Buletin Ilmiah Sarjana Teknik Elektro
Vol. 6, No. 1, March 2024, pp. 83-90
ISSN: 2685-9572, DOI: 10.12928/biste.v6i1.10120

Performance Evaluation of Deep Learning Techniques in Gesture Recognition Systems

Gregorius Airlangga
Information System Study Program, Universitas Katolik Indonesia Atma Jaya, Indonesia

ARTICLE INFORMATION

Article History:

Submitted 23 February 2024
Revised 23 March 2024
Accepted 23 April 2024

Keywords:

Gesture Recognition;
Deep Learning Variants;
Machine Learning;
Human-Computer Interaction;
Deep Learning Comparison

Corresponding Author:

Gregorius Airlangga,
Universitas Katolik Indonesia Atma Jaya, Jakarta, Indonesia.
Email: gregorius.airlangga@atmajaya.ac.id

ABSTRACT

Tabel 1. Comparison Results

Methods	Accuracy	Precision	Recall	F1
LSTM	0.95	0.95	0.95	0.95
GRU	0.95	0.96	0.95	0.95
CNN	0.93	0.93	0.93	0.93
RNN	0.93	0.93	0.93	0.93
MLP	0.60	0.59	0.60	0.59
BILSTM	0.95	0.95	0.95	0.95
TCN	0.95	0.95	0.95	0.95

As human-computer interaction becomes increasingly sophisticated, the significance of gesture recognition systems has expanded, impacting diverse sectors such as healthcare, smart device interfacing, and immersive gaming. This study conducts an in-depth comparison of seven cutting-edge deep learning models to assess their capabilities in accurately recognizing gestures. The analysis encompasses Long Short-Term Memory Networks (LSTMs), Gated Recurrent Units (GRUs), Convolutional Neural Networks (CNNs), Simple Recurrent Neural Networks (RNNs), Multi-Layer Perceptrons (MLPs), Bidirectional LSTMs (BiLSTMs), and Temporal Convolutional Networks (TCNs). Evaluated on a dataset representative of varied human gestures, the models were rigorously scored based on accuracy, precision, recall, and F1 metrics, with LSTMs, GRUs, BiLSTMs, and TCNs outperforming others, achieving an impressive score bracket of 0.93 to 0.95. Conversely, MLPs trailed with scores around 0.59 to 0.60, underscoring the challenges of non-temporal models in processing sequential data. This study pinpoints model selection as pivotal for optimal system performance and suggests that recognizing the temporal patterns in gesture sequences is crucial. Limitations such as dataset diversity and computational demands were noted, emphasizing the need for further research into models' operational efficiencies. Future studies are poised to explore hybrid models and real-time processing, with the prospect of enhancing gesture recognition systems' interactivity and accessibility. This research thus provides a foundational benchmark for selecting and implementing the most suitable computational methods for advancing gesture recognition technologies.

Document Citation:

G. Airlangga, "Performance Evaluation of Deep Learning Techniques in Gesture Recognition Systems," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 6, no. 1, pp. 83-90, 2024, DOI: 10.12928/biste.v6i1.10120.

1. INTRODUCTION

In the vanguard of the digital revolution, gesture recognition technology emerges as an essential interface, crafting more organic interactions between humans and machines [1]–[3]. As an integral facet of applications ranging from immersive virtual environments to the nuanced control of smart home devices, this technology also serves critical roles in healthcare, allowing for contactless monitoring systems, and in education, facilitating adaptive learning experiences [4][5]. The evolution of gesture recognition technology has been rapid, thanks to advances in sensor technologies and computational methods [6]. Then, the successful deployment of gesture recognition within these sectors exemplifies its transformative potential. However, the journey towards fully intuitive systems is fraught with challenges: the inherent variability in human gestures, the nuanced interpretation of complex movements, and the stringent demands of real-time operation within computational constraints [7]. The challenges are manifold, including the variability in human gestures, the complexity of capturing and interpreting subtle movements [8], and the need for systems to operate in real-time with minimal computational resources [9]. Deep learning has catalyzed a paradigm shift in gesture recognition, equipping models to autonomously decipher intricate data patterns [10]. Traditional machine learning methods, such as Decision Trees [11], Support Vector Machines (SVMs) [7], and Random Forests, have been pivotal in early studies, offering solid frameworks for understanding gesture data through handcrafted features [12]. These methods, while effective in certain contexts, often fall short when dealing with the high-dimensional, temporal nature of gesture data, necessitating extensive preprocessing and feature extraction efforts [13].

The advent of deep learning has revolutionized gesture recognition, introducing models capable of learning complex, hierarchical representations of data directly from raw inputs [14]. Convolutional Neural Networks (CNNs) have shown exceptional prowess in extracting spatial features from static images and sequences of frames, making them a popular choice for image-based gesture recognition tasks [10]. Recurrent Neural Networks (RNNs), including their more sophisticated variants like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), have become the go-to models for capturing temporal dynamics in gesture sequences [15]. Despite their success, these models demand substantial computational resources and large volumes of labeled data, posing challenges for deployment in resource-constrained environments [16]. The literature also reveals an emerging interest in hybrid models and novel architectures, such as Temporal Convolutional Networks (TCNs) and Bidirectional LSTMs, aiming to combine the strengths of existing models while mitigating their weaknesses [17]. TCNs, for instance, offer an attractive alternative to RNNs for handling sequential data, providing comparable or even superior performance with the added benefits of parallelism and a flexible receptive field [18]. However, the exploration of TCNs in gesture recognition is still in its nascent stages, with much potential for discovery and application [19]. The urgency for advancing gesture recognition technology is driven by the increasing demand for more sophisticated human-computer interaction modalities across various sectors.

In healthcare, for instance, gesture recognition can enable touchless control systems, reducing the risk of infection transmission [20]. In education, it can facilitate interactive learning environments that cater to diverse learning needs [21]. The state-of-the-art in gesture recognition is characterized by a rapid adoption of deep learning models, which have significantly pushed the boundaries of what's possible, achieving unprecedented levels of accuracy and efficiency [22]. Yet, the quest for models that can quickly adapt to new gestures, perform reliably across different users and environments, and operate in real-time on low-power devices remains [23]. This highlights the need for innovative approaches that not only leverage the full potential of deep learning but also address its inherent limitations [24]. This research aims to bridge the gap in gesture recognition by conducting a comprehensive comparison of various computational models' performance, ranging from traditional machine learning to the latest deep learning architectures [25]. By evaluating models such as CNNs, LSTMs, GRUs, Simple RNNs, MLPs, Bidirectional LSTMs, and TCNs on a unified dataset, this study seeks to uncover insights into the optimal strategies for gesture recognition [26]. Beyond mere performance comparison, the research investigates the models' computational efficiency, robustness to variations in gesture execution, and adaptability to new gestures, providing a holistic view of their applicability in real-world scenarios [27]. Despite the wealth of research on gesture recognition, a comprehensive analysis comparing a broad spectrum of computational models is conspicuously absent [14]. Most studies focus on a limited range of models or specific aspects of gesture recognition, such as accuracy, neglecting other critical factors like computational efficiency and generalizability.

This research gap impedes the development of gesture recognition systems that are not only accurate but also practical for real-world applications [28]–[30]. By addressing this gap, this study aims to provide a more nuanced understanding of the trade-offs involved in selecting and deploying gesture recognition models, guiding future research and development efforts.This article makes several significant contributions to the field of gesture recognition. Firstly, it presents one of the first extensive comparative analyses of seven different computational models, shedding light on their performance nuances in the context of gesture recognition.

Secondly, it introduces a novel preprocessing pipeline that significantly enhances model performance, demonstrating the importance of effective data preparation. Lastly, the findings from this study offer valuable guidelines for researchers and practitioners in selecting the most suitable models for their specific needs, potentially accelerating the development of more advanced and user-friendly gesture recognition systems. Following this introduction, Section 2 details the methodology, including the dataset, data preprocessing techniques, model architectures, and evaluation criteria. Section 3 presents a comprehensive analysis of the experimental results, highlighting key findings and their implications for gesture recognition research. In addition, we also discuss the broader implications of these findings, exploring their relevance to current challenges and future directions in the field. The article concludes with Section 4, which summarizes the study's contributions and outlines avenues for future research, emphasizing the ongoing need for innovation in gesture recognition technologies.

2. METHODS

2.1. Data Preparation

The study utilizes a unified gesture dataset [31] comprising a diverse range of gestures captured through motion sensors or vision-based systems. Each gesture is represented by a series of frames, with each frame encapsulating spatial and/or temporal features relevant to gesture dynamics. The dataset includes four primary gesture categories: 'rock', 'paper', 'scissors', and 'ok', with each category containing several hundred instances to ensure statistical significance. The gestures were performed by participants from varied demographic backgrounds to introduce diversity in gesture execution styles. This variability is crucial for assessing the robustness and generalizability of the models under comparison.

2.2. Data Preprocessing

Preprocessing plays a pivotal role in preparing the raw gesture data for model training and evaluation. Initially, the dataset undergoes a cleaning process to remove any noisy or incomplete instances, ensuring data quality. Subsequently, feature scaling is applied using the MinMaxScaler to normalize the feature values across all samples, enhancing model convergence during training. For deep learning models that require specific input shapes (e.g., CNNs and RNNs), the data is reshaped accordingly. For instance, data intended for CNNs is reshaped into 2D arrays representing image frames, while for RNNs, sequences are maintained in their temporal form. Additionally, a novel preprocessing step introduced in this study involves augmenting the dataset with synthetic gestures generated through slight modifications of existing samples, aiming to increase the robustness of models to variations in gesture execution.

2.3. Model Architecture

In this study, we undertake a comprehensive comparison of seven distinct computational models, each uniquely tailored for the task of gesture recognition, reflecting the diversity and complexity inherent in interpreting human gestures through machine learning technologies. The first model we explore is the Convolutional Neural Network (CNN) as presented in the equation (1), which is built upon multiple layers of convolutional filters that are adept at extracting spatial hierarchies from gesture data. This architecture, complemented by dense layers for classification, excels in processing image-based representations of gestures, making it a cornerstone in the field of gesture recognition. Following the CNN, we delve into the domain of sequential data processing with Long Short-Term Memory Networks (LSTMs) as presented in the equation (2) – equation (7). LSTMs are engineered to recognize and retain long-term dependencies within sequential data through a sophisticated system of memory cells. This capability renders them particularly effective for time-series gesture data analysis, where the chronological sequence of movements is paramount for accurate recognition.

Gated Recurrent Units (GRUs) are then examined as a simpler, yet potent, alternative to LSTMs. GRUs as presented in the equation (8) – equation (11) streamline the gating mechanism involved in processing sequential data, potentially offering a more efficient route for gesture recognition tasks without compromising on the ability to capture temporal dependencies. The study also evaluates Simple Recurrent Neural Networks (Simple RNNs), representing the foundational form of RNNs. Despite their ability to capture the temporal dynamics in gesture sequences, Simple RNNs face challenges with long-term dependencies, highlighting the trade-offs between model complexity and performance. Turning to Multi-Layer Perceptrons (MLPs), we investigate this feedforward neural network's prowess in classifying gestures from flattened data representations. MLPs concentrate on deciphering the intricate relationship between input features and gesture categories, showcasing the versatility of neural networks in handling diverse data formats.

Bidirectional LSTMs (Bi-LSTMs) as presented in the equation (12) are also featured in our comparison, with their dual-direction data processing capability. This approach enhances the model's understanding of

context within gesture sequences, promising improvements in recognition accuracy by leveraging information from both past and future states. Lastly, Temporal Convolutional Networks (TCNs) as presented in the equation (13) are evaluated for their innovative use of dilated convolutions to process sequential data efficiently across extended sequences. TCNs present a compelling alternative to conventional RNNs, aiming to capture temporal patterns in gesture data with greater efficiency and potentially superior performance. Each model in this study is meticulously implemented with an architecture optimized for gesture recognition, incorporating specific layer sizes, activation functions, and other hyperparameters. These configurations are informed by both preliminary experiments and a thorough review of existing literature, ensuring a robust foundation for evaluating and comparing the effectiveness of these diverse computational models in understanding and classifying human gestures.

$$f_{i,j}^{(l)} = \sigma\left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} w_{m,n}^{(l)} \cdot x_{i+m,j+n}^{(l-1)} + b^{(l)}\right) \tag{1}$$

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big) \tag{2}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4}$$

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C_t} \tag{5}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t \times \tanh(C_t) \tag{7}$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{8}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{9}$$

$$\widehat{h_t} = \tanh(W \cdot [r_t \times h_{t-1}, x_t]) \tag{10}$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \widehat{h_t} \tag{11}$$

$$y = \sigma(W_n \cdots \sigma(W_2\sigma(W_1 x + b_1) + b_2) \cdots + b_n) \tag{12}$$

$$F(s) = \sum_{i=0}^{k-1} f(i) \cdot x(s - d \cdot i) \tag{13}$$

2.4. Training Procedure

Models are trained using the prepared dataset, split into training and testing sets to evaluate generalization performance. The training process for each model involves optimizing a loss function specific to the gesture recognition task, typically using variants of cross-entropy for classification tasks. Models are trained using the

Adam optimizer, a popular choice for deep learning tasks due to its adaptive learning rate properties. Training is conducted over a fixed number of epochs, with early stopping implemented to prevent overfitting based on the validation set performance. Additionally, dropout layers are incorporated into the models as needed to further mitigate overfitting by randomly omitting a subset of neurons during training.

2.5. Evaluation Metrics

The performance of each model is evaluated using a combination of metrics suited for classification tasks, including accuracy, precision, recall, and F1-score. Accuracy provides a straightforward measure of overall performance, while precision, recall, and F1-score offer deeper insights into the models' ability to correctly recognize gestures across different categories. The evaluation extends to computational efficiency, assessing models based on their training time and resource consumption, which are critical factors for real-world deployment.

3. RESULT AND DISCUSSION

The Table 1 presented showcases a comparative analysis of several machine learning models applied to gesture recognition, evaluated based on four key performance metrics: Accuracy, Precision, Recall, and F1 Score. These metrics collectively provide insights into the models' classification strengths and weaknesses. At the outset, it's observed that the LSTM, GRU, BiLSTM, and TCN models demonstrate exemplary performance, uniformly achieving 0.95 across all metrics. This uniformity suggests that these models are not only accurate overall (as reflected in the Accuracy metric) but also show a high level of reliability in their predictions (Precision), a strong ability to identify relevant instances (Recall), and a balanced harmonic mean of Precision and Recall (F1 Score). The high scores can be attributed to these models' architectural strengths in processing sequential and temporal data, which is intrinsic to gesture recognition. LSTMs and BiLSTMs, with their memory cells and bidirectional processing, respectively, are adept at capturing long-range dependencies within the gesture sequences. GRUs, with a more streamlined structure, still capture essential temporal features without the complexity of LSTMs, which might contribute to their equivalent performance. Similarly, TCNs leverage dilated convolutions to effectively manage temporal hierarchies in data, affirming their suitability for time-series analysis.

Tabel 1. Comparison Results

Methods	Accuracy	Precision	Recall	F1
LSTM	0.95	0.95	0.95	0.95
GRU	0.95	0.96	0.95	0.95
CNN	0.93	0.93	0.93	0.93
RNN	0.93	0.93	0.93	0.93
MLP	0.60	0.59	0.60	0.59
BILSTM	0.95	0.95	0.95	0.95
TCN	0.95	0.95	0.95	0.95

Conversely, CNNs and simple RNNs display a marginally lower performance with scores of 0.93, which, while commendable, indicate certain limitations. The CNNs, primarily renowned for their spatial feature extraction capabilities, may not fully encapsulate the temporal aspect of gesture sequences, possibly accounting for the slight dip in performance compared to models that specialize in sequence data. Simple RNNs, although designed for temporal data, are known to falter with long-term dependencies, which may explain their inability to match the performance of their more advanced counterparts. The MLP model lags significantly behind the others, with its scores hovering around the 0.60 mark. This stark contrast underscores the challenges faced by traditional feedforward architectures in handling the complexities of gesture data. Without the mechanisms to process the temporal sequences inherent in gestures, the MLP struggles to achieve the high standards set by recurrent and convolutional architectures.

The F1 Score is particularly revealing, as it is a measure that conveys the balance between Precision and Recall. The high F1 Scores achieved by LSTMs, GRUs, BiLSTMs, and TCNs emphasize their capability to maintain this balance, making them robust choices for applications where misclassifications can be costly. The modest F1 Scores for CNNs and RNNs suggest that while these models are still quite capable, they may not be as reliable as the others when Precision and Recall are equally important. The low F1 Score for the MLP further highlights its limitations in the context of gesture recognition. In essence, the results indicate a clear hierarchy in model performance, with sequence-processing models at the top, spatial-feature-oriented and simple temporal models in the middle, and the traditional, non-sequential MLP at the bottom. This analysis not only affirms the importance of architectural alignment with the nature of the task but also suggests areas for potential

improvement in model design and application. For instance, enhancements to CNNs that integrate temporal processing could bridge the performance gap, while innovations in MLP structures might better capture the complexities of gesture data. These insights are crucial for advancing the field of gesture recognition, guiding future research toward the development of more sophisticated and specialized models.

4. Conclusion

The research presented in this article aimed to critically evaluate and compare the performance of various computational models in the domain of gesture recognition. Through rigorous experimentation and analysis, the study has yielded insightful conclusions that not only enhance our understanding of the capabilities of these models but also illuminate the path forward for future explorations in the field. The findings reveal a noteworthy disparity in the performance of the examined models. Long Short-Term Memory Networks (LSTMs), Gated Recurrent Units (GRUs), Bidirectional LSTMs (BiLSTMs), and Temporal Convolutional Networks (TCNs) have demonstrated exceptional proficiency, as evidenced by their uniformly high accuracy, precision, recall, and F1 scores. This success can be largely attributed to their inherent design, which effectively captures the temporal dependencies and nuances present in gesture data. Their ability to model sequences makes them particularly adept for tasks that require an understanding of the context and progression of movements over time. In contrast, the Multi-Layer Perceptron (MLP) model exhibited a considerable decline in performance. This outcome underscores the significance of model selection in accordance with the nature of the dataset and the task at hand. MLPs, with their feedforward architecture, are not naturally equipped to handle the sequential and spatial complexities of gesture recognition, leading to their diminished effectiveness as reflected in the lower metric scores. Furthermore, the slightly lower scores of Convolutional Neural Networks (CNNs) and Simple Recurrent Neural Networks (RNNs) suggest that while they hold potential, there may be room for optimization in their architectures or training processes to fully harness their capabilities for gesture recognition tasks. The implications of these results are profound for the development of interactive technologies that require robust and accurate gesture recognition. The superior models identified by this research could be employed to enhance user experience in various applications, from virtual reality to assistive technologies, ensuring more seamless and natural human-computer interactions. For future work, it is recommended to explore hybrid models that combine the strengths of the high-performing architectures identified in this study. Such models could potentially address any existing limitations and push the boundaries of gesture recognition technology further. Additionally, investigating the impact of larger and more diverse datasets on the performance of these models would provide deeper insights into their scalability and adaptability to real-world scenarios. Finally, the exploration of real-time processing capabilities and the implementation of these models in edge devices present exciting avenues for research, promising to bring gesture-based interaction closer to ubiquitous adoption.

REFERENCES

[1] D. Sayers *et al.*, "The Dawn of the Human-Machine Era: A forecast of new and emerging language technologies," hal-03230287, 2021, https://doi.org/10.17011/jyx/reports/20210518/1.

[2] Z. Ding, Y. Ji, Y. Gan, Y. Wang, and Y. Xia, "Current status and trends of technology, methods, and applications of Human--Computer Intelligent Interaction (HCII): A bibliometric research," *Multimed. Tools Appl.*, pp. 1–34, 2024, https://doi.org/10.1007/s11042-023-18096-6.

[3] M. Tzampazaki, C. Zografos, E. Vrochidou, and G. A. Papakostas, "Machine Vision—Moving from Industry 4.0 to Industry 5.0," *Applied Sciences*, vol. 14, no. 4, p. 1471, 2024, https://doi.org/10.3390/app14041471.

[4] C. Holloway and G. Barbareschi, *Disability interactions: creating inclusive innovations*. Springer Nature, 2022, https://doi.org/10.1007/978-3-031-03759-7.

[5] A. Taghian, M. Abo-Zahhad, M. S. Sayed, and A. H. Abd El-Malek, "Virtual and augmented reality in biomedical engineering," *Biomed. Eng. Online*, vol. 22, no. 1, p. 76, 2023, https://doi.org/10.1186/s12938-023-01138-3.

[6] S. Berman and H. Stern, "Sensors for gesture recognition systems," *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.*, vol. 42, no. 3, pp. 277–290, 2011, https://doi.org/10.1109/TSMCC.2011.2161077.

[7] D. Sarma and M. K. Bhuyan, "Methods, databases and recent advancement of vision-based hand gesture recognition for hci systems: A review," *SN Comput. Sci.*, vol. 2, no. 6, p. 436, 2021, https://doi.org/10.1007/s42979-021-00827-x.

[8] S. Wang *et al.*, "Hand gesture recognition framework using a lie group based spatio-temporal recurrent network with multiple hand-worn motion sensors," *Inf. Sci. (Ny).*, vol. 606, pp. 722–741, 2022, https://doi.org/10.1016/j.ins.2022.05.085.

[9] J. Zhao, X. Sun, Q. Li, and X. Ma, "Edge caching and computation management for real-time internet of vehicles: An online and distributed approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 4, pp. 2183–2197, 2020, https://doi.org/10.1109/TITS.2020.3012966.

[10] Z. Wang *et al.*, "A study on hand gesture recognition algorithm realized with the aid of efficient feature extraction method and convolution neural networks: design and its application to VR environment," *Soft Comput.*, pp. 1–24, 2023, https://doi.org/10.1007/s00500-023-09077-w.
[11] Y. L. Coelho, F. de A. S. dos Santos, A. Frizera-Neto, and T. F. Bastos-Filho, "A lightweight framework for human activity recognition on wearable devices," *IEEE Sens. J.*, vol. 21, no. 21, pp. 24471–24481, 2021, https://doi.org/10.1109/JSEN.2021.3113908.
[12] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognit.*, vol. 108, p. 107561, 2020, https://doi.org/10.1016/j.patcog.2020.107561.
[13] S. Ahmed, K. D. Kallu, S. Ahmed, and S. H. Cho, "Hand gestures recognition using radar sensors for human-computer-interaction: A review," *Remote Sens.*, vol. 13, no. 3, p. 527, 2021, https://doi.org/10.3390/rs13030527.
[14] Z. Lv, F. Poiesi, Q. Dong, J. Lloret, and H. Song, "Deep learning for intelligent human--computer interaction," *Appl. Sci.*, vol. 12, no. 22, p. 11457, 2022, https://doi.org/10.3390/app122211457.
[15] J. V Tembhurne and T. Diwan, "Sentiment analysis in textual, visual and multimodal inputs using recurrent neural networks," *Multimed. Tools Appl.*, vol. 80, pp. 6871–6910, 2021, https://doi.org/10.1007/s11042-020-10037-x.
[16] V. Kamath and A. Renuka, "Deep Learning Based Object Detection for Resource Constrained Devices-Systematic Review, Future Trends and Challenges Ahead," *Neurocomputing*, 2023, https://doi.org/10.1016/j.neucom.2023.02.006.
[17] Z. Ma and G. Mei, "A hybrid attention-based deep learning approach for wind power prediction," *Appl. Energy*, vol. 323, p. 119608, 2022, https://doi.org/10.1016/j.apenergy.2022.119608.
[18] Y. Shen, J. Wang, C. Feng, and Q. Wang, "Dual attention-based deep learning for construction equipment activity recognition considering transition activities and imbalanced dataset," *Autom. Constr.*, vol. 160, p. 105300, 2024, https://doi.org/10.1016/j.autcon.2024.105300.
[19] Y. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Inf. Fusion*, p. 101869, 2023, https://doi.org/10.1016/j.inffus.2023.101869.
[20] C. Eze and C. Crick, "Learning by Watching: A Review of Video-based Learning Approaches for Robot Manipulation," *arXiv Prepr. arXiv2402.07127*, 2024, https://doi.org/10.48550/arXiv.2402.07127.
[21] E. Murphy. *In Service to Security: Constructing the Authority to Manage European Border Data Infrastructures*. Copenhagen Business School [Phd]. 2023. https://research.cbs.dk/en/publications/in-service-to-security-constructing-the-authority-to-manage-europ.
[22] H. Zhou *et al.*, "Deep-learning-assisted noncontact gesture-recognition system for touchless human-machine interfaces," *Adv. Funct. Mater.*, vol. 32, no. 49, p. 2208271, 2022, https://doi.org/10.1002/adfm.202208271.
[23] T. Valtonen *et al.*, "Learning environments preferred by university students: A shift toward informal and flexible learning environments," *Learn. Environ. Res.*, vol. 24, pp. 371–388, 2021, https://doi.org/10.1007/s10984-020-09339-6.
[24] L. A. Al-Haddad, W. H. Alawee, and A. Basem, "Advancing task recognition towards artificial limbs control with ReliefF-based deep neural network extreme learning," *Comput. Biol. Med.*, vol. 169, p. 107894, 2024, https://doi.org/10.1016/j.compbiomed.2023.107894.
[25] N. Schizas, A. Karras, C. Karras, and S. Sioutas, "TinyML for Ultra-Low Power AI and Large Scale IoT Deployments: A Systematic Review," *Futur. Internet*, vol. 14, no. 12, p. 363, 2022, https://doi.org/10.3390/fi14120363.
[26] X. Xu *et al.*, "Enabling hand gesture customization on wrist-worn devices," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2022, https://doi.org/10.1145/3491102.3501904.
[27] W. Li *et al.*, "A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: Theories, applications and challenges," *Mech. Syst. Signal Process.*, vol. 167, p. 108487, 2022, https://doi.org/10.1016/j.ymssp.2021.108487.
[28] R. E. Nogales and M. E. Benalcázar, "Hand gesture recognition using machine learning and infrared information: a systematic literature review," *Int. J. Mach. Learn. Cybern.*, vol. 12, no. 10, pp. 2859–2886, 2021, https://doi.org/10.1007/s13042-021-01372-y.
[29] H. Mohyuddin, S. K. R. Moosavi, M. H. Zafar, and F. Sanfilippo, "A comprehensive framework for hand gesture recognition using hybrid-metaheuristic algorithms and deep learning models," *Array*, vol. 19, p. 100317, 2023, https://doi.org/10.1016/j.array.2023.100317.
[30] M. Lee and J. Bae, "Real-time gesture recognition in the view of repeating characteristics of sign languages," *IEEE Trans. Ind. Informatics*, vol. 18, no. 12, pp. 8818–8828, 2022, https://doi.org/10.1109/TII.2022.3152214.
[31] M. Simão, P. Neto, and O. Gibaru, "EMG-based online classification of gestures with recurrent neural networks," *Pattern Recognition Letters*, vol. 128, pp. 45-51, 2019, https://doi.org/10.1016/j.patrec.2019.07.021.

AUTHOR BIOGRAPHY

GREGORIUS AIRLANGGA is Received the B.S. degree in information system from the Yos Sudarso Higher School of Computer Science, Purwokerto, Indonesia, in 2014, and the M.Eng. degree in informatics from Atma Jaya Yogyakarta University, Yogyakarta, Indonesia, in 2016. He got Ph.D. degree with the Department of Electrical Engineering, National Chung Cheng University, Taiwan. He is also an Assistant Professor with the Department of Information System, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia. His research interests include data science, artificial intelligence and software engineering include path planning, machine learning, natural language processing, deep learning, software requirements, software design pattern and software architecture.