JOIV : Int. J. Inform. Visualization, 7(2) - June 2023 494-501

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Implementation of Convolutional Neural Network and Long Short-Term Memory Algorithms in Human Activity Recognition Based on Visual Processing Video

Andi Nur Rachman [a], Husni Mubarok [b], Euis Nur Fitriani Dewi [b,*], Rama Edwinda Putra [b]

[a] *Department of Information System, Universitas Siliwangi, Tasikmalaya, Indonesia*
[b] *Department of Informatics, Universitas Siliwangi, Tasikmalaya, Indonesia*
Corresponding author: [*]euis.nurfitriani@unsil.ac.id

***Abstract*—Human Activity Recognition (HAR) is an interesting research topic, especially in identifying human movement actions focusing on video-based security surveillance. Symptom of an illness from a movement. The use of HAR in this research is the key to better understanding the various semantics contained in the video to find out the pattern of a human movement, especially in sports movements. In this study, a combination of the CNN and LSTM method algorithms was applied by using several variations of the model parameter values on the dropout layer and batch size to convert the pattern in the video into image form to produce a HAR model. Data processing at the convolution layer is used to extract spatial features in the frame. The extraction results are fed to the LSTM layer on each network for modeling the temporal sequence of human movement. In this way, the network on the model will learn spatiotemporal features directly in end-to-end data training tests to produce a robust model. The test data used are 10 sports activities obtained from related research from the University of Central Florida (UCF). The results showed that the performance was quite good, although there were still errors in the classification of sports activities because they had similarities in the movements of the activities carried out. The classification results show a loss value of 0.4 and an accuracy of 0.94. In further research, what needs to be corrected is the loss value which is still high so that several times the test results show an error in the classification of sports activities that have similarities in the movements of the activities.**

***Keywords*— Human activity recognition (HAR); classification, convolutional neural network; long short-term memory.**

Manuscript received 24 Dec. 2022; revised 18 Mar. 2023; accepted 1 Apr. 2023. Date of publication 30 Jun. 2023.

I. INTRODUCTION

The analysis process of human activities and event detection in the video is conducted manually by repeatedly observing the motion of human action from CCTV, which is very time-consuming and energy. HAR video-based is a challenging research field today, specifically observing sports activities [1], [2], [3]. HAR is a key to understanding various semantics of video content [4]. The recognition of human activities on a video and others else, such as taking video for the content, system of area surveillance for office security, private services, and human-computer interaction, has gained significant attention in computer vision today [5], [6].

Deep learning is the most used technique in machine learning, with many high-end features to identify actions and human behavior based on the video. The Convolutional Neural Network (CNN) is a deep learning architecture used in HAR of convolutional operation to learn video frames in every training process [7].

The activities recognition using pre-trained weight of machine learning architecture to represent video frame visually in the training stage which affect difference feature determination, such as the difference of visual and temporal. The deep learning in HAR using Bi-Directional Long Short-Term Memory (BiLSTM) and Dilated Convolutional Neural Network (DCNN) architectures, which were focused on input feature frame that effective for recognizing several human actions in a video [8]. Recurrent convolutional architecture can be end-to-end trained for a vast scale of visual understanding in recognizing activities, text, images, and video descriptions based on camera on HAR, like in health services and social caring. The RNN gained 99.55% in average accuracy and recognized twelve human activities [9].

HAR uses the joints skeleton technique that focuses on noise that causes joint points to be irrelevant and unsteady,

becoming the main performance deterioration problem. Research that used the transfer learning method of VGG-16 to obtain image features and classification processes by CNN in human activities recognition. The accuracy proposed by that research reached 96.95% [10].

The subsequent study is about a two-stream structure using LSTM for the spatial data stream when extracting the motion in the video by utilizing spatial and temporal features on an RGB frame. Furthermore, there was a study of HAR that focused on multi-class classifications to increase accuracy with low computation and reduced model complexity by eliminating processes that are needed for feature advanced techniques [11], [12].

On the other hand, Liu et al. [13] studied the architecture model by combining 3DCNN and LSTM methods. The proposed model can stack video frames, extract time and spatial features, and also perform data training for action video to gain performance of reliable recognition. The LSTM model becomes a bridge between frames at different times to obtain better information about data from the previous frame.

This study was designed based on previous research with implementation and combining CNN and LSTM methods. Then, in its processes, convolutional layers used to extract the action of spatial features from video frames became input variables for LSTM. After that process, the network architecture of the temporal HAR model will be created. In this way is believed that the network architecture of machine learning classifications will learn spatiotemporal features directly at training end-to-end for gaining a reliable HAR model to identify human action.

II. Materials and Method

Several studies regarding HAR on video have been carried out in previous studies using various methods, such as the research entitled Human action recognition using attention-based LSTM network with dilated CNN features [4]. This research discusses video focus on motion action recognition techniques by using pre-training weighting identification methods, then from the results of machine learning architectural designs in the visual representation of video frames processed at the training stage can affect the differences in features, such as visual difference with temporal video. This research proposes a machine learning architecture for HAR using Bi-Directional Long Short-Term Memory (BiLSTM) algorithms and Dilated Convolutional Neural Network (DCNN) in a reduced video detection manner focusing on managing input frame identification features that can effectively recognize various human actions in videos.

Related research discusses Human action recognition using LSTM and fully connected LSTM by providing different data input indicators. This study aims to discover the features of motion video action recognition with the STDAN network architecture, combining Convolutional LSTM and Fully Connected LSTM. Another study was entitled Long-term Recurrent Convolutional Networks for Visual Recognition and Description [5]. The research focuses on a class of iterative convolution architectures in images that can be trained end-to-end and are suitable for big data-scale visual identification for motion recognition activities, text, images, and video descriptions.

Park et al. [6] used a depth camera based RNN method on HAR for health and social care services. The proposed method achieves an average recognition accuracy of 99.55% and can reliably recognize twelve human activities. Arif et al. [7] proposed combining the 3D-CNN method with LSTM. In the process, the 3-dimensional convolution network will combine the raw information from the video into motion identification, called a motion map. Combining motion maps and video frames can increase the length of training videos iteratively. Applying a linear weighted fusion scheme combines the identification of motion in network features into spatiotemporal features and applying an encoder-decoder in LSTM to carry out the final prediction process.

Another study on HAR used the skeleton joints technique, which was carried out with the title Human Activity Recognition Based On Optimal Skeleton Joints Using Convolutional Neural Networks [8]. This study focuses on noise that causes irrelevant and immovable joint points, which is the main cause of decreased HAR performance [8]. Zheng et al. [9] conducted a study using the VGG-16 transfer learning method to get deep image features and a machine learning classification process trained on CNN in human activity recognition. The accuracy of the method proposed in this study reached 96.95%. Research conducted by Zhao et al. [10] improved a two-stream model for human action recognition and examines a two-stream structure by using LSTM for spatial data streams in extracting video motion by utilizing spatial and temporal features in RGB frames [28], [29].

There is research on HAR that focuses on multi-class classification processes in increasing accuracy with low computational costs [11], [12], [30], [31], [32] and reducing model complexity by process of elimination required for advanced features techniques. Some previous studies designed the architectural model using a combination of the 3DCNN and LSTM methods [13], [26], [27]. In this study, the proposed model is capable of stacking video frames, extracting time and spatial features [14], [15], [16]. As well as carrying out the video movement data training process to achieve good recognition performance [17], [18]. The LSTM model is a link between frames at different times to get better data information on previous frames [19], [20], [21].

Based on related research, in each stage of his research, using the HAR method has advantages and disadvantages for designing machine learning model architectures in identifying human movements. This study was designed based on the results of previous research studies using a combination of CNN and LSTM methods. In the HAR stage process, where the convolution layer is used to extract the movement of spatial features from the frame, it will be fed to the LSTM layer [22], [23], [24]. After that, the network architecture of the temporal HAR model was made. In this way, the network architecture of the model will classify machine learning into studying spatiotemporal features directly in end-to-end training to produce a good HAR model in identifying human movement [25].

III. Results and Discussion

The sample or research data used in this study is the UCF50 dataset. The research carried out is to build a machine learning model using a combination of CNN and LSTM algorithms on

HAR and test quality parameters using the confusion matrix method, which shows the success of implementing algorithms, hyperparameters, and architecture models used. The research stages are presented as a whole, as in Figure 1.

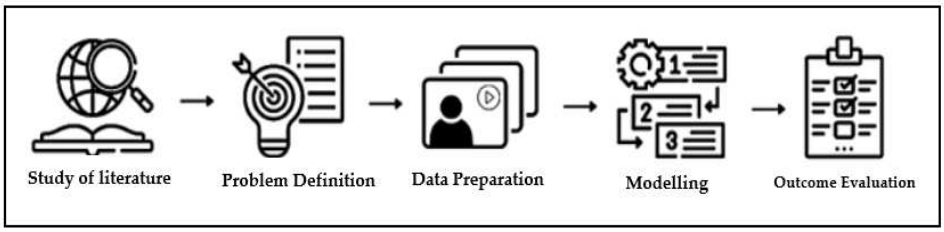

Fig. 1 Research Stages

The modeling process begins with designing the machine learning model architecture. Figure 2 is the architecture of the machine learning model in this study. Four convolutional layers use ReLU activation, which makes the limiting value at zero to determine whether or not the neurons are active in the neural network, so only neurons related to objects are selected and followed by Maxpooling2D to reduce the number of input parameters spatially and layers. Dropout to reduce overfitting problems. The convolution and flattening layers are wrapped in a Time Distributed layer which is used to process sequence or time-series data, and it is possible to apply the layer to each temporal slice of input in parallel to the training process. The extracted features in the Conv2D layer will be converted using the Flatten layer and will be fed to the LSTM layer. The activation function used in the Dense layer or fully connected is SoftMax which will use the output from the LSTM layer to predict the action to be taken, presented as a whole as in Figure 2.

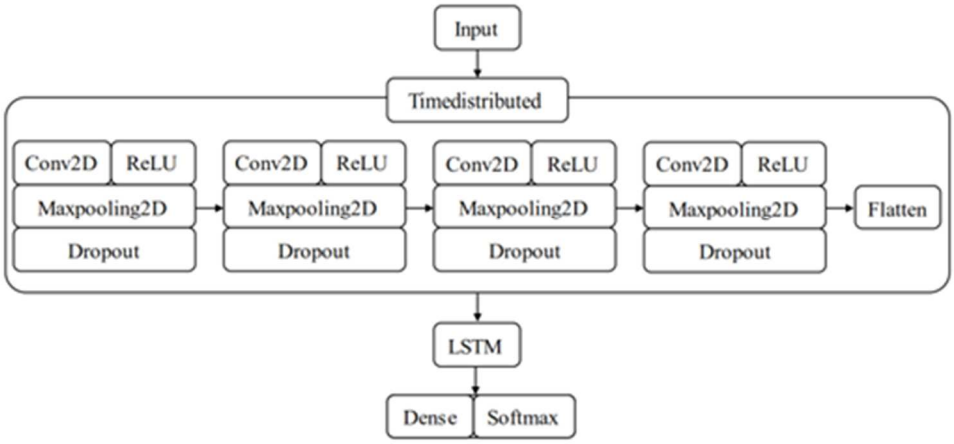

Fig. 2 Machine Learning Model Design

The system is tested with some testing data to determine the performance of the model that has been built using the confusion matrix method. The confusion matrix is used to determine the value of accuracy, precision, recall, f1-score, and support. The following is the equation used to determine the value of accuracy, precision, recall, and f1-score using the confusion matrix. The confusion matrix is presented as a whole, as in Table 1.

TABLE I
CONFUSION MATRIX

Confusion Matrix		Predictions	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Accuracy is a performance value in the model based on the degree of closeness between the predicted value and the actual value. Accuracy can be interpreted as an illustration of how precise the model is in carrying out the classification process correctly. Determine the accuracy value can be done with the following formula 1:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

Precision is the ratio of the amount of relevant information selected by the system to the total amount of information selected. Precision can be interpreted as a match between the information requested by the system and the predicted results provided by the model. Determine the precision value can be done with the following formula 2:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall is the ratio of the amount of relevant information the system selects to the total amount of relevant information available. Recall can be described as the success of the model in finding relevant information. Recall can be calculated using the following formula 3:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1-Score is a metric used to measure model performance by comparing the average precision and recall values. F1-Score can be calculated using the following formula 4:

$$F1 - Score = 2 \times \frac{(Recall \times Precision)}{(Recall + Precision)} \quad (4)$$

The model's design consists of several processes for classifying the HAR, starting with the input data or dataset from the extraction process on the video that is used as a learning resource. The input data that is loaded is the result of dividing the dataset in the form of training, validation, and test datasets. The model design to be built uses a combination of CNN and LSTM algorithms. After the design stage of the machine learning model is complete, the process will be carried out training to produce machine learning models that can perform the classification process on the HAR. The Machine Learning Models are presented as a whole, as in Figure 3.

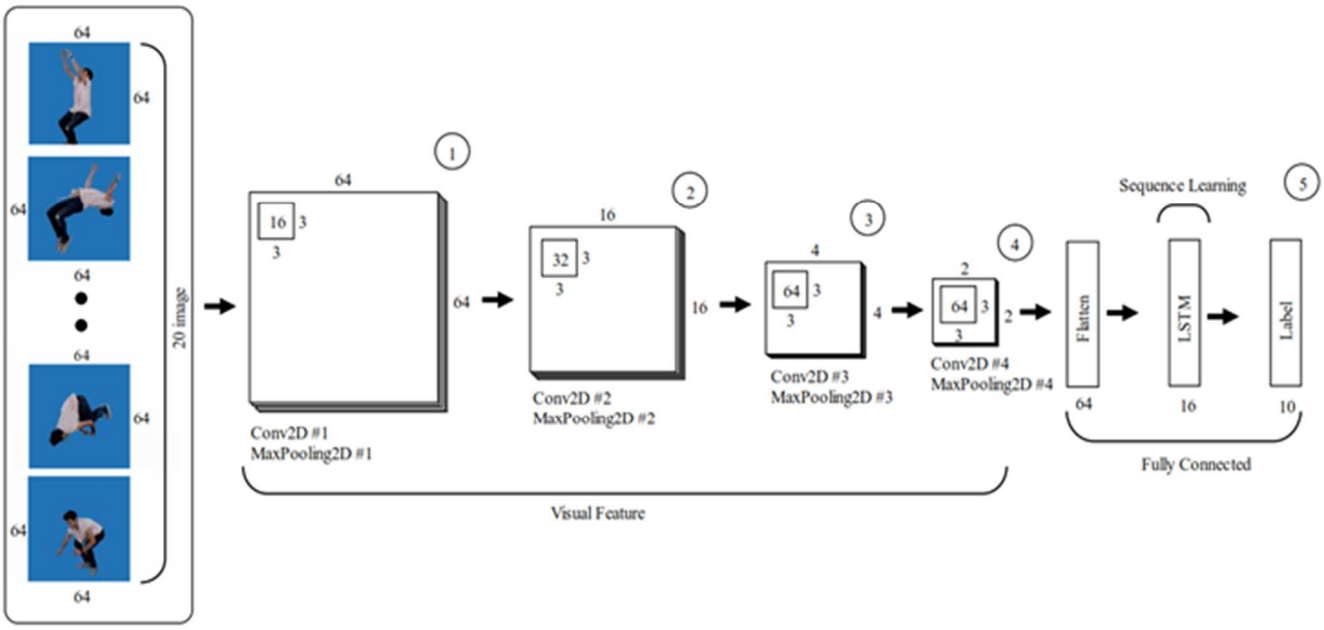

Fig. 3 Machine Learning Models

The machine learning model in this study uses four layers convolution, maxpooling2D layer, dropout layer, flatten layer, LSTM layer, and dense layer, as seen in Figure 4.6. The convolutional and flattened layers were wrapped by Time Distributed, which is used for managing data sequences or time series and can implement a layer to every temporal slice from input in parallel with the training process.

Data input is an image with a length of 64 pixels, a width of 64 pixels, and three channels of RGB. The data were processed by the first layer of the convolutional (see Figure 3 point 1) that extracted features with 16 filters, kernel 3x3, padding with 'same' parameter and used ReLU for activation function.

The second convolutional layer (See Figure 3 point 2) extracted the feature of the image with 32 filters, kernel 3x3, padding with the 'same' parameter, and still used ReLU as an activation function. As previously processed in this layer, Maxpooling2D is used with 4x4 of size and dropout layer as additional.

The third convolutional layer acquired data input as 4x4 pixels of the image. This layer (Figure 3 point 3) extracted image features with 64 filters, kernel 3x3, padding with the 'same' parameter, and ReLU as an activation function. The Maxpooling2D in this layer with 2x2 in size and dropout layer is added in this layer. The fourth convolutional layer is identical to the previous one, except this layer processes image 2x2 pixels as data input.

The feature extracted on the Conv2D layer is converted to a vector using Flatten layer, whose result becomes data input for the LSTM layer. In this layer (See Figure 3 point 5), the result is reduced become 16 outputs which are processed at the Dense layer suitable with the number of classes in the category or dataset label. The activation function used in the Dense layer or fully connected layer was SoftMax, which was to calculate the probability of all labels obtained from the LSTM output to forecast action taken by humans.

The model processes data input as categorical data, thereby using categorical cross-entropy loss. This model's optimizer uses ADAM because it is generally better than other optimizer algorithms in processing more data and has good efficiency in computation time and memory usage.

A. Training Process

The parameter configurations are needed in the training process to gain the optimum machine learning model. The experiment was conducted with various parameter values at the dropout layer and batch size. The six models' variations are shown in Table 1, and the visualization can be seen in Figure 4.

TABLE II
VARIATION PERFORMANCE RESULTS

Model type	Dropout layer	Batch sized value	Number of epochs	Accuracy Value results	Loss	Acc Val	Loss Val	Exec time
A	-		50	1.000	0.003	0.909	0.500	94.14s
B	0.2	4	50	0.963	0.135	0.781	0.973	85.78s
C	0.4		50	0.972	0.076	0.824	0.675	86.31s
D	-		46	1.000	0.012	0.856	0.674	74.64s
E	0.2	8	46	0.995	0.046	0.882	0.509	58.36s
F	0.4		50	0.914	0.306	0.754	0.806	63.27s

According to Table 1, it clearly can be seen that both models A and D, which are not applying a dropout layer, have a higher accuracy value and lower loss value than the opposite side but need more time for training. Also, the value of low batch size needs a longer execution time.

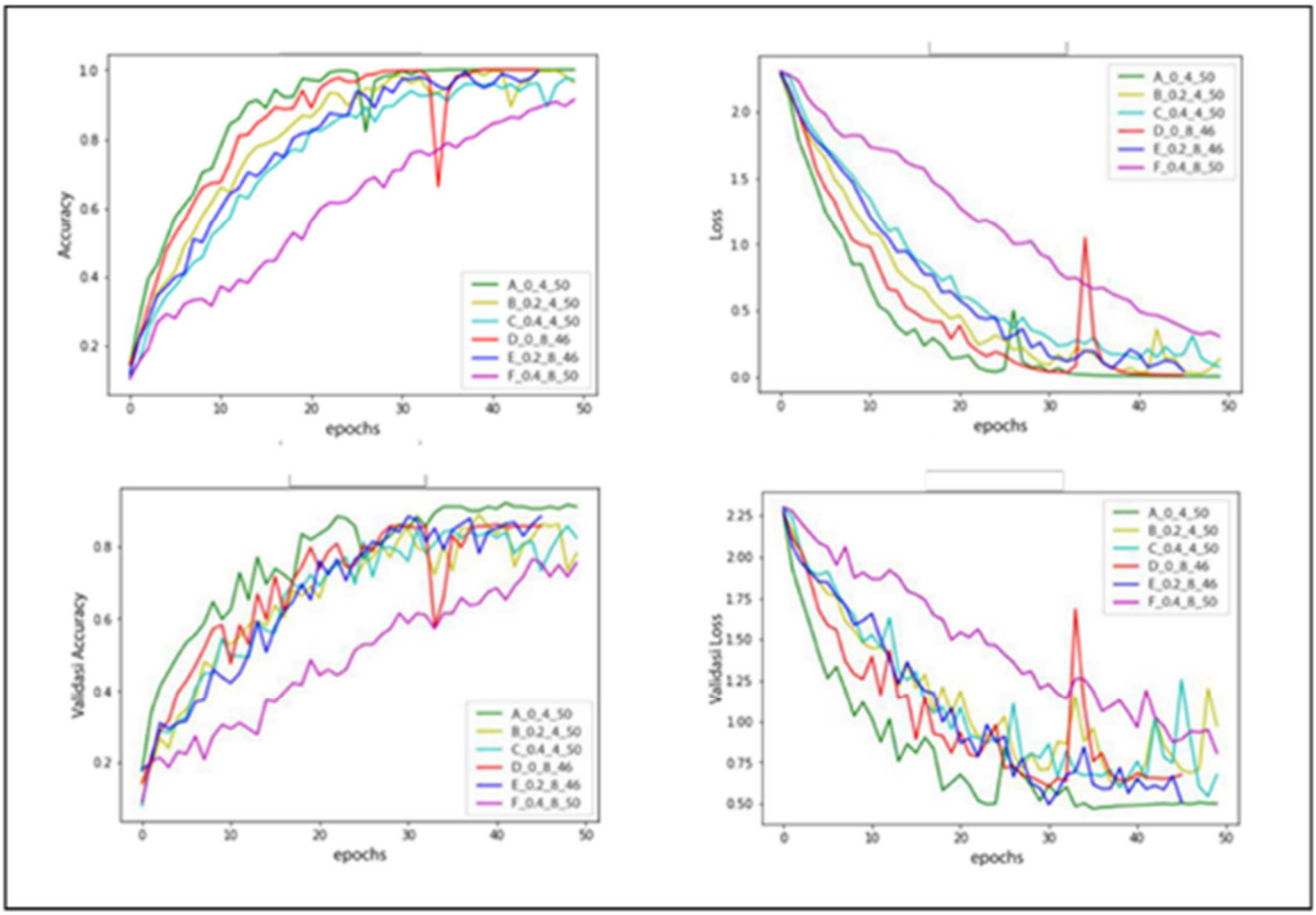

Fig. 4 Variation Performance Graph

Figure 4 describes the increasing trend for accuracy and validation values, and in contrast with those, for loss graphic and validation loss have decreasing trend. The weight trend obtained from data validation was not more stable than the training data because the model had never learned the data used before. If the weight trend deviates consistently, the training process will be stopped early, as in models D and E, to avoid overfitting.

B. Results Evaluation

A confusion matrix was used for analyzing the performance of the machine learning model. It was used to know the quality parameters of the classification model by counting the number of true and false predictions for all classes.

TABLE III
CLASSIFICATION REPORT

Variation	Loss	Accuracy	Precision	Recall	f1-score
A	0.4087	0.9375	0.94	0.94	0.94
B	0.5492	0.8600	0.86	0.86	0.86
C	0.4442	0.9050	0.90	0.91	0.90
D	0.3151	0.9275	0.93	0.93	0.93
E	0.3750	0.9050	0.91	0.91	0.90
F	0.5711	0.8450	0.84	0.84	0.84

Table 2 shows the report of classification using the testing dataset. Higher accuracy was obtained from the variation A dan D, which were without the applied dropout layer. On the variation that using a bigger batch size yields lower loss value and vice versa because the bigger batch size will slowly give the training process more convergent with accuracy on predicting.

The confusion matrix result is the actual data and data prediction using the testing dataset of the machine learning model that was built. It can be seen in the confusion matrix graphic variations A and D, which are without applied dropout layers, showed significant performance with the greatest number of classes of true positive, even though there was still misclassification, such as in jumping-jack activity becoming basketball. This matter happened because of close similarities in the motion of both sides.

The following confusion matrix evaluation describes how the performance measure was calculated, and the detailed description is illustrated using one class output, namely class 0. If we compare the accuracy between all variations in the experiment, then variations A and D have had higher accuracy than the others. Likewise, if the whole class is calculated on average, then Variations A and D consistently still have the highest average accuracy for the entire experiment, as previously seen in Table III.

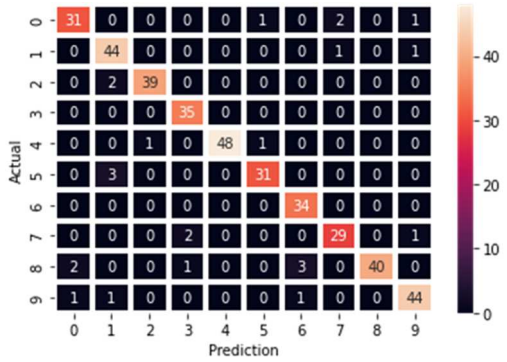

Fig. 5 Confusion Matrix Variation (A)

Figure 5 describes the evaluation result from the multi-class confusion matrix. The columns represent actual classes, while the rows are for prediction classes. To obtain accuracy, precision, recall, and F1-score value, the interpretation of TP, FP, TN, and FN must be performed first. As an illustration, we calculate that value for class 0. The TP value is picked from $cell_{1,1}$ as 31 data, meaning the data were correctly classified as expected. Figure 5 shows the TP value for all classes located at diagonal cells. The TN value of class 0 can be calculated by adding all matters of the cell except for cells in column 1 and row 1, so it accounts for 362 data. The FP of the class 0 obtained by adding value from $cell_{1,2}$ to $cell_{1,10}$ accounted for 4 data. The last, for the FN, can be calculated from the value of $cell_{2,1}$ through $cell_{2,10}$, which is 3 data. Based on formula numbers 1,2,3, and 4, the value of accuracy, precision, recall, and F1-score of class 0, respectively, are obtained as 98.25%, 88.57%, 91.18%, and 89.86%.

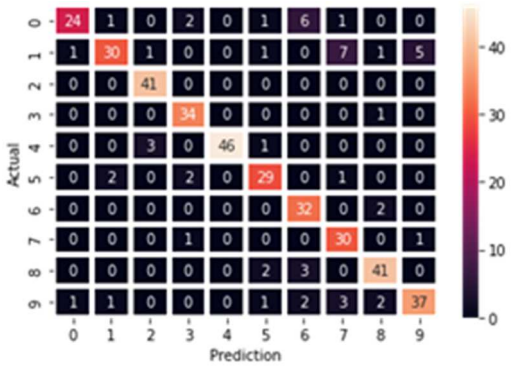

Fig. 6 Confusion Matrix Variation (B)

The confusion matrix evaluation for Variation B can be interpreted from Figure 4 using the same calculation method. The TP, FP, TN, and FN of class 0 are 24, 11, 363, and 2, respectively. So, we obtained the accuracy, precision, recall, and F1-score of the class 0 in Variation B consecutively are 96.75%, 68.57%, 92.31%, and 78.68%.

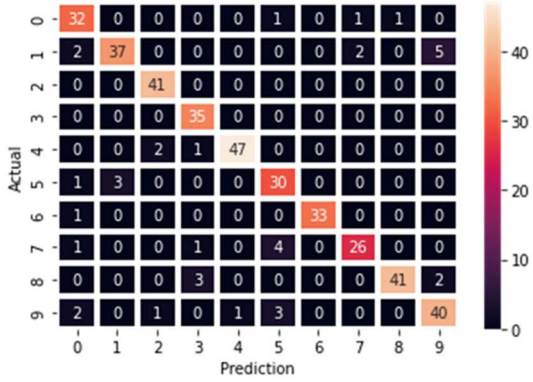

Fig. 7 Confusion Matrix Variation (C)

The evaluation for Variation C is depicted in Figure 7, where the values of TP, FP, TN, and FN of the class 0 consecutively is 32, 3, 358, and 7. Using the same formula as previously described, the accuracy was 97.50%, the precision was 91.43%, the recall was 82.05%, and finally, the F1-score was 86.49%.

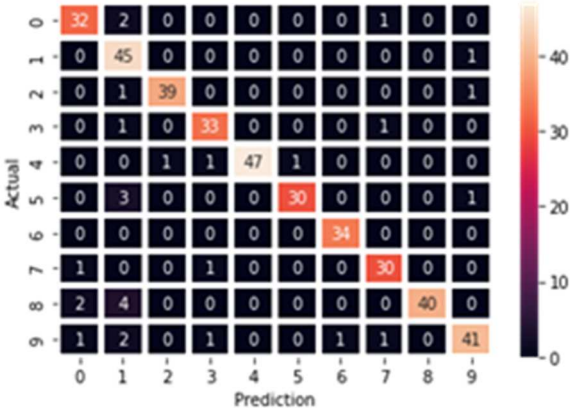

Fig. 8 Confusion Matrix Variation (D)

The fourth confusion matrix is the evaluation of Variation D, as can be seen in Figure 8, where the value of TP, FP, TN, and FN consecutively is 32, 3, 361, and 4 with 98.25%, 91.43%, 88.89%, and 90.14 % is value for accuracy, precision, recall and F1-score for class 0 in Variation D, respectively.

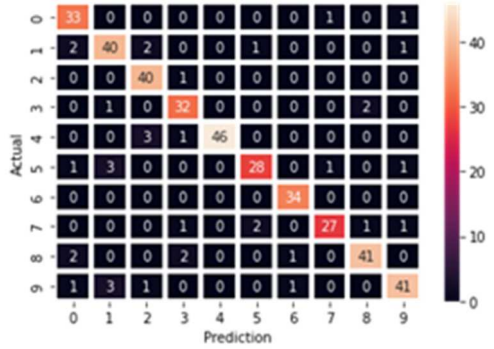

Fig. 9 Confusion Matrix Variation (E)

The subsequent evaluation is for Variation E, which obtained an accuracy of 98.00%, a precision of 94.29%, a recall of 84.62%, and an F1-score of 89.19% of class 0. Those performance calculations were gained from the value of TP 33, FP 2, TN 359, and FN 6.

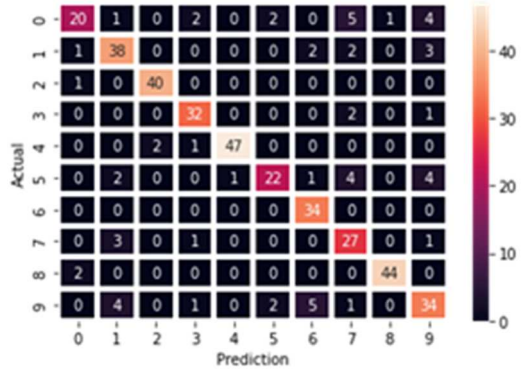

Fig. 10 Confusion Matrix Variation (F)

The last evaluation is for Variation F, where the confusion matrix resulted in an accuracy of 95.25%, a precision of 57.14%, a recall of 83.33%, and an F1-score of 67.80% of class 0. Those results are based on TP, FP, TN, and FN values, respectively, as 20, 15, 361, and 4. Based on the experiment result, the model performance in training and testing was affected by dropout layer implementations and batch size with an appropriate measure. The main function of the dropout layer is to prevent overfitting. However, the bigger dropout parameter size will impact the model's inability to fit at training time because of reduced model capability correctly. Furthermore, removing neurons in the hidden layer and visible layer in the network affects the bad result of the model.

C. Threat of Validity

The experiment model was only performed using the UCF50 dataset with ten types of sports activities. The used dataset is a video with three channels of color, 64x64 pixels, and a number of video frames processed by the model are 20 sequences. Lacking a number of the dataset in this experiment impacted the model learning capabilities toward dataset training. This matter is a reason for the experiment's variation not being optimum. The experiment used a Google Collaboratory environment with a hardware-sharing scheme to affect the experiment's performance. For the following research, using a dedicated machine with suitable specifications to gain maximum performance is highly recommended.

IV. Conclusion

The study resulted in six models with several variations in the value of parameters of the model, especially on dropout layer and batch size. According to the experiment result, the highest accuracy was obtained from the variation that did not implement a dropout layer with batch size four accounting for 0.94 and loss value 0.4. Whereas the lowest accuracy was obtained from the variation that implemented dropout layer as 0.4 with batch size accounting for 8, the accuracy and the loss are 0.84 and 0.57, respectively.

Following the experiment result, the accuracy trends and its validation are increasing, while the loss and validation loss is decreasing. This matter showed that the model has good performance. Both variations in training or testing processes that were not implemented dropout layer obtained high accuracy and low loss value but needed more execution time for the training process. On the contrary, the model that implemented the dropout layer behaves otherwise.

Because the value loss is still high and the occurrence of misclassification in activities with a similar motion, the future study must focus on parameter and hyperparameter tunings with a sufficient dataset. The transfer learning method also must be considered, such as using pre-trained architecture like VGGNet, ResNet, and DenseNet to gain the optimum result.

References

[1] Żelawski, Marcin and Hachaj, Tomasz. "The application of topological data analysis to human motion recognition" International Journal Technical Transactions, vol.118, no.1, 2021, pp.-. doi: 10.37705/TechTrans/e2021011.

[2] Z. Zhang, Z. Lv, C. Gan, and Q. Zhu, "Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions," International Journal Neurocomputing, vol.410, pp.304–316, 2020, doi: 10.1016/j.neucom.2020.06.032.

[3] Mokari, M., Mohammadzade, H., & Ghojogh, B. (2020). Recognizing involuntary actions from 3D skeleton data using body states. International Journal *Scientia Iranica*, *27*(3), 1424-1436. doi: 10.24200/sci.2018.20446.

[4] K. Muhammad et al., "Human action recognition using attention based LSTM network with dilated CNN features," International Journal Future General Computing System, vol. 125, pp. 820–830, 2021, doi: 10.1016/j.future.2021.06.045.

[5] J. Donahue et al., "Long-term Recurrent Convolutional Networks for Visual Recognition and Description," pp. 1–14, 2016.

[6] S. U. Park, J. H. Park, M. A. Al-Masni, M. A.Al-Antari, M. Z. Uddin, and T. S. Kim, "A Depth Camera-based Human Activity Recognition via Deep Learning Recurrent Neural Network for Health and Social Care Services," Procedia Computing Science., vol. 100, pp. 78–84, 2016, doi: 10.1016/j.procs.2016.09.126.

[7] S. Arif, J. Wang, T. Ul Hassan, and Z. Fei, "3D-CNN-based fused feature maps with LSTM applied to action recognition," Journal Future Internet, vol.11, no. 2, 2019, doi: 10.3390/fi11020042.

[8] N. Surayahani, M. Norzali, and M. Razali, "Human Activity Recognition Based on Convolutional Neural Network," Journal International Science Technology., vol. 2018-Augus, pp. 48–57, 2018, doi: 10.1109/ICPR.2018.8545435.

[9] S. Deep and X. Zheng, "Leveraging CNN and Transfer Learning for Vision-based Human Activity Recognition," 2019 29th International

Telecommunication Networks Application Conference ITNAC 2019, pp.35–38, 2019, doi: 10.1109/ITNAC46935.2019.9078016.

[10] Y. Zhao, K. L. Man, J. Smith, K. Siddique, and S. U. Guan, "Improved two-stream model for human action recognition," Eurasip Journal Image Video Process, vol. 2020, no. 1, 2020, doi: 10.1186/s13640-020-00501-x.

[11] W. Xu, Y. Pang, Y. Yang, and Y. Liu, "Human Activity Recognition Based On Convolutional Neural Network," in 2018 24th International Conference on Pattern Recognition (ICPR), Aug. 2018, vol. 11742 LNAI, pp. 165–170, doi: 10.1109/ICPR.2018.8545435.

[12] R. Mutegeki and D. S. Han, "A CNN-LSTM Approach to Human Activity Recognition," 2020 International Conference Artificial Intelligent Information Communication. ICAIIC 2020, pp. 362–366, 2020, doi: 10.1109/ICAIIC48513.2020.9065078.

[13] Y.-C. Liu, J.-J.Ding, Y.-J. Chang, C.-Y. Wang, and J.-C. Wang, "Action recognition using three dimension convolution and long short term memory," in 2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), Jun. 2017, pp. 83–84, doi: 10.1109/ICCE- China.2017.7991006.

[14] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions", *Journal of Big Data* (Vol. 8, Issue 1). Springer International Publishing, 2021, doi: 10.1186/s40537-021-00444-8.

[15] Batta, M., "Machine Learning Algorithms - A Review", *International Journal of Science and Research (IJ*, *9*(1), 381-undefined, 2020, doi: 10.21275/ART20203995.

[16] Caron, M., Bojanowski, P., Joulin, A., & Douze, M., "*Deep Clustering for Unsupervised Learning of Visual Features*", 2019, https://arxiv.org/abs/1807.05520.

[17] Nima, R., & Shila, F., "Crack classification in rotor-bearing system by means of wavelet transform and deep learning methods: an experimental investigation", *Journal of Mechanical Engineering, Automation and Control Systems*, *1*(2), 102–113, 2020 doi: 10.21595/jmeacs.2020.21799.

[18] Rebala, G., A, R., & S, C., "Machine Learning Definition and Basics", *Springer, Cham*, 2019, doi: 10.1007/978-3-030-15729-6_1.

[19] Wildan, M., Aldi, P., & Aditsania, A., "Analisis dan Implementasi Long Short Term Memory Neural Network untuk Prediksi Harga Bitcoin", *E-Proceeding of Engineering*, *5*(2), 3548–3555, 2018, https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/6739.

[20] Reddy, K. K., & Shah, M., 'Recognizing 50 human action categories of web videos", *Journal Machine Vision and Applications*, *24*(5), 971–981, 2013, doi: 10.1007/s00138-012-0450-4

[21] Ghosh, A., Sufian, A., Sultana, F., Chakrabarti, A., & De, D. "Fundamental concepts of convolutional neural network", *Journal Intelligent Systems Reference Library* (Vol. 172, Issue January), 2019, doi: 10.1007/978-3-030-32644-9_36.

[22] François-lavet, V., Henderson, P., Islam, R., Bellemare, M. G., François-lavet, V., Pineau, J., & Bellemare, M. G. "An Introduction to Deep Reinforcement Learning", *Foundations and Trends in Machine Learning*, *II*(3–4), 1–140, 2018, doi: 10.1561/2200000071.

[23] Firmansyah, R., "Implementasi Deep Learning Menggunakan Convolutional Neural Network Untuk Klasifikasi Bunga", *Fakultas Sains Dan Teknologi UIN Syarif Hidayatullah Jakarta*, 2020, https://repository.uinjkt.ac.id/dspace/handle/123456789/55347.

[24] Apaydin, H., Feizi, H., Sattari, M. T., & Colak, M. S., "Comparative Analysis of Recurrent Neural Network", *Water (Switzerland)*, *12*, 1–18, 2020, https://www.mdpi.com/2073-4441/12/5/1500.

[25] Hochreiter, S., & Schmidhuber, J., "Long Short-Term Memory". *Journal Neural Computation*, *9*(8), 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[26] Bhaskar, D., Manhart, A, Milzman, J, Nardini, J. T, Storey, K. M., Topaz, C. M., & Ziegelmeier, L. (2019). Analyzing collective motion with machine learning and topology. Chaos: An Interdisciplinary Journal of Nonlinear Science, 29(12), 123–125.

[27] Ko, J. H., Han, D. W., & Newell, K. M. (2018). Skill level changes the coordination and variability of standing posture and movement in a pistol-aiming task. Journal of Sports Sciences, 36(7), 809–816.

[28] Alwin Poulose, Jung Hwan Kim, Dong Seog Han, "HIT HAR: Human Image Threshing Machine for Human Activity Recognition Using Deep Learning Models", *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 1808990, 21 pages, 2022. doi: 10.1155/2022/1808990.

[29] M. Ronald, A. Poulose, and D. S. Han, "iSPLInception: an inception-ResNet deep learning architecture for human activity recognition," *IEEE Access*, vol. 9, pp. 68985–69001, 2021.

[30] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Device-free human activity recognition using commercial WiFi devices," IEEE Journal on Selected Areas in Communications, vol. 35, no. 5, pp. 1118–1131, 2017.

[31] F. Wang, W. Gong, and J. Liu, "On spatial diversity in WiFi-based human activity recognition: a deep learning-based approach," IEEE Internet of Things Journal, vol. 6, no. 2, pp. 2035–2047, 2019.

[32] Y. Wang, J. Wu, and H. Li, "Human detection based on improved mask R-CNN," *Journal of Physics: Conference Series*, vol. 1575, no. 1, Article ID 012067, 2020.