ICIC Express Letters
Volume **17**, Number **7**, July **2023**

ICIC International ©2023 ISSN 1881-803X
pp. 809–816

DEAF SIGN LANGUAGE TRANSLATION SYSTEM WITH POSE AND HAND GESTURE DETECTION UNDER LSTM-SEQUENCE CLASSIFICATION MODEL

Ridwang[1,3], Ingrid Nurtanio[2], Amil Ahmad Ilham[2,*] and Syafaruddin[1]

[1]Department of Electrical Engineering
[2]Department of Informatics
Universitas Hasanuddin
Jalan Poros Malino Km. 6, Gowa 92171, Indonesia
ridwang17d@student.unhas.ac.id; { ingrid; syafaruddin }@unhas.ac.id
*Corresponding author: amil@unhas.ac.id

[3]Department of Electrical Engineering
Universitas Muhammadiyah Makassar
Jl. Sultan Alauddin No. 259, Makassar 90221, Indonesia
ridwang@unismuh.ac.id

Received July 2022; accepted October 2022

Abstract. *Hand gesture recognition (HGR) is a fundamental mode of human communication and engagement. While HGR can be used to improve user interaction in human-computer interaction (HCI), it can also be used to overcome linguistic barriers. For example, HGR can be used to recognize sign language, which is a visual language expressed by hand movements and used as a fundamental mode of communication by the deaf and mute all over the world. The goal of this study is to create a novel way for detecting hand gestures in deaf sign language translation systems. To recognize dynamic hand gestures, a combination of the long short-term memory (LSTM) approach and a sequential method is used. In this study, nine dynamic gestures that are suited to the context were devised to solve the challenge of gesture recognition. Sequence and image processing data were collected using MediaPipe Holistic, preprocessed, and trained using an LSTM method. The model was practiced using training and validation data, and the test set was utilized to evaluate it. The experimental result revealed that the accuracy of the proposed model was 98 percent for nine kinds of gesture testing.*

Keywords: LSTM, Gesture, MediaPipe, Language, Sign

1. **Introduction.** A gesture is a type of nonverbal or non-vocal communication in which visible gestures are used instead of, or in addition to, words to express certain messages. Gesture is one of the most essential human communication tools for efficiently expressing user requests [1]. According to [2], the hand gestures are recognized as a natural and fundamental human contact and communication technique, as they have been used to transmit information even before the invention of language. Hand gestures allow information to be transmitted and complex procedures to be completed with ease using a series of hand movements and finger placements. Consequently, hand gestures can be used as a highly adaptable interface for human-computer interaction (HCI), allowing interaction to be expedited by eliminating physical touch between the mechanical device and its users.

In addition, according to [2,3], hand gestures are the major means of communication for those who are deaf or have hearing loss. They were applied with communicative gestures as an implicit metric of embodiment [4]. Hand gestures are essentially a sign language for communication [5]. People who are deaf or have hearing loss face challenges in communicating with the broader public. In order to understand the meaning of sign language

DOI: 10.24507/icicel.17.07.809

movement, hand gesture is applied in formal and casual communication. As a result, establishing a hand gesture detection system is a must because it may be used to break through communication barriers.

In general, there are two types of hand gesture recognition: static and dynamic [2,6]. According to [6], the static hand gesture recognition is concerned with a picture form representation of hand gestures. In this case, typical techniques like template matching can be used to complete the recognition process. Dynamic hand gesture is a method for extracting the skeleton sequence's temporal dynamic properties [7]. A dynamic gesture is one in which the posture of a finger, palm, or arm changes over time [1]. The following methods are used to recognize dynamic hand gestures [8]: the convolutional neural networks (CNN) [9], the recurrent neural networks (RNN) [10], long short-term memory (LSTM) [11]; 3D-convolutional neural network (3D-CNN) [12] and deep gesture recognition utility (DeepGRU) [13]. This approach cannot directly train for a set of all data. Before training, the data extraction for this approach still translates the input video type into a frame.

Recently, some studies have investigated LSTM methods to detect hand gestures as sign language. [14] applied a convolutional LSTM recurrent neural network (called CNN-LSTM) in gesture recognition. This model looked at qualitative evaluation based on internal representation visualization and examined temporal classification output at a frame level to see if it matched the cognitive sense of a gesture. Finally, this research demonstrates that in gesture recognition, CNNLSTM beats both plain CNN and LSTM. Authors in [15] presented a new sensor-based continuous hand gesture detection algorithm that uses LSTM. This method is used to generate an output path from a sequence of sensory data inputs. The final classification results are then determined using a maximum a posteriori estimation according to the observed path. The performance study was carried out using a prototype system based on smartphones. According to the results, the proposed method is a viable alternative for reliable and precise hand-gesture recognition. [16] evaluated the effectiveness of recurrent deep learning network in classifying electromyograms (EMGs) because they can learn long-term and non-linear time series dynamics. In this study, LSTM was used to produce multiclass classification on six grip gestures performed by nine amputees at three distinct force levels using a neural network, such as low, medium and high. Our findings reveal that an LSTM-based neural network can deliver consistent results across all nine amputee and force levels, with an average categorization error rate of around nine per cent. We show that deep learning can be used to control upper-limb prostheses.

In addition, [12] investigated 3D-CNN to recognize dynamic hand gestures for Indian sign language (ISL) modeling. In this research, a dataset created by reproducing the movements of 20 gestures from conventional ISL to train this model. Consequently, the network model produced good results in terms of precision, accuracy, recall, and f1-scores. [17] investigated CNN to classify hand gesture recognition. This model recognizes hand motions in videos or images. The ResNeXt-101 model is used to classify hand motions, which makes it more sophisticated. As a result, training accuracy for gesture recognition ranged from 95% to 99%, and the testing accuracy was 94.35 per cent. Authors in [13] illustrated the Deep GRU method for gesture and action recognition. This study looked at the seventh publicly available dataset, which has a significant number of samples and covers a wide spectrum of interactions (full-body, multi-actor and hand motions). On cross-subject and cross-view tests of the NTU RGB+D dataset, it achieves recognition accuracy of 84.9 per cent and 92.3 per cent, respectively, and 100 per cent recognition accuracy on the UT-Kinect dataset.

However, none of the studies listed above used MediaPipe Holistic as a multi-stage pipeline that addresses distinct regions with appropriate picture resolution for each region. The main contributions of this paper are highlighted below.

1) The proposed method can be applied to large series dataset with higher accuracy and quickly processes without additional layers to input data in LSTM.
2) The proposed method can be applied to identifying dynamic hand gestures in deaf sign language translation systems with higher accuracy and faster extract data.

In this study, MediaPipe Holistic uses the position and hand landmark models to generate a total of 54 landmarks, including 33 pose landmarks and 21 hand landmarks. This work is split into four sections: Section 1 is the introduction; Section 2 outlines research method; Section 3 covers results and discussion, and Section 4 presents the conclusions.

2. Research Method.

2.1. **Architecture design.** Figure 1 demonstrates the architectural design of the proposed system. According to [18], learning both spatial and temporal information for dynamic gesture recognition is difficult using a handmade feature extraction method. We presented a model to address this problem, as shown in Figure 1.

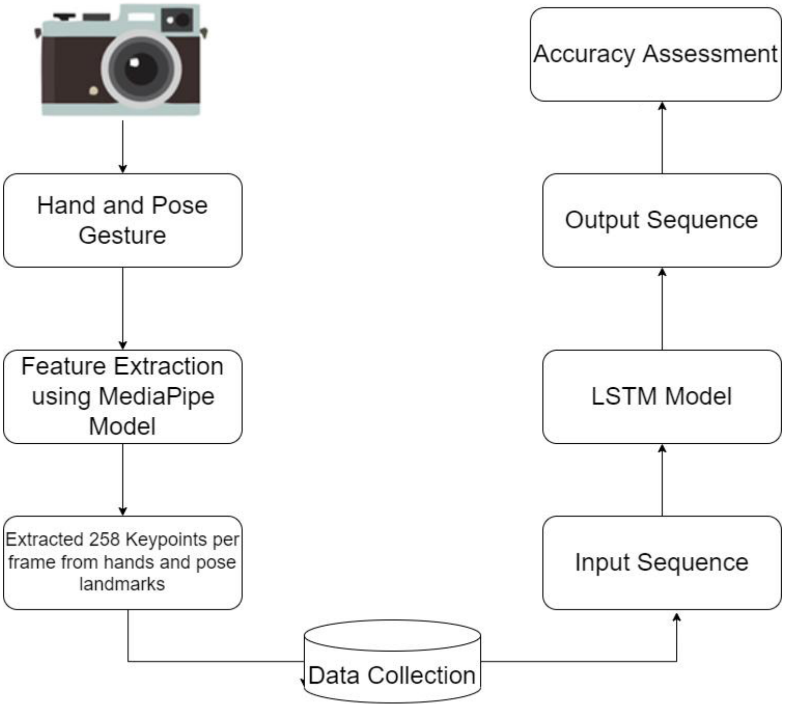

Hand and Pose Gesture
Feature Extraction using MediaPipe Model
Extracted 258 Keypoints per frame from hands and pose landmarks
Data Collection
Input Sequence
LSTM Model
Output Sequence
Accuracy Assessment

FIGURE 1. Architectural design of model

2.2. **Data collection.** In this step, we record the gestures of five different people to generate a varied dataset. Since all of these people are right-handed, users must follow several processes to remove bias and improve the dataset. Users are divided into five groups of two people before recording their movements. The motion sequence recorded by each user consists of a five second dynamic movement with 30 frames. Each move must be performed 10 times by the user. There is no need to start with hands in a resting or neutral position all the time (hands down). Each user must act differently for each attempt, adjusting the speed and position of the hands. However, during training data collection we found some invalid data. Before pre-processing, we eliminate invalid data to improve classification.

To speed up the data collection process, basic data collection tools were used, and the data collected consisted of motion sequences generated by real human actions. Furthermore, movements were captured using a camera with various training data collection scenarios: using left hand gestures and poses without right hands, right hand gestures and poses without left hands, and also using right hands together with left hands and poses. To incorporate modality data into the proposed model, a high-resolution webcam camera is used as the main tool for data recording. Hands and poses are extracted from the body using the camera and MediaPipe application as input to the model during the preprocessing step. The extraction results are in the form of key points from several right and left hands points and poses that are sorted into a dataset that has the same number as the number of frames generated from one movement.

2.3. **LSTM-sequence classification method.** The LSTM approach is a recurrent neural network that can represent sequential data and detect anomalous values effectively [19-21]. LSTM is a deep learning method that is the advancement of recurrent neural network (RNN), which has been shown in many studies to have a high ability in time-series forecasting and to handle problems in long-term dependency models [22,23]. In the recurrent layer of a standard LSTM, there are input gates, output gates, forget gates, and memory blocks. Memory blocks include memory cells that use self-connections to retain the network's temporal state, as well as gates that control the flow of information. The input gate regulates the flow of activations into memory cells. The output gate regulates the flow of cell activation output to the remainder of the network. The forget gate scales the cell's internal state before adding it as input to the cell via the self-recurrent link, allowing the cell's memory information to be adaptively forgotten.

To verify accurately all data, sequence classification is applied to validating and recognizing data output from the LSTM. Figure 2 (below) illustrates the proposed LSTM-sequence classification. The proposed method uses five layers network composed of three LSTM layers and two dense layers to consolidate input from the third LSTM layers to final predictive value. Layers 1, 2, 3, 4 and 5 contain 64, 128, 64, 64 and 32 units, respectively. In order to define the output of classification, the softmax function is frequently utilized as the final activation function of a network.

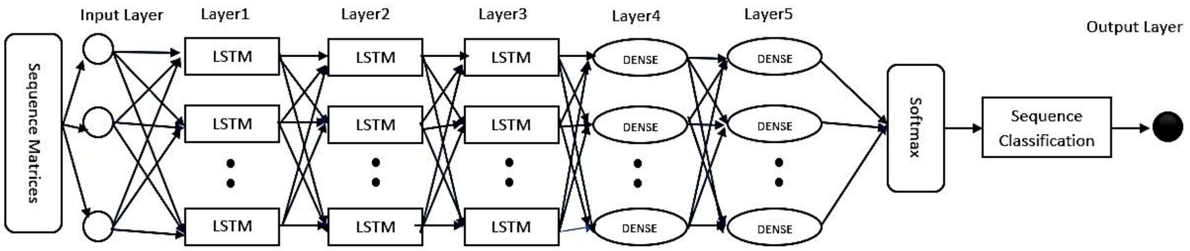


FIGURE 2. The proposed LSTM-sequence classification

2.4. **Model validation.** For training and validation, 80 per cent and 20 per cent of the dataset would be used in each trial, respectively. The data set was separated into five groups (folds), and five trials were carried out. One of the folds was designated as the testing set for each trial, while the others were designated as training sets. Following that, the model was practiced using the training sets and validated using the testing set. Each experiment extracted the total accuracy and a confusion matrix for nine classes for validation. The 9-class confusion matrix is converted into another matrix, which includes true positive (TP), true negative (TN), false positive (FP), and false negative (FN), as illustrated in Table 1.

For each class, the estimated TP, TN, FP, and FN can be used to calculate accuracy (ACC), sensitivity (Se), and specificity (Sp). The model's accuracy relates to its ability to correctly detect instances. The proportion of "actual" positives that are correctly identified as positives is referred to as sensitivity, whereas the proportion of "genuine" negatives that are correctly identified as negatives is referred to as specificity. The accuracy, sensitivity, and specificity equations are represented in terms of TP, TN, FP, and FN as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Se = \frac{TP}{TP + FN} \tag{2}$$

$$Sp = \frac{TN}{TN + FP} \tag{3}$$

Table 1. Average ACC, Se and Sp for 9-class

No	Class	Accuracy (%)	Sensitivity (%)	Specificity (%)
1	Assalamu alaikum	100	100	100
2	Good morning	94	67	100
3	Hello	100	100	100
4	How are you	93	100	93
5	I	94	50	100
6	Study	100	100	100
7	Write	100	100	100
8	Television	100	100	100
9	Bath	100	100	100

The TP, TN, FP, and FN can also be used to calculate the Matthews correlation coefficient (MCC), Fowlkes-Mallows index (FM), and Bookmaker informedness (BM) to demonstrate the statistical relevance of each class. MCC is a metric for comparing observed and predictable binary categorization. FM is used to compare the similarity of observed and estimated binary classifications. BM is applied for approximating the probability of an informed decision.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

$$FM = \sqrt{\frac{TP}{TP + FP} \times \frac{TP}{TP + FN}} \tag{5}$$

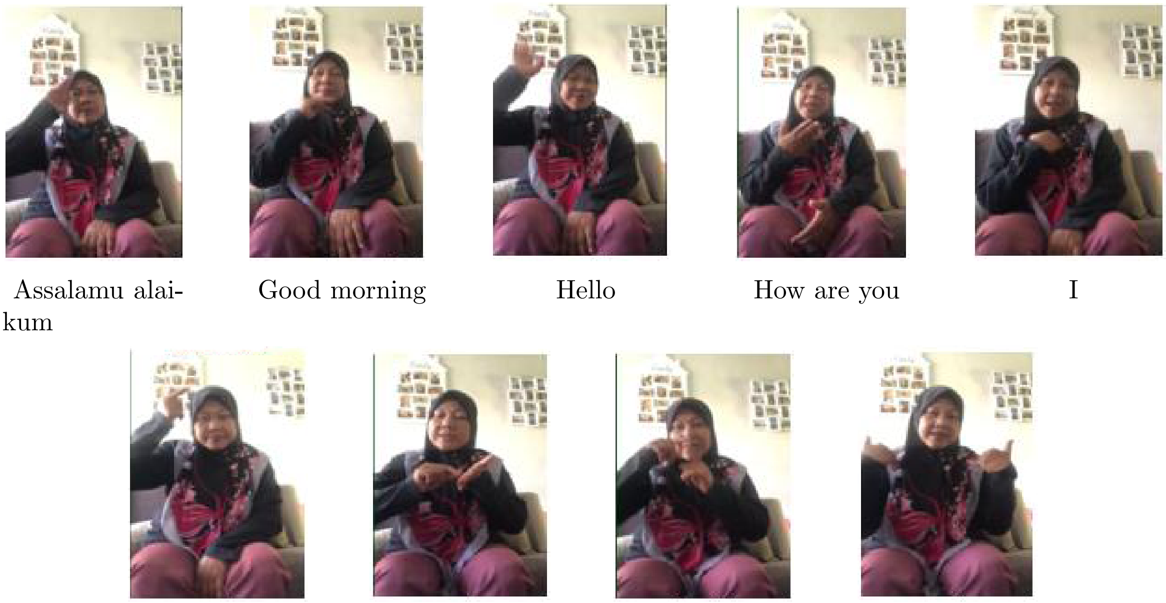
$$BM = Se + Sp - 1 \tag{6}$$

3. **Results and Discussion.**

To justify this model, nine examples of visualization of dynamic gesture are selected, as illustrated in Figure 3.

In this study, 9-class confusion matrices were created for each of the five trials and then translated into TP, TN, FP, and FN matrices. As a consequence, accuracy, sensitivity and specificity were computed. To appropriately analyze the results, an average of over five trials were collected for each letter for ACC, Se and Sp, as shown in Table 1. The model's per-class accuracy and sensitivity were calculated to be 100 per cent, with the exception of the expressions *how are you*, *good morning* and *I*, which amounted to just 50%, showing that the model has a high chance of correctly predicting and identifying a favorable result in each of the nine classes. As a result, the proportion of accurately identified instances would be higher. For the word signs for, *how are you*, sensitivity only reaches 100%.

In the numerical experiments, all of the algorithms acquire higher accuracy (100%) in the classes of *assalamu alaikum*, *hello*, *study* and *write*, *television* and *bath*. For the sign words for *good morning*, *how are you* and *I*, we only received 94, 93, and 94 percent, respectively. As a result, the proposed method's average accuracy for 9-classes is 98 per cent. Furthermore, the proposed technique predicted the 9-class categorization with a



Assalamu alaikum

Good morning

Hello

How are you

I

Study

Write

Television

Bath

Figure 3. The visualization of dynamic gesture

Table 2. Statistical significance for 9-class

No	Class	MCC (%)	FM (%)	BM (%)
1	Assalamu alaikum	100	100	100
2	Good morning	79	82	67
3	Hello	100	100	100
4	How are you	68	71	93
5	I	68	71	50
6	Study	100	100	100
7	Write	100	100	100
8	Television	100	100	100
9	Bath	100	100	100

statistically good accuracy, a higher level of similarity between observed and expected binary classifications, and a higher likelihood of calculating an informed decision. Table 2 introduces statistical significance of the proposed method. The word sign for *assalamu alaikum*, *hello*, *study*, *write*, *television* and *bath* achieved 100%; only *good morning*, *how are you* and *I* achieved less than 100%. Consequently, we can conclude that the proposed method has a better accuracy and good prediction in the numerical analysis.

In contrast, model accuracy was estimated to be much lower than expected after 100 epochs of training; hence, 80 sessions of training were chosen. A graph on model accuracy over epochs was generated to evaluate the adequacy of selection, as illustrated in Figure 4. As can be seen, the model's accuracy grows with the number of epochs, and the graph for the testing set gradually flattens between the 100 and 300 epochs. Model loss, on the other hand, fell dramatically in the first 100 epochs and then narrowed, but continued, as seen in Figure 5. The loss graph for the testing set eventually becomes flat soon before the 300th epoch.

The suggested work employs LSTM and a jump motion controller to recognize 9-word signs. The proposed approach appears to have a somewhat better overall performance since it is higher precise and quickly process. The higher ability of neural networks to handle big series datasets can certainly be attributed to the LSTM in particular.

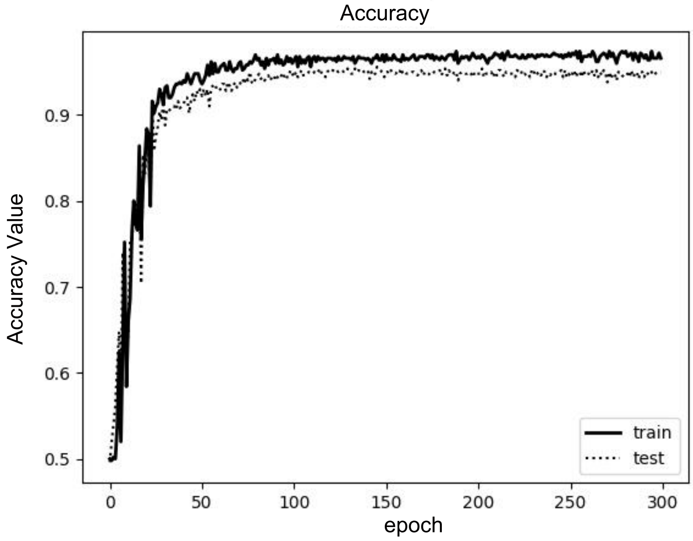


Figure 4. Model accuracy over epochs

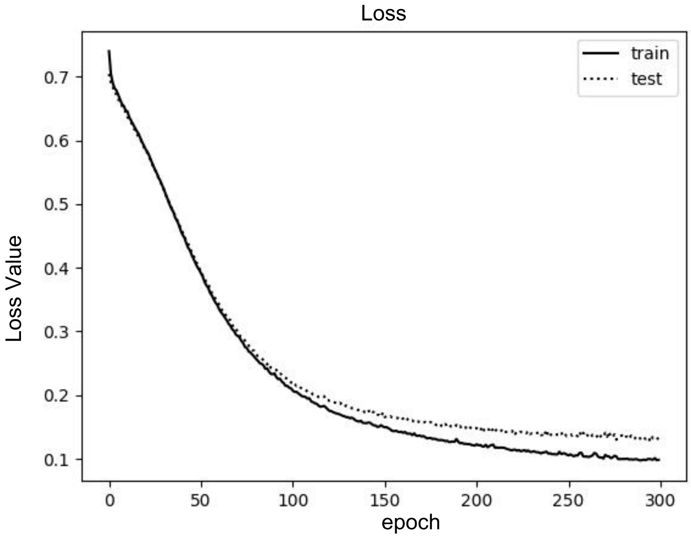


Figure 5. Model loss over epochs

4. **Conclusions.** The proposed research has developed an innovative model for detecting hand gestures in deaf sign language translation systems. The spatial-temporal properties of a gesture sequence may be extracted using a combination of LSTM and an algorithm sequence model, which was particularly useful for dynamic gesture recognition. In terms of employing it in a real-time application, integrating the LSTM and sequence classification model might reduce the model's gesture classification search to a smaller one, making the model's work easier and improving the accuracy result. Based on the result of research it indicated that the accuracy of the proposed method achieved 98 per cent for nine kinds of gesture testing. Consequently, the proposed method is appropriate to identifying dynamic hand gestures in deaf sign language translation systems. For the future research, it is recommended to develop a new strategy that can be used to eliminate motion transition from each gesture in order to improve the accuracy of dynamic sign detection from deaf sign language.

Acknowledgment. The authors would like to take this opportunity to thank the Ministry of Education, Culture Research and Technology of the Republic of Indonesia, and Universitas Hasanuddin for research facilities.

REFERENCES

[1] Y. Li, J. Huang, F. Tian, H. A. Wang and G. Z. Dai, Gesture interaction in virtual reality, *Virtual Reality and Intelligent Hardware*, vol.1, pp.84-112, 2019.

[2] Y. S. Tan, K. M. Lim and C. P. Lee, Hand gesture recognition via enhanced densely connected convolutional neural network, *Expert Systems with Applications*, vol.175, 114797, 2021.

[3] P. K. Athira, C. J. Sruthi and A. Lijiya, A signer independent sign language recognition with co-articulation elimination from live videos: An Indian scenario, *Journal of King Saud University – Computer and Information Sciences*, vol.34, no.3, pp.771-781, 2022.

[4] R. O. M. Mor, E. Obasi, J. Lu, N. Odeh, S. Kirker, M. M. Sweeney, S. G. Meadow and T. R. Makin, Talking with your (artificial) hands: Communicative hand gestures as an implicit measure of embodiment, *iScience*, vol.23, no.11, 101650, 2020.

[5] V. Adithya and R. Rajesh, Hand gestures for emergency situations: A video dataset based on words from Indian sign language, *Data in Brief*, vol.31, 106016, 2020.

[6] U. T. Salim and S. A. Dawwd, Systolic hand gesture recognition/detection system based on FPGA with multi-port BRAMs, *Alexandria Engineering Journal*, vol.58, pp.841-848, 2019.

[7] Y. Li, D. Ma, Y. Yu, G. Wei and Y. Zhou, Compact joints encoding for skeleton-based dynamic hand gesture recognition, *Computers and Graphics*, vol.97, pp.191-199, 2021.

[8] K. Lai and S. N. Yanushkevich, CNN+RNN depth and skeleton based dynamic hand gesture recognition, *The 24th International Conference on Pattern Recognition (ICPR)*, Beijing, China, pp.1-4, 2018.

[9] S. Lata and O. Surinta, An end-to-end Thai fingerspelling recognition framework with deep convolutional neural networks, *ICIC Express Letters*, vol.16, no.5, pp.529-536, 2022.

[10] 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.

[11] J. C. Núñez, R. Cabido, J. J. Pantrigo, A. S. Montemayor and J. F. Vélez, Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition, *Pattern Recognition*, vol.76, pp.80-94, 2018.

[12] D. K. Singh, 3D-CNN based dynamic gesture recognition for Indian sign language modeling, *Procedia Computer Science*, vol.189, pp.76-83, 2021.

[13] M. Maghoumi and J. J. LaViola, Deep GRU: Deep gesture recognition utility, *Advances in Visual Computing*, pp.16-31, 2019.

[14] E. T. P. Barros, C. Weber and S. Wermter, An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition, *Neurocomputing*, vol.268, pp.76-86, 2017.

[15] T. M. Tai, Y. J. Jhang, Z. W. Liao, K. C. Teng and W. J. Hwang, Sensor-based continuous hand gesture recognition by long short-term memory, *IEEE Sensors Letters*, vol.2, pp.1-10, 2018.

[16] M. Jabbari, R. N. Khushaba and K. Nazarpour, EMG-based hand gesture classification with long short-term memory deep recurrent neural networks, *The 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Canada, pp.1-5, 2020.

[17] K. Anand, S. Urolagin and R. K. Mishra, How does hand gestures in videos impact social media engagement – Insights based on deep learning, *International Journal of Information Management Data Insights*, vol.1, no.2, 100036, 2021.

[18] Q. Xiao, X. Chang, X. Zhang and X. Liu, Multi-information spatial-temporal LSTM fusion continuous sign language neural machine translation, *IEEE Access*, vol.8, pp.216718-216728, 2020.

[19] Q. Zhang, J. Zhang, J. Zou and S. Fan, A novel fault diagnosis method based on stacked LSTM, *IFAC-PapersOnLine*, vol.53, pp.790-795, 2020.

[20] M. Z. Islam, M. M. Islam and A. Asraf, A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images, *Informatics in Medicine Unlocked*, vol.20, 100412, 2020.

[21] A. Sherstinsky, Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network, *Physica D: Nonlinear Phenomena*, vol.404, 132306, 2020.

[22] A. H. Manurung, W. Budiharto and H. Prabowo, Algorithm and modeling of stock prices forecasting based on long short-term memory (LSTM), *ICIC Express Letters*, vol.12, no.12, pp.1277-1283, 2018.

[23] T. Wahyono, Y. Heryadi, H. Soeparno and B. S. Abbas, Enhanced LSTM multivariate time series forecasting for crop pest attack prediction, *ICIC Express Letters*, vol.14, no.10, pp.943-949, 2020.