A Convolved Self-Attention Model for IMU-based Gait Detection and Human Activity Recognition

Shuailin Tao[1,2,3], Wang Ling Goh[1] and Yuan Gao[3]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Republic of Singapore
[2]AI-X, Interdisciplinary Graduate Programme, Nanyang Technological University, Republic of Singapore
[3]Institute of Microelectronics (IME), Agency for Science, Technology and Research (A*STAR), Singapore 138634, Republic of Singapore
Email: shuailin001@e.ntu.edu.sg, ewlgoh@ntu.edu.sg, gaoy@ime.a-star.edu.sg

2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS) | 979-8-3503-3267-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/AICAS57966.2023.10168654

***Abstract*—This paper presents a convolved self-attention neural network model for gait detection and human activity recognition (HAR) tasks using wearable inertial measurement unit (IMU) sensors. By embedding a convolved window inside the self-attention module, prior time step knowledge is utilized by self-attention layer to improve accuracy. Moreover, a streamlined fully connected (FC) layer without hidden layers is proposed for the feature mixer. This arrangement enables significant reduction of overall network parameters, since hidden layers occupy the majority of the parameters in a transformer encoder. Compared to the other state-of-art neural networks, the proposed method achieved better accuracy of 95.83% and 96.01% with the smallest network size on HAR datasets UCI-HAR and MHEALTH respectively,**

Keywords—Human Activity Recognition, Wearable sensor, Transformer Model, Time-series Data Processing

I. Introduction

Gait detection plays an important role in the diagnosis of various neurologic disorders [1]. It is also widely used in the rehabilitation progress assessment for stoke patient and the patient with lower limb amputation [2]. The current practice of gait detection are mainly image-based approaches [3] or radar systems [4]. The key limitation of image-based approach is the requirement of sufficient ambient illumination for image capture and the constraint on patient's movement to be within the camera's field of view. Recently, wearable sensor with Inertial Measurement Unit (IMU) draw a lot of attention since IMU can capture the details of gait motion which images cannot provide and more importantly, the system can be miniaturized as a wearable device for long-term continuous monitoring without constraints to the user [5].

On the other hand, various methods have been developed for human activity recognition (HAR) tasks. Adaptive methods such as peak detection and fuzzy logic model [6] achieved remarkable classification accuracy but they need to be optimized for specific application scenario. Threshold detection and time-frequency analysis [7] has requirement on hardware capability. Random Forest is suitable for hardware implementing on FPGA devices [8], but it has limited inference accuracy.

Recently, deep learning models such as Recurrent Neural Network (RNN) [9] and Convolution Neural Network (CNN) are applied for locomotion intent recognition [10, 11]. RNN models encounter challenges in retaining long-term prior time step information due to their limited memory capacity, which encompasses merely a few time steps. The problem is partially addressed by Long Short-Term Memory (LSTM) method, but at the cost of increased number of parameters to be learned. Because of the correlation between current gait status and that of prior time steps, prior time step information should be utilized to improve the classification performance. Self-attention model is more effective in capturing long-term dependencies than RNN models since self-attention can be shifted to certain part of the input sequence, without the constraint of sequential processing. For different applications and models, CNN needs to rearrange the feature map size. In contrast, self-attention is more flexible and it does not require padding and cropping.

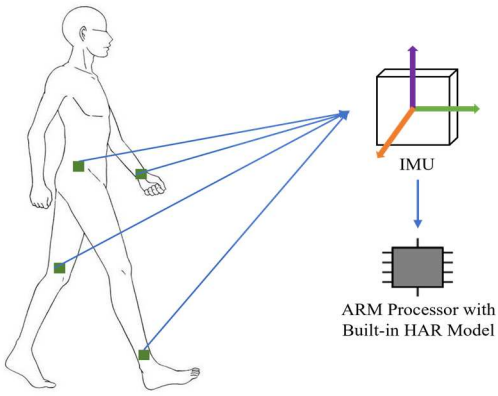

Fig. 1. Illustration of human activity recognition with IMU and edge device.

Generally speaking, transformer encoder using self-attention has two major limitations. First, the number of learning parameters is larger than other type of models. To reach the goal of online processing, it is challenging to control the size of parameter and computational cost for edge devices. Secondly, the performance of transformer model is not comparable to that of CNN when the size of training dataset is not sufficiently large [12]. The reason is that, different from CNN, transformer does not consider prior knowledge [14]. Recently, Swin-Transformer used a shifted window attention feature to provide additional data position information and it achieved better results in image processing tasks [14]. The data position information is also known as a typical prior knowledge. Since the feature map of sensor-based HAR has similar characteristics to image classification, hence it is

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on August 11,2023 at 14:20:20 UTC from IEEE Xplore. Restrictions apply.

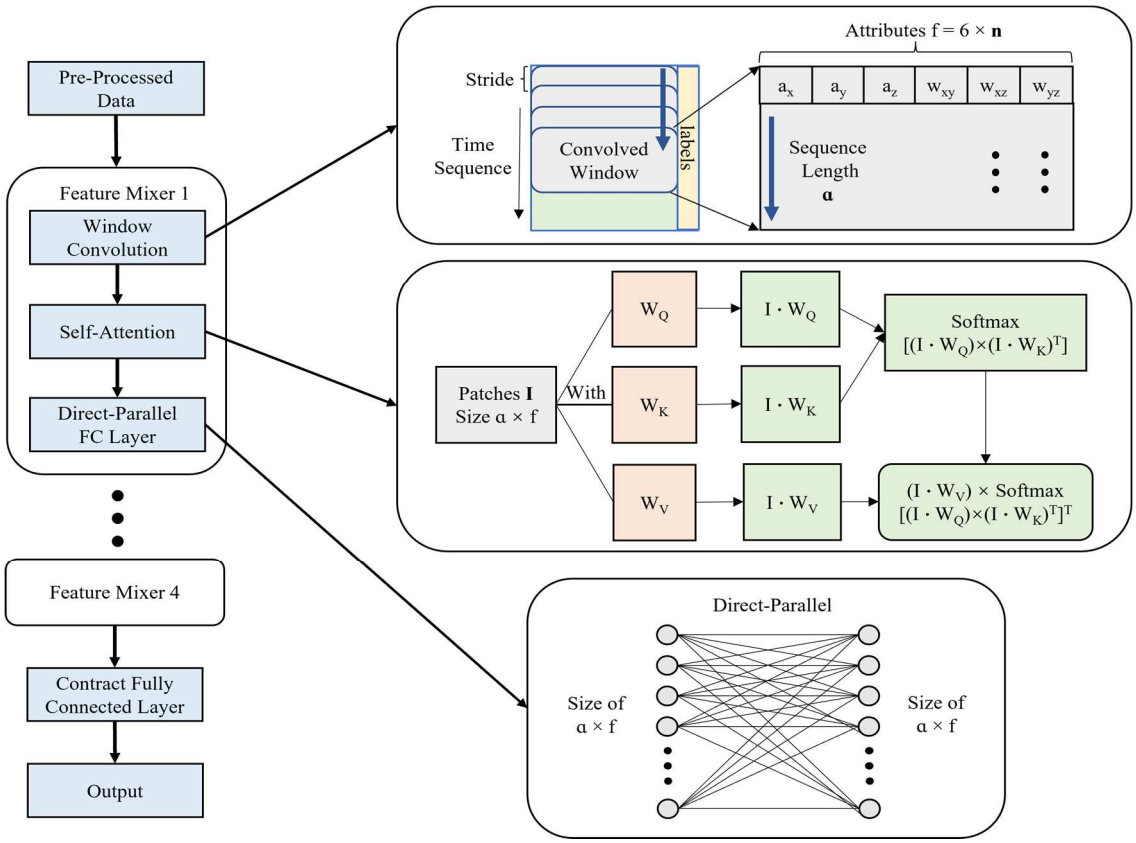

Fig. 2. Flow chart of the proposed self-attention model with convolved window.

promising to apply attention-based learning models to achieve improved accuracy and less processing latency for HAR tasks.

This paper presents a new convolved self-attention model for IMU-based gait detection and HAR. The proposed model incorporates convolved windows to provide prior time step knowledge to the self-attention operation. Furthermore, it requires the minimal number of learning parameters, rendering it suitable for implementation on edge devices. This paper is organized as follows, Section II introduces the proposed model and the implementation of key blocks; Section III presents the results with different datasets. Conclusion is summarized in Section IV.

II. Proposed Model

A. Overall Architecture

The flow chart of the proposed convolved window self-attention model is shown in Fig. 2. This model consists of four cascaded feature mixers and one contraction layer. Each feature mixer contains one convolved window layer, one self-attention layer and one direct-parallel fully connected (FC) layer. The convolved window captures consequential time steps into adjacent patches to provide local position information. The direct-parallel FC layer is a unique type of fully connected layer that lacks hidden layers and possesses equal input and output dimensions. Meanwhile, the contract FC layer executes the classification operation. During network training, activity classes are one-hot encoded into binary sequences. For each convolved window, the category label is set as the majority in each time step. The attention layer takes into each convolved window and allocate respective size's parameter matrix W_Q, W_K, and W_V to them.

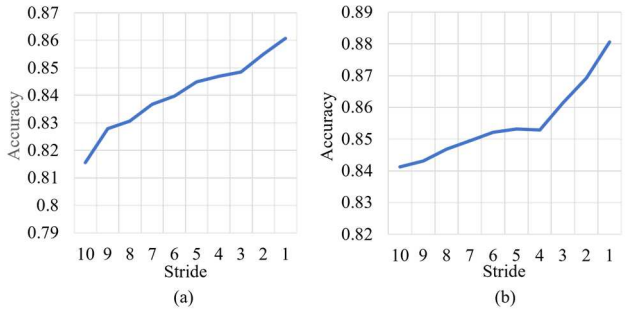

Fig. 3. The impact of convolved stride to prediction accuracy. (a) and (b) shows the result form IMU HAR dataset [5] and [10] accordingly.

B. Data Segmentation and Convolution

IMU data is in 2D array format with attributes by instances sequence. For example, the typical attributes for a 6-axis IMU are three-axis acceleration data (a_x, a_y, a_z) and three-axis angular velocity (w_{xy}, w_{yz}, w_{xz}). Other attributes are based on these features, such as minmax, standard derivation, and energy. Only 6 main attributes above are used for model training. Then the feature map has the size of $b\alpha \times 6n$, where b is the number of convolved windows, α refers to sequence length, while 6 is main attributes and n represents the number of IMU. Normalization is executed on each column to ensure each feature equally contributed to the classification result.

Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on August 11,2023 at 14:20:20 UTC from IEEE Xplore. Restrictions apply.

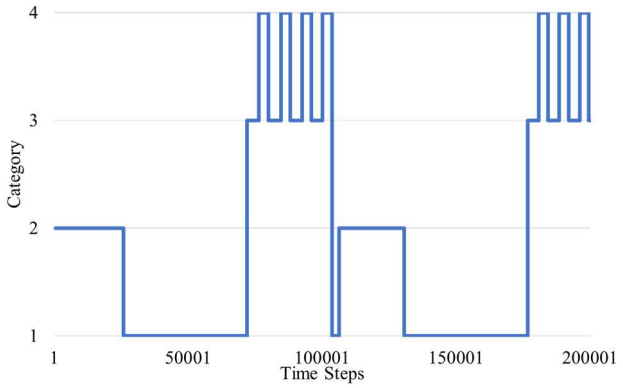

Fig. 4. Gait activity category with time step changes. Each time step is 100ms length. 1, 2, 3, and 4 represents W, US, DS, and R accordingly.

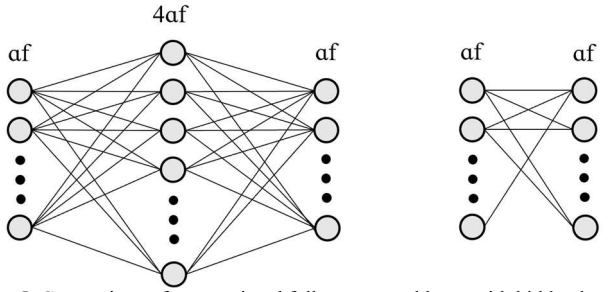

Fig. 5. Comparison of conventional fully connected layer with hidden layer and the streamlined FC layer. The local parameter amount reduces from $8f^2\alpha^2$ to $f^2\alpha^2$, which is 12.5% of the original model.

The feature map is not directly separated into patches and sent to models. Instead, a shifted window is used to record each data patch. The time sequence α is set as 10 and the stride varies from 1 to 10. As depicted in Fig. 3, the prediction accuracy generally increases with smaller convolved window strides. The upward trend is approximately linear, and the rate of increase is marginally higher at the beginning of window overlap and near full overlap. The total accuracy increases 4.51% in the first dataset and 3.93% in the second dataset comparing from direct separation to closest convolution.

Participants' locomotion patterns change infrequently with the natural time steps as shown in Fig. 4. Thus, random permutation between batch windows is necessary to reduce partial influence.

C. Feature mixer layers

After window convolution, patches are sent to the attention layer. Similar to conventional self-attention, query ($Q_{\alpha\times 6n}$), key ($K_{\alpha\times 6n}$), and value ($V_{\alpha\times 6n}$) are used as token mixers. W_Q, W_K, and W_V, three parameter matrices are set as the same size with input patches $\alpha \times 6n$ and perform dot product with input patches. Depending on datasets and applications, more attention layers can be added in series after layer normalization. The output of an attention layer in this model can be expressed as:

$$Attention(\alpha \times 6n) = \left(I \cdot W_V\right) \times Softmax\left[\left(I \cdot W_Q\right) \times (I \cdot W_K)^T\right]^T \quad (1)$$

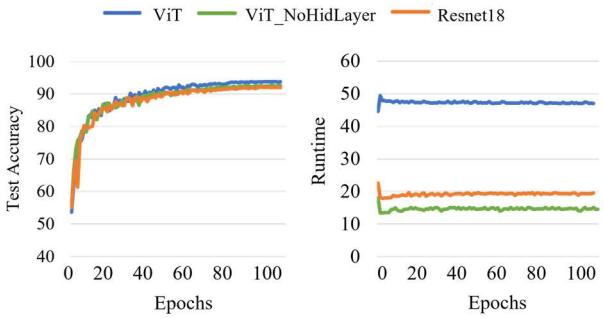

Fig. 6. Comparison of test accuracy and epoch run time of ViT, ResNet18, and the proposed model on CIFAR-10 dataset.

TABLE I CONFUSION MATRIX OF CLASSIFICATION TASK UCI-HAR DATASET

Activity	W	US	DS	ST	SD	L	Recall (%)
W	490	1	5	0	0	0	98.79
US	1	462	8	0	0	0	97.88
DS	0	12	408	0	0	0	97.14
ST	2	8	0	447	33	1	91.04
SD	0	10	3	41	534	0	90.82
L	0	0	0	0	0	537	100
Precision (%)	99.39	92.03	96.23	87.13	88.27	100	95.83

A direct-parallel FC layer with equal size of input and output is used after the attention layer and normalization as illustrated in Fig. 5. Different from conventional transformer encoder [13], the direct-parallel FC does not have hidden layers. Fully connected hidden layer occupies most parameters in transformer encoder structure but affect little on classification accuracy. Traditional transformer encoder uses a quadruple size hidden layer as FC input and output [13]. The parameter amounts in transformer encoder is shown in (2).

$$W_{TfEndoer} = hl \times (3f\alpha + 8f^2\alpha^2) + f\alpha t \quad (2)$$

After FC layer simplification, the amount of parameter is reduced to

$$W_{New} = hl \times (3f\alpha + f^2\alpha^2) + f\alpha t \quad (3)$$

Where h is the number of multi-head amounts, l is the number of encoder layer, f is the feature size, a is the sequence size, and t is the classification types. $3f\alpha$ is the amount of single attention layer parameters, which refers to W_Q, W_K, and W_V. $8f^2\alpha^2$ is the number of parameters in fully connected layer after the self-attention layer. It can be observed that the reduction from $8f^2\alpha^2$ to $f^2\alpha^2$ greatly reduces the total parameter amounts, and results will be shown in the following experiment session.

III. EXPERIMENT AND DISCUSSION

A. Streamlined FC Layer

Firstly, the impact of the proposed streamlined FC layer to the overall model performance is investigated. To validate the generalizability of the approach of reducing (FC) layers, a study is conducted to examine its effectiveness across datasets in different field. A standard Vision Transformer (ViT) is used as the test vehicle [12]. The fully connected layer with hidden

Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on August 11,2023 at 14:20:20 UTC from IEEE Xplore. Restrictions apply.

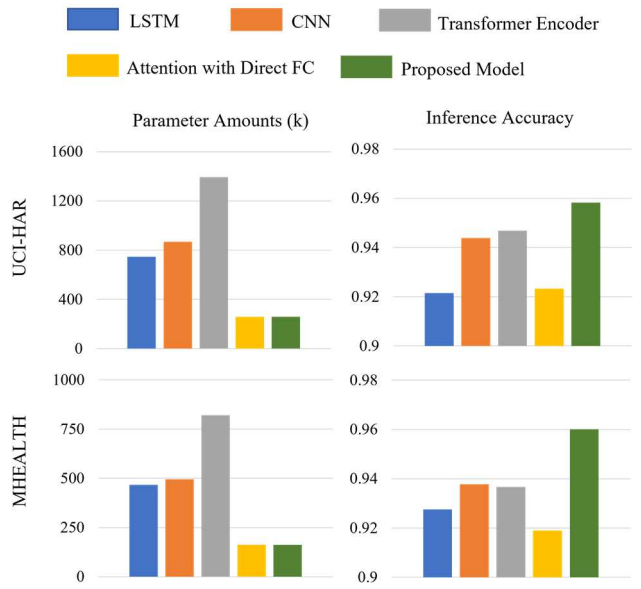

Fig. 7. Comparison of different model parameter amount and inference accuracy for HAR datasets.

layer in feature mixer is replaced with the proposed streamlined FC layer. CIFAR-10 image dataset is used to validate the performance because it has similar features to sensor-based data in the format of 2D arrays. Pre-trained results with JFT-300M [12] are used for parameter initiation.

As shown in Fig. 6, the accuracy for standard ViT, ResNet18 [15] and the modified ViT with streamlined FC layer are 93.74%, 92.62%, and 91.98%. Compared to the original ViT, less than 2% accuracy degradation is achieved with 68.07% parameter amounts reduced and the average running time for each epoch is reduced from 48s to 14s.

B. Model Performance on HAR Datasets

Next step, the complete self-attention model including convolved window and direct-parallel FC layer is evaluated with two publicly available HAR datasets UCI-HAR and MHEALTH. The convolved window does not cost additional parameter amount but only increases the training step time. Stride 1 is used for the convolved window. The parameter amount and inference accuracy are evaluated in comparison to other state-of-art models. The number of layer and layer sizes are adjusted to attain comparable accuracy, ensuring a fair comparison of parameter amounts. Fig. 7 shows when the inference accuracy reach similar score, the proposed model requires the lowest parameter amounts, which is 65.23% less than LSTM, 70.07% less than CNN, and 81.34% less than transformer encoder.

Recently there are a few hybrid models reported for HAR tasks [16-20]. Convolutional layers provide prior knowledge and more local relevancy, and recurrent based neural networks such as GRU and LSTM are fit for sequence data. The combination of these models gives better accuracy but at the cost of increased parameter size. The performance of the proposed model is compared with these hybrid models. The results are summarized in Table II and Table III for HCI-HAR and MHEALTH datasets, respectively. From the tables, we can observe that the proposed model achieved 95.83% and 96.01% accuracy with lowest parameter amounts as CNN-GRU method reaches similar accuracy, but it uses the highest number of hidden layers with 2.93 times size of parameters. RCNN with attention method can perform semi-supervised learning, but the accuracy is 14.51% lower than the proposed model.

TABLE II A COMPARATIVE ANALYSIS OF HCI-HAR DATASET

	[16]	[17]	[18]	[19]	[20]	**This Model**
Topology	CNN-LSTM	RCNN with Attention	CNN-GRU	Binary AOA	CNN-LSTM with self-attention	**Convolved Attention**
Epochs	30	100	100	3000	150	**100**
Number of Heads	1	2	3	1	1	**2**
Number of Classes	6	6	6	6	6	**6**
Number of Layers	6	10	14	9	13	**4**
Number of Parameters	280k	513k	761k	442k	634k	**260k**
F1 Score	91.89%	80.18%	94.39%	95.33%	93.11%	**95.08%**
Accuracy	92.13%	81.32%	94.48%	95.23%	93.11%	**95.83%**

TABLE III A COMPARATIVE ANALYSIS OF MHEALTH DATASET.

	[16]	[17]	[18]	[19]	[20]	**This Model**
Topology	CNN-LSTM	RCNN with Attention	CNN-GRU	Binary AOA	CNN-LSTM with self-attention	**Convolved Attention**
Epochs	30	100	100	3000	150	**100**
Number of Heads	1	2	3	1	1	**2**
Number of Classes	12	4	12	12	12	**12**
Number of Layers	6	10	14	9	13	**4**
Number of Parameters	177k	351k	525k	340k	449k	**162k**
F1 Score	92.54%	93.31%	94.29%	95.76%	93.79%	**95.87%**
Accuracy	93.30%	94.05%	94.55%	95.80%	94.09%	**96.01%**

IV. CONCLUSION

This paper presents a novel self-attention neural network model for IMU-based gait detection and human activity recognition. This model employs convolved windows to supply prior time-step knowledge to attention layer and utilizes a streamlined FC layer without hidden layer to reduce parameter amount. Owing to the incorporation of the convolved window and the streamlined direct fully connected layer, the proposed model achieves 95.83% and 96.01% accuracy on the UCI-HAR and MHEALTH datasets, respectively. This performance is attained with the smallest parameter size compared to other state-of-the-art models, rendering the model particularly suitable for implementation on edge devices.

ACKNOWLEDGEMENT

This work was supported by Agency for Science, Technology and Research (A*STAR), Singapore under the Nanosystems at the Edge programme (Grant No. A18A1b0055)

Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on August 11,2023 at 14:20:20 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] W. Shao et al., "A multi-modal gait analysis-based detection system of the risk of depression," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 10, pp. 4859-4868, Oct. 2022.

[2] B. -Y. Su, J. Wang, S. -Q. Liu, M. Sheng, J. Jiang and K. Xiang, "A CNN-based method for intent recognition using inertial measurement units and intelligent lower limb prosthesis," *IEEE Trans. Neural Syst. Rehabilitation Eng.*, vol. 27, no. 5, pp. 1032-1042, May 2019.

[3] X. Gu, Y. Guo, G. -Z. Yang and B. Lo, "Cross-domain self-supervised complete geometric representation learning for real-scanned point cloud based pathological gait analysis," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 3, pp. 1034-1044, Mar. 2022.

[4] K. Long, C. Rao, X. Zhang, W. Ye and X. Lou, "FPGA accelerator for radar-based human activity recognition," *IEEE Int. Conf. Artif. Intell. Circuits and Syst. (AICAS)*, Incheon, Korea, Republic of, pp. 391-394, 2022

[5] M. Zhang, Q. Wang, D. Liu, B. Zhao, J. Tang and J. Sun, "Real-time gait phase recognition based on time domain features of multi-MEMS inertial sensors," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-12, 2021.

[6] Y. C. Han, K. I. Wong and I. Murray, "Gait phase detection for normal and abnormal gaits using IMU," *IEEE Sens. J.*, vol. 19, no. 9, pp. 3439-3448, 1 May. 2019.

[7] Y. Yang, L. Chen, J. Pang, X. Huang, L. Meng and D. Ming, "Validation of a spatiotemporal gait model using inertial measurement units for early-stage parkinson's disease detection during turns," *IEEE. Trans. Biomed. Eng.*, vol. 69, no. 12, pp. 3591-3600, 2022.

[8] D. Watanabe et al., "An architectural study for inference coprocessor core at the edge in IoT sensing," *IEEE Int. Conf. Artif. Intell. Circuits and Syst. (AICAS)*, Genova, Italy, pp. 305-309, 2020.

[9] F. Sherratt, A. Plummer, and P. Iravani, "Understanding LSTM network behaviour of IMU-based locomotion mode recognition for applications in prostheses and wearables," *Sensors*, vol. 21, no. 4, p. 1264, Feb. 2021.

[10] I. Klein, "Smartphone location recognition: A deep learning-based approach," *Sensors*, vol. 20, no. 1, p. 214, Dec. 2019.

[11] O. Dehzangi, M. Taherisadr, and R. ChangalVala, "IMU-based gait recognition using convolutional neural networks and multi-sensor fusion," *Sensors*, vol. 17, no. 12, p. 2735, Nov. 2017.

[12] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," *Proc. Int. Conf. Learn. Represent. (ICLR)*, pp. 1–11, 2021.

[13] T. B. Brown et al. , "Language models are few-shot learners," *Adv. Neural Inf. Process. Syst. (NIPS)*, 2020.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[16] D. Nidhi, S. N. Singh and V. B. Semwal, "Multi-input CNN-GRU based human activity recognition using wearable sensors," *Comput. Arch. Inform. and Numer. Comput. (Computing)*, vol. 103, (7), pp. 1461-1478, 2021.

[17] R. Mutegeki and D. S. Han, "A CNN-LSTM approach to human activity recognition," *Int. Conf. Artif. Intell. Inform. and Commun. (ICAIIC)*, Fukuoka, Japan, pp. 362-366, 2020.

[18] M. A. Khatun et al., "Deep CNN-LSTM with self-attention model for human activity recognition using wearable sensor," *IEEE J. Transl. Eng. Health Med.*, vol. 10, pp. 1-16, 2022.

[19] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang and F. Nie, "A semisupervised recurrent convolutional attention model for human activity recognition," *IEEE Trans. Neural. Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1747-1756, May. 2020.

[20] A. Dahou, M. A. A. Al-qaness, M. Abd Elaziz, and A. Helmi, "Human activity recognition in IoHT applications using arithmetic optimization algorithm and deep learning," *Meas.*, vol. 199, p. 111445, 2022.

Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on August 11,2023 at 14:20:20 UTC from IEEE Xplore. Restrictions apply.