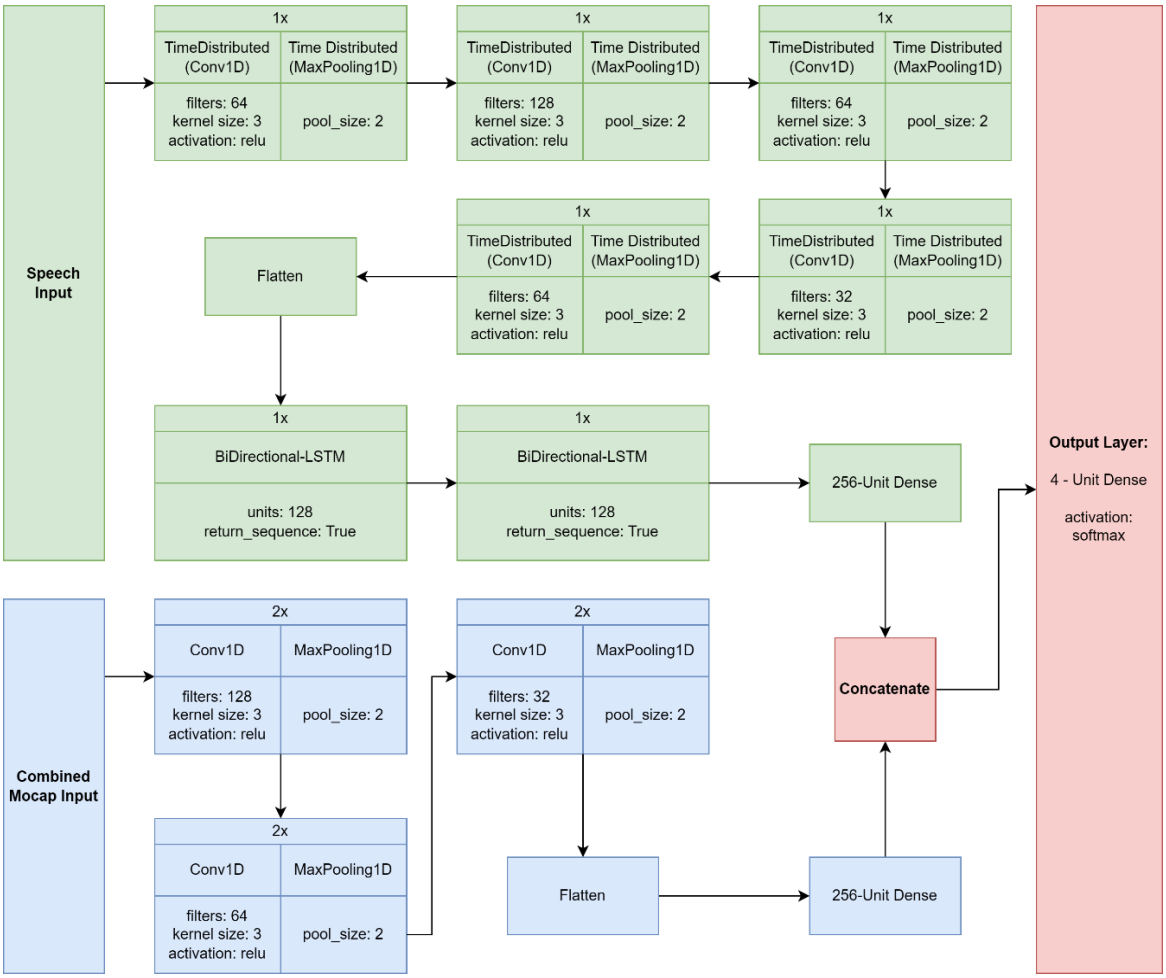

Figure 1. Proposed base model

The model for combined mocap data that uses CNN consists of a pair of two blocks of Conv1D and MaxPooling1D. With filter values as follows: 128, 128, 64, 64, 32, 32, and pool_size of 2. After that, the data is passed onto the 256-unit dense layer. The next step is to concatenate the results of the two models and pass it onto the final classification layer using dense layer with 4 unit and softmax activation function. The baseline model uses the default value of learning rate of Adam optimizer.

3.3. Proposed optimization using hyperparameter tuning approach for experiment

For the model optimization using hyperparameter tuning approach, this experiment will try to experiment on using hyperparameter tuning approaches such as dropout layer, l2 kernel regularization, batch normalization and learning rate schedule. The hyperparameter tuning approaches are done in manual or brute force method by testing the approaches values and whether they are used or not. This research using manual hyperparameter tuning in mind of limited computational resources. For the dropout layer, the dropout layer will be attached to the model after the dense layer of each of the modality models. The values for the dropout layer that are experimented with is 0.3 and 0.5. For the l2 kernel regularization, the regularization will be used in every convolutional layer of each modality model. The values that are experimented with are 0.0001,

0.001, and 0.01. For batch normalization, it will be inserted into each convolutional layer on the motion capture model and audio model. Then for the learning rate scheduler, this experiment uses base value of 0.001 and then we will be experimenting with reducing continuously by applying ReduceLROnPlateau with patience of 5 and with the monitor is set on the validation loss. Using ReduceLROnPlateau will help the model to not over learn a parameter when there is a stagnant learning detected. The evaluations metric for the model and the hyperparameter tuning approaches will be using accuracy metric and F1-score.

4. EXPERIMENT AND RESULT DISCUSSION

The research experiment was conducted using Google Colab Pro platform utilizing its V100 graphics processing unit (GPU) to avoid out-of-memory. The experiment hyperparameter for the model is epoch of 100, batch size of 32 and 48 for testing. The model also uses Adam optimizer and the modalities that are used and focused are the combined motion capture and sound. Before we get into the proposed model result, there is the need of explanation that some of the other research model for comparison doesn't report the complete value of the training accuracy, training loss, validation accuracy, validation loss, and F1-score. That is why we wrote only the reported averaged accuracy for some the model comparison.

From Table 1, we could see the proposed base model result is very high on the training accuracy reaching 99.96%, but the validation accuracy is quite low. Then in the hyperparameter tuning experiment, we try to use various method of hyperparameter tuning approaches and get the better generalized model with validation accuracy of 73.67% and f1-score of 0.73. For comparison, other models with two modalities with motion capture+audio or video+audio with 4 classes of emotion, our proposed base model doesn't have a much different in the result with another research [9], [11], [12]. But our proposed optimized model has a better generalization on the data with two modalities reaching training validation accuracy of 73.67% with f1-score of 0.73, sacrificing the training accuracy down to 84.13%. We also run the model in limitation of the computing power, the model trained with batch size of 32 at first and then increased to 48 but never been able to run with batch size of 64 because we will encounter out-of-memory error from lacking the GPU video random access memory (VRAM). This is caused by the audio data that is bigger in size than the motion capture data.

Table 1. Multimodal model result and comparison

Model	Modalities	Emotion classess	Accuracy	Loss	Validation accuracy	Validation loss	F1-score
CNN and LSTM [8]	Text+Motion Capture+Audio	4	0.9666	0.0836	0.6731	2.1394	-
CNN+RNN+ 3DCNN [9]	Audio+Video	4 and 3	-	-	0.5194 and 0.7175	-	-
MESM [10]	Text+Audio+Video	6	Avg accuracy=0.8440				0.5730
EmbraceNet+ [11]	Face+Audio+Text and Face+Audio	4	Avg accuracy=0.7760 and 0.5990				0.7900 and 0.5900
CNN [12]	Motion Capture+Audio	4	0.9758	0.0738	0.695	1.7082	0.7
Proposed base model	Motion Capture+Audio	4	0.9966	0.0198	0.6986	2.2242	0.69
Proposed optimized model	Motion Capture+Audio	4	0.8413	0.6173	0.7367	0.8993	0.73

The hyperparameter tuning approaches resulted in the best dropout layer value is 0.3. Then for the l2 regularizations, we find that different configurations for each modality model resulted in better results, 0.001 for combined motion capture model and 0.01 for audio model. Batch normalization layer first applied to all convolutional layer for both motion capture and audio model. But in the end, we removed the batch normalization for the audio model because the limitations of hardware that is causing out of memory error. The motion capture model still uses batch normalization for all the convolutional layers. For the learning rate, we start in 0.001 and we apply ReduceLROnPlateau with patience of 5 while monitoring the validation loss. We get that the learning rate value where the validation loss is at the lowest and the validation accuracy is high is 0.00001.

We see that generalization is supposedly better than training accuracy because the limitation that comes from the dataset such as emotion classes imbalance. Therefore, we try to push up the validation accuracy by allowing the training accuracy to get lower. Also, other findings worth mentioning ReduceLROnPlateau is really helping the model to get better generalization. The base model which can reach high training accuracy up 99.66% shows that the model is acutely overfit to the training data and the learning

rate must be adjusted. First, the optimized model is run in epoch of 50 and already reach 71.49% validation accuracy, we try to extend the epoch to 100 and reach the current validation accuracy.

Then the validation loss is also worth noting that it is the lowest among other compared research models that reported the validation loss. Our optimized model also has a higher F1-score. Lastly, we observed on each other research models, the models that use or includes text modality will certainly have boost on accuracy, as in their reports on their papers. The results on text modality are always quite high on accuracy.

5. CONCLUSION

In this paper we experiment on optimizing deep learning model for multimodal emotion recognition using IEMOCAP dataset from overfitting and focusing on two modalities, motion capture, and audio modalities. We try to utilize various hyperparameter tuning approaches such as dropout layer, l2 kernel regularization, batch normalization, and learning rate scheduler. Our optimized model can reach up to 73.67% validation accuracy and 0.73 f1-score but by sacrificing the training accuracy which is allowed because our focus is on making the model more generalized towards the data and especially when using only two modalities. Also, it is worth noting that our model runs on a smaller batch size of 48 and some of the research models run on bigger batch size of 64 which is quite a difference for the resource needed. Although our focus is only on making the model more generalized, there are limitations that can be improved in the future such as using more computer power to allow the experimentation on more sophisticated and powerful models. Other suggestions for future works are the usage of the hyperparameter tuning optimization could be experimented on model with three modalities (motion capture, audio, and text) and also new data augmentation methods can also be introduced and experimented with for IEMOCAP dataset in hopes that it can eliminate data inbalance to promote better model generalization.

ACKNOWLEDGEMENTS

The authors would like to thank and acknowledge the support of Bina Nusantara University. The support, guidance, and resources provided are very helpful in finishing this paper. Without them this paper would not have been completed successfully.

REFERENCES

[1] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowledge and Information Systems*, vol. 62, no. 8, pp. 2937–2987, Aug. 2020, doi: 10.1007/s10115-020-01449-0.

[2] D. Liu, Z. Wang, L. Wang, and L. Chen, "Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning," *Frontiers in Neurorobotics*, vol. 15, Jul. 2021, doi: 10.3389/fnbot.2021.697634.

[3] A. Landowska, M. Szwoch, W. Szwoch, M. R. Wróbel, and A. Kołakowska, "Emotion recognition and its applications," *Human-computer systems interaction: Backgrounds and applications*, vol. 3, pp. 51–62, 2014, doi: 10.1007/978-3-319-08491-6_5.

[4] S. Salman and X. Liu, "Overfitting Mechanism and Avoidance in Deep Neural Networks," *arXiv*, Jan. 2019, doi: 10.48550/arXiv.1901.06566.

[5] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. M. Sadeeq, and S. Zeebaree, "Multimodal Emotion Recognition using Deep Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 02, pp. 52–58, Apr. 2021, doi: 10.38094/jastt20291.

[6] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion Recognition from Multimodal Physiological Signals for Emotion Aware Healthcare Systems," *Journal of Medical and Biological Engineering*, vol. 40, no. 2, pp. 149-157, 2020, doi: 10.1007/s40846-019-00505-7.

[7] C. J. Yang, N. Fahier, W. C. Li, and W. C. Fang, "A Convolution Neural Network Based Emotion Recognition System using Multimodal Physiological Signals," in *2020 IEEE International Conference on Consumer Electronics - Taiwan, ICCE-Taiwan 2020*, 2020, pp. pp. 1-2, doi: 10.1109/ICCE-Taiwan49838.2020.9258341.

[8] S. Tripathi, S. Tripathi, and H. Beigi, "Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning," *arXiv*, Apr. 2018, doi: 10.48550/arXiv.1804.05788.

[9] M. Singh and Y. Fang, "Emotion Recognition in Audio and Video Using Deep Neural Networks," *arXiv*, Jun. 2020, doi: 10.48550/arXiv.2006.08129.

[10] W. Dai, S. Cahyawijaya, Z. Liu, and P. Fung, "Multimodal End-to-End Sparse Model for Emotion Recognition," in *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 2021, pp. 5305–5316, doi: 10.18653/v1/2021.naacl-main.417.

[11] J. Heredia *et al.*, "Adaptive Multimodal Emotion Detection Architecture for Social Robots," *IEEE Access*, vol. 10, pp. 20727-20744, 2022, doi: 10.1109/ACCESS.2022.3149214.

[12] Kenny and A. Chowanda, "Multimodal Approach for Emotion Recognition Using Feature Fusion," *ICIC Express Letters*, vol. 17, no. 2, pp. 181–189, Feb. 2023, doi: 10.24507/icicel.17.02.181.

[13] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176274-176285, 2020, doi: 10.1109/ACCESS.2020.3026823.

[14] B. Xie, M. Sidulova, and C. H. Park, "Article robust multimodal emotion recognition from conversation with transformer-based crossmodality the title fusion," *Sensors*, vol. 21, no. 14, 2021, doi: 10.3390/s21144913.

[15] P. P. Ippolito, "Hyperparameter Tuning: The Art of Fine-Tuning Machine and Deep Learning Models to Improve Metric Results," *Applied data science in tourism: Interdisciplinary approaches, methodologies, and applications*, pp. 231-251, 2022, doi: 10.1007/978-3-030-88389-8_12.
[16] L. Liao, H. Li, W. Shang, and L. Ma, "An Empirical Study of the Impact of Hyperparameter Tuning and Model Optimization on the Performance Properties of Deep Neural Networks," *ACM Transactions on Software Engineering and Methodology*, vol. 31, no. 3, 2022, doi: 10.1145/3506695.
[17] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295-316, 2020, doi: 10.1016/j.neucom.2020.07.061.
[18] T. Yu and H. Zhu, "Hyper-Parameter Optimization: A Review of Algorithms and Applications," *arXiv*, Mar. 2020, doi: 10.48550/arXiv.2003.05689.
[19] M. Kaveh and M. S. Mesgari, "Application of Meta-Heuristic Algorithms for Training Neural Networks and Deep Learning Architectures: A Comprehensive Review," *Neural Processing Letters*, vol. 55, no. 4, pp. 4519–4622, Aug. 2023, doi: 10.1007/s11063-022-11055-6.
[20] H. Lim, "A study on dropout techniques to reduce overfitting in deep neural networks," in *Lecture Notes in Electrical Engineering*, pp. 133-139, 2021, doi: 10.1007/978-981-15-9309-3_20.
[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929-1958, 2014.
[22] A. Thakkar and R. Lohiya, "Analyzing fusion of regularization techniques in the deep learning-based intrusion detection system," *International Journal of Intelligent Systems*, vol. 36, no. 12, pp. 7340-7388, 2021, doi: 10.1002/int.22590.
[23] Y. Li, X. Ren, F. Zhao, and S. Yang, "A zeroth-order adaptive learning rate method to reduce cost of hyperparameter tuning for deep learning," *Applied Sciences (Switzerland)*, vol. 11, no. 21, 2021, doi: 10.3390/app112110184.
[24] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, no. 1, pp. 1-74, Dec. 2021, doi: 10.1186/S40537-021-00444-8.
[25] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008, doi: 10.1007/S10579-008-9076-6/METRICS.

BIOGRAPHIES OF AUTHORS

Fernandi Widardo is a student of Master Track Information Technology in Bina Nusantara University (Indonesia, now). He is currently working on his graduation research paper thesis. His research interest is in machine learning, deep learning, and human signal processing. He can be contacted at email: fernandi.widardo@binus.ac.id.

Andry Chowanda earned his Bachelor's degree in Computer Science from Bina Nusantara University (Indonesia, 2009), a Master's in Business Management from BINUS Business School (Indonesia, 2011), and a Ph.D. in Computer Science from Nottingham University (England, 2017). He is now a Computer Science Lecturer at Bina Nusantara University. His research is in agent architecture and machine (and deep) learning. His work is mainly on how to model an agent that can sense and perceive the environment based on the perceived data and build a social relationship with the user over time. In addition, he is also interested in serious game and gamification design. He can be contacted at email: achowanda@binus.edu.