RESEARCH ARTICLE

REVISED MSTCN: A multiscale temporal convolutional network for user independent human activity recognition [version 2; peer review: 2 approved, 1 approved with reservations]

Sarmela Raja Sekaran[1], Ying Han Pang[1], Goh Fan Ling[2], Ooi Shih Yin[1]

[1]Faculty of Information Science and Technology, Multimedia University, Ayer Keroh, Melaka, 75450, Malaysia
[2]Millapp Sdn Bhd, Bangsar South, Kuala Lumpur, 59200, Malaysia

v2 **First published:** 08 Dec 2021, **10**:1261
https://doi.org/10.12688/f1000research.73175.1
Latest published: 18 May 2022, **10**:1261
https://doi.org/10.12688/f1000research.73175.2

Abstract

Background: In recent years, human activity recognition (HAR) has been an active research topic due to its widespread application in various fields such as healthcare, sports, patient monitoring, etc. HAR approaches can be categorised as handcrafted feature methods (HCF) and deep learning methods (DL). HCF involves complex data pre-processing and manual feature extraction in which the models may be exposed to high bias and crucial implicit pattern loss. Hence, DL approaches are introduced due to their exceptional recognition performance. Convolutional Neural Network (CNN) extracts spatial features while preserving localisation. However, it hardly captures temporal features. Recurrent Neural Network (RNN) learns temporal features, but it is susceptible to gradient vanishing and suffers from short-term memory problems. Unlike RNN, Long-Short Term Memory network has a relatively longer-term dependency. However, it consumes higher computation and memory because it computes and stores partial results at each level.
Methods: This work proposes a novel multiscale temporal convolutional network (MSTCN) based on the Inception model with a temporal convolutional architecture. Unlike HCF methods, MSTCN requires minimal pre-processing and no manual feature engineering. Further, multiple separable convolutions with different-sized kernels are used in MSTCN for multiscale feature extraction. Dilations are applied to each separable convolution to enlarge the receptive fields without increasing the model parameters. Moreover, residual connections are utilised to prevent information loss and gradient vanishing. These features enable MSTCN to possess a longer effective history while maintaining a relatively low in-network computation.
Results: The performance of MSTCN is evaluated on UCI and WISDM datasets using a subject independent protocol with no overlapping subjects between the training and testing sets. MSTCN achieves accuracies of 97.42 on UCI and 96.09 on WISDM.

Open Peer Review

Approval Status ✓ ? ✓

	1	2	3
version 2 (revision) 18 May 2022	✓ view		✓ view
version 1 08 Dec 2021	? view	? view	

1. **Cheng-Yaw Low**, Institute for Basic Science, Seoul, South Korea
2. **Sultan Daud Khan**, National University of Technology, Islamabad, Pakistan
3. **Xinghua Li**, Wuhan University, Wuhan, China

Any reports and responses or comments on the article can be found at the end of the article.

Conclusion: The proposed MSTCN dominates the other state-of-the-art methods by acquiring high recognition accuracies without requiring any manual feature engineering.

Keywords
human activity recognition, smartphone, temporal convolutional network, dilated convolution, one-dimensional inertial sensor

This article is included in the Research Synergy Foundation gateway.

Corresponding author: Ying Han Pang (yhpang@mmu.edu.my)

Author roles: Raja Sekaran S: Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Pang YH**: Data Curation, Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing; **Ling GF**: Data Curation, Software; **Yin OS**: Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This study was funded by the Fundamental Research Grant Scheme (FRGS) from the Ministry of Education Malaysia – FRGS/1/2020/ICT02/MMU/02/7. The funder had no role in study design, data collection and analysis, decision to publish, or manuscript preparation.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Raja Sekaran S, Pang YH, Ling GF and Yin OS. **MSTCN: A multiscale temporal convolutional network for user independent human activity recognition [version 2; peer review: 2 approved, 1 approved with reservations]** F1000Research 2022, **10**:1261 https://doi.org/10.12688/f1000research.73175.2

First published: 08 Dec 2021, **10**:1261 https://doi.org/10.12688/f1000research.73175.1

REVISED **Amendments from Version 1**

In the revised version, the work's problem statement and contributions have been revised to ensure a clearer description. We have also revised Figures 1, 2, 3, 4 and 5 to provide more details. In addition, the authors have included another literature work in the Related Work section. Further, the Methods and Results sections have been separated into two sections. The Method section has been revised to include a detailed explanation regarding the architecture of the proposed model, mainly the justification of the use of the Inception model, 1×1 convolution, dilated convolution, feature concatenation, global average pooling, causality, regularisation, softmax equation and categorical cross-entropy function. Besides, the Experiment and Results section has been revised to include input dimensionality and hyperparameter values (optimizer, learning rate, batch size, etc.) used in the proposed model. Finally, the Discussion section was revised and written in a paragraph.

Any further responses from the reviewers can be found at the end of the article

Introduction

Human activity recognition (HAR) is extensively applied in various applications such as personal health monitoring,[1,2] geriatric patient monitoring,[3] ambient assisted living,[4] etc. The widespread use of smartphone-based HAR is due to the ubiquity of smartphones and low-cost sensors. Additionally, sensor-based HAR provides a non-intrusive solution.

Over the years, numerous algorithms have been proposed, including handcrafted feature (HCF) methods[5-7] and deep learning (DL) methods.[8,9] HCF methods require complex signal pre-processing and manual feature engineering to extract essential features. In contrast, DL methods, such as convolutional neural network (CNN),[8,9] recurrent neural network (RNN), and long-short term memory network (LSTM),[10,11] can automatically extract crucial discriminative features from input signals without manual feature engineering. Besides, the architecture is adaptable to different applications.

Though the existing methods produce satisfactory performances, there are several challenges which hinder the HAR models from achieving potential performance:

- HCF methods require manual feature extraction where the extracted features are highly dependent on prior knowledge. This may lead to high bias and missing of essential implicit patterns.

- CNN is good at extracting spatial features. It is suboptimal in learning temporal features. Temporal features are crucial in motion analysis.

- Although recurrent models are feasible for time-series data, RNN is prone to short-term memory problems, leaving out important information at the beginning if the input sequence is too long.

- LSTM prevails over RNN. LSTM has a longer-term dependency and is less susceptible to vanishing gradient. However, LSTM requires higher computation due to multiple gate operations and more memory to store partial results throughout the training phase.

To address the aforementioned challenges, this work proposes a multiscale temporal convolutional network (MSTCN) for HAR. As illustrated in Figure 1, MSTCN is constituted by multiscale dilation (MSD) blocks, global average pooling and softmax. The details of the components will be described in the later section. The contributions of this work are:

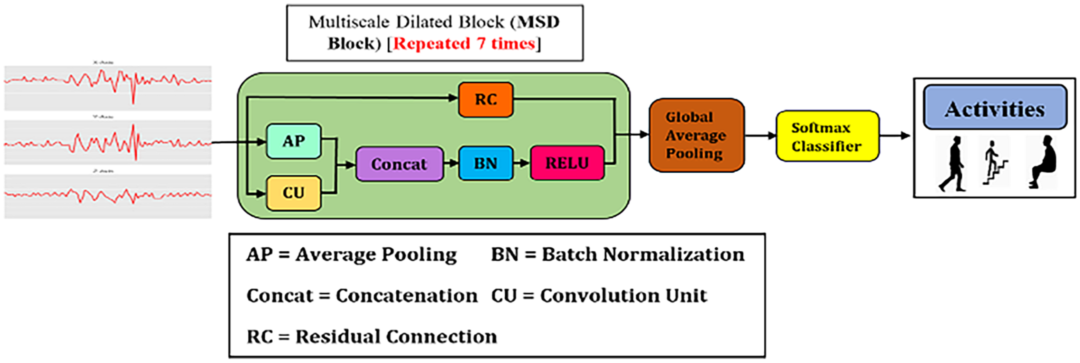

Figure 1. **Architecture of MSTCN.**

- A deep analytic model, amalgamating Inception model and Temporal Convolutional Network (TCN), is developed to extract spatial-temporal features from inertial data. MSTCN requires minimal data pre-processing and no manual feature engineering.

- MSTCN incorporates multiple different-sized convolutions to perform multiscale feature extraction. These multiscale features provide richer information for data analysis.

- To retain longer effective history, dilated convolution is implemented to increase the receptive field without raising the overall parameters.

- A comprehensive experimental analysis is conducted using two popular public databases, UCI[5] and WISDM.[12] Subject independent protocol is implemented where different subjects are used for training and testing. In other words, there is no overlap in subject in the training and test sets.

Related work

One-dimensional inertial data undergoes a complicated pre-processing in HCF methods to extract salient statistical feature vectors in time and/or frequency domains. The manually extracted features are then fed into standard machine learning classifiers, such as support vector machine (SVM),[5,6] ADA Boost,[7] Random Forest,[13] C4.5 decision tree,[14] etc., for activity classification. He and Jin[15] proposed a discrete cosine transform method to extract features and classify the features using multiclass SVM. Lara *et al.*,[16] developed an additive logistic regression, boosting with an ensemble of 10 decision stump classifiers. In the works of Ronao and Cho,[17,18] the authors explored the Continuous Hidden Markov Model (HMM) to perform activity recognition in two stages, where the first stage is for static and dynamic classification and the second stage is for course classification. Although these methods produce satisfactory performances, they are highly dependent on the effectiveness of the manual feature engineering techniques.

Recently, researchers leaned towards DL methods since DL requires minimal to zero pre-processing and feature engineering. Ronao *et al.*,[8] Yazdanbakhsh *et al.*,[9] and Huang *et al.*,[19] proposed a CNN-based deep learning system to perform HAR. The reported empirical results show the feasibility of the CNN-based method in analysing motion data. Besides, three-layer LSTM was proposed to classify human activities.[20] In addition, Ullah *et al.* proposed a HAR algorithm that classified the normalised inertial data signals using stacked LSTM into respective classes.[11] Further, LSTM variant, known as Bidirectional LSTM, was employed in HAR.[10] This model uses richer information, i.e. previous and subsequent information, to perform activity recognition. Nair *et al.*, proposed two variations of TCN, namely Dilated-TCN and Encoder-Decoder TCN in HAR.[21] In addition, another two TCN-based models are proposed in Ref. 22, namely TCN-FullyConnectedNetwork and deepConvTCN. Both works of Nair *et al.*,[21] and Garcia *et al.*,[22] concluded that the TCN-based models achieved better performance than existing recurrent models in HAR application due to the longer-term dependencies in TCN.

Methods

In the proposed HAR, the raw inertial signals were firstly pre-processed to remove noise. Next, the pre-processed signals were segmented using sliding window technique. In specific, the signals were partitioned into fixed-sized time windows and each window did not intersect with another window. Then, the segmented data was fed into MSTCN for feature analysis and classification. MSTCN comprises of MSD blocks (green box in Figure 1), global average pooling and softmax classifier.

Figure 2 illustrates the structure of a MSD block, comprising convolution unit (CU), average pooling, residual connection, batch normalization etc. The design of MSD is inspired by Inception module[23] in such a way that multiple kernels/filters are applied simultaneously to the input time series data, as shown in the CU in Figure 3. These kernels are in varying lengths which allow multiscale feature extraction, i.e. extracting features from short and long time series.[24] In the subsequent MSD blocks, the input of CU is processed via one-by-one causal convolution for channel-wise pooling and dimensionality reduction.[25] The padding preserves the input sequence's length and order, preventing information leakage from the future into the past. Next, the produced feature maps are further processed parallelly by separable convolutions (SepConv) with three different-sized filters to extract features at multiple scales. The ordinary Inception module is using multiple standard convolutions with smaller kernel sizes, i.e., 3 and 5.[23] However, bigger kernel sizes are required in HAR application in order to capture longer time series and preserve longer-term dependencies of the input.[24] The authors also claimed that the increasing kernel size leads to the rise of the number of network parameters, which may cause overfitting of the model. Hence, SepConv was used since it reduces the number of parameters in convolution process, while demanding lesser memory compared to standard convolutions.[26] Figure 4 shows the operation of SepConv through decoupling standard convolution.

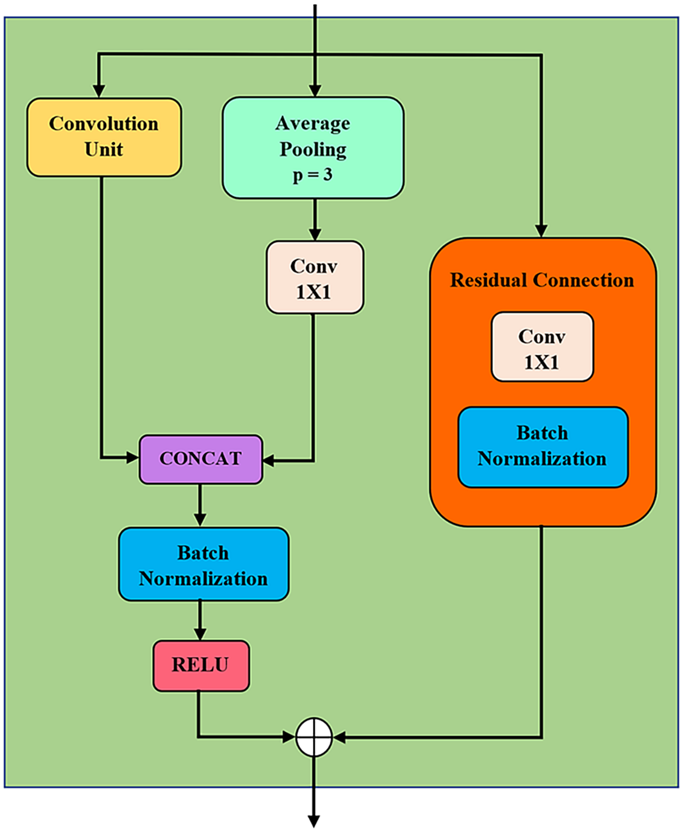

Figure 2. **MSD block.** (concat = concatenation, conv = convolution and p = pooling factor).

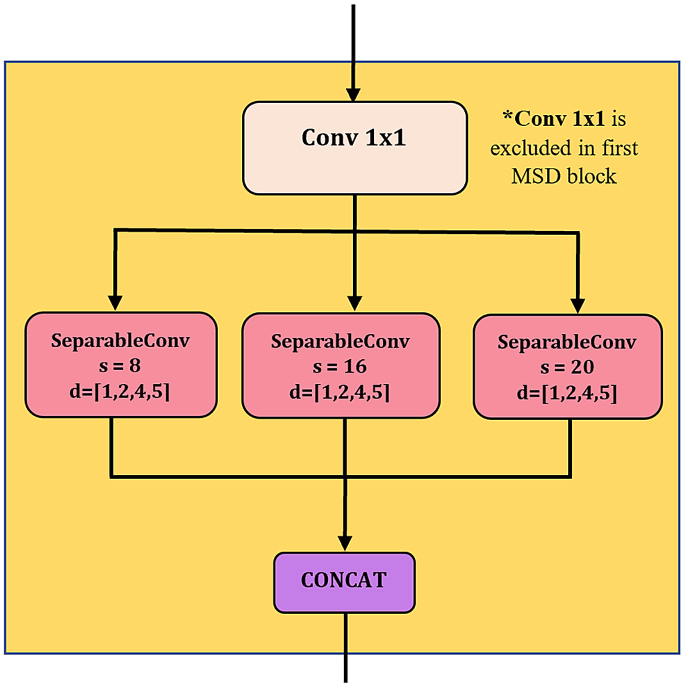
Conv 1x1
*Conv 1x1 is excluded in first MSD block
SeparableConv s = 8 d=[1,2,4,5]
SeparableConv s = 16 d=[1,2,4,5]
SeparableConv s = 20 d=[1,2,4,5]
CONCAT

Figure 3. **Convolutional unit in MSD block.** (concat = concatenation, conv = convolution, s = kernel size and d = dilation rate).

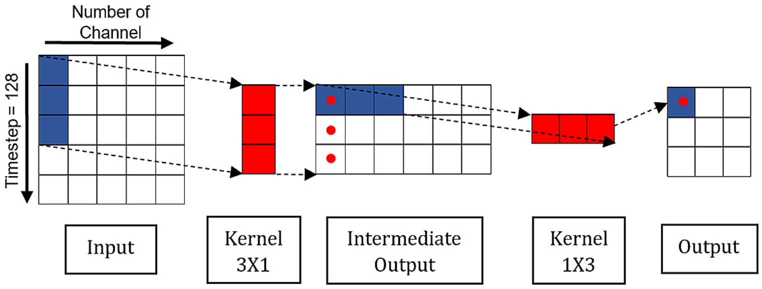

Figure 4. Separable convolution.

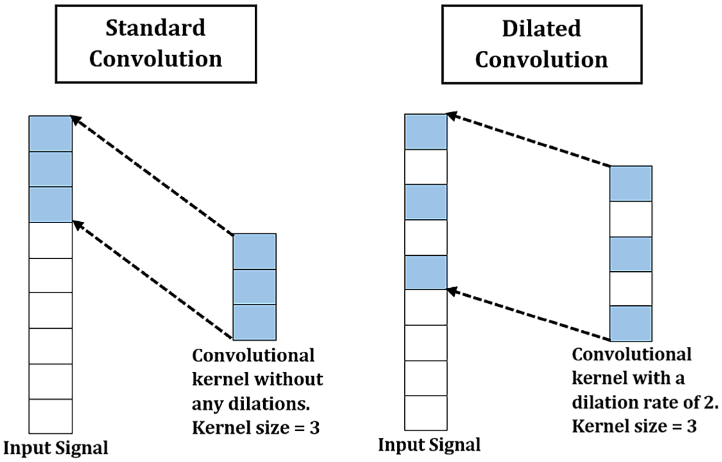
Standard Convolution
Dilated Convolution
Convolutional kernel without any dilations. Kernel size = 3
Input Signal
Convolutional kernel with a dilation rate of 2. Kernel size = 3
Input Signal

Figure 5. Comparison between standard and dilated convolution.

One of the ways to capture longer time dependent features is by introducing dilations to the convolutions for improving the receptive fields without drastically increasing the model's parameters.[27] The difference between dilated and standard convolution is shown in Figure 5. Receptive field, or field of view, is the region of an input space which is visible to a convolution kernel at a time. A model can capture longer underlying patterns from input data using a convolution kernel with a larger receptive field. The receptive field size of a kernel can be enlarged by increasing the dilation rate. Hence, dilated convolutions were applied in this work to enlarge the receptive field without requiring extra parameters. After the parallel convolutions, the produced feature maps of each SepConv are concatenated by stacking them together, see Figure 3.

In the MSD block, average pooling (in Figure 2) down-samples the feature map to reduce noise and dimensionality. Additionally, it also preserves localisation. The pooling's output is fed into a one-by-one convolution. Next, the features of CU are stacked with the one-by-one convolution output. As illustrated in Figure 2, a residual connection is formed by passing the input into a one-by-one convolution, followed by a batch normalisation. This residual connection ensures longer-term dependencies and prevents information loss. Further, it also reduces the vanishing gradient effects. On the other hand, batch normalisations in MSD block are to reduce the internal covariate shift in the model during training. Furthermore, ReLU activation is chosen for its non-linearity, and the gradient vanishing is minimised.

The features extracted from the series of MSD blocks are further fed into the global average pooling (GAP) for feature pooling. Next, softmax classifier is implemented for data classification. The softmax activation formula for the i^{th} input vector, $\sigma(z)_i$, is defined:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{1}$$

where z_i is the i^{th} input vector, $e^{(z_i)}$ is the exponential function of the i^{th} input vector, K is the number of classes and $e^{(z_j)}$ is the exponential function of the j^{th} output vector. This function outputs a probability of each human activity class, ranging from zero to one, and the target/predicted class will have the highest probability. Then, softmax loss is computed by implementing categorical cross-entropy loss function to the softmax output.

$$CE_{general} = -\sum_{i}^{K} t_i \log z_i \tag{2}$$

where t_i are the ground truths and z_i are the predicted values for i^{th} class in classes K

$$CE_{softmax} = -\log \frac{e^{z_p}}{\sum_j^K e^{z_j}} \tag{3}$$

where z_p is the softmax score for the positive class p. The details can be referred to Ref. 28.

Experiments and results

Model configuration and experimental setup

The proposed MSTCN was implemented using Tensorflow, an open-source machine learning platform, with Keras library (a high-level deep learning API written in Python). MSTCN is learned for 100 epochs according to the parameter settings in Table 1. These parameters were fine-tuned based on the validation data from the training set with 10% random data of the training samples.

The experiments were conducted on a desktop with Intel® Core™ i7-8750H CPU with 2.20 GHz, 16GB RAM and NVIDIA GeForce GTX 1050 Ti with Max-Q Design and 4GB memory. Two public databases, UCI[5] and WISDM[12] were used to assess the reliability of the proposed model. In this work, subject independent protocol was implemented to facilitate impersonal solution. There is no overlap in subject between the training and testing sets. This protocol is relatively challenging since there are some extent of discrepancies of gaits towards the motion patterns in same activities. Details of the databases are recorded in Table 2. The evaluation metrics used in this work include precision, recall, F1 score and classification accuracy.

Table 1. Parameter settings of the proposed model.

	UCI	WISDM
Input dimension	(128,9)	(128,3)
Batch size	64	64
Number of MSD blocks	7	7
Number of filters	64	64
Filter size	8, 16 and 20	8, 16 and 20
Dilation rate	1, 2, 4 and 5	1, 2, 4 and 5
Stride	1	1
Regularisation	L1 and L2	L1 and L2
Number of epoch	100	100
Initial learning rate	0.001	0.001
Reduce learning rate on plateau function	Patience: 5 Minimum learning rate: 0.0001 Factor: 0.5 Mode: Validation loss	Patience: 5 Minimum learning rate: 0.0001 Factor: 0.5 Mode: Validation loss
Optimizer	Adam	Adam
Loss function	Categorical cross-entropy	Categorical cross-entropy

Table 2. Description of UCI and WISDM datasets.

	UCI	WISDM
Sensor	Accelerometer and Gyroscope	Accelerometer
Segment size	128 ms^{-2}	128 ms^{-2}
Segment interval	50 ms^{-2}	20 ms^{-2}
Channel size	9	3
Activities (class labels)	Walking, Upstairs, Downstairs, Sitting, Standing and Laying	Walking, Jogging, Upstairs, Downstairs, Sitting and Standing
Training testing split	21 training users: 9 testing users	31 training users: 5 testing users
Validation split	10% of the training set	10% of the training set

Table 3. Performance of MSTCN using different convolutions.

	Dilated 1D causal convolution	Dilated 1D separable convolution
Number of parameters	6 062 086	3 750 406
Precision	0.9357	0.9764
Recall	0.9375	0.9744
F1 score	0.9356	0.9747
Accuracy	93.62	97.42

Experiments

Experiments were conducted on UCI dataset to study the effects of (1) convolution, (2) pooling and (3) regularisation on MSTCN's performance. Table 3 shows the proposed model's performances using dilated one-dimensional (1D) causal convolution (CC) and dilated 1D separable convolution (SC). From the empirical results, it was observed that the parameters of SC are approximately half of the parameters of CC. Usually, models with more parameters perform better since maximal data patterns are captured and learned. However, when the training sample size is limited, these models might tend to overfit and not generalise properly to the unseen data, leading to poor performance. In this study, SC obtains ~4% higher accuracy than CC.

Next, the performances of max-pooling and average pooling were studied. From Table 4, average pooling excels max-pooling with ~3% higher accuracy. Average pooling performs better in this domain because it takes every value into account. With this, the information leakage is prevented, and feature localisation is preserved.

Table 5 shows the performance of MSTCN with different regularisation settings. The regularisation is performed at the one-by-one causal convolution in MSTCN. L1 is good at dealing with outliers since it takes the absolute values of all the weight instead of squared value.[35] On the other hand, L2 forces weights toward zero, but never exactly zero. The non-sparseness of L2 is useful as a prediction performance. By combining the usage of L1 and L2, we can leverage the benefits of both with achieving ~97.5% accuracy.

Table 4. Performance of MSTCN using different pooling layers.

	Max pooling	Average pooling
Precision	0.9478	0.9764
Recall	0.9468	0.9744
F1 score	0.9463	0.9747
Accuracy	94.67	97.42

Table 5. **Performance of MSTCN using different regularisation settings.**

	L1	L2	L1 and L2	Without regularisation
Precision	0.9485	0.9666	0.9764	0.9529
Recall	0.9464	0.9650	0.9744	0.9521
F1 score	0.9459	0.9649	0.9747	0.9517
Accuracy	94.60	96.44	97.42	95.28

Further, we also conducted the performance comparison between the proposed MSTCN and the other state-of-the-art methods. Tables 6 and 7 records the classification accuracy performance of the methods on UCI and WISDM datasets, respectively.

Table 6. **Accuracy for user independent UCI dataset.**

	Type	Accuracy (%)
Statistical features + SVM[5]	HCF	96.00
Statistical features + Continuous HMM[17]	HCF	91.76
Statistical features + HMM Ensemble[29]	HCF	83.51
Statistical features + RF[13]	HCF	78.00
Statistical features + Linear SVM[6]	HCF	86.00
Statistical features + Hierarchical Continuous HMM[18]	HCF	93.18
Statistical features + Dropout Classifiers[30]	DL	~76.00
Statistical features + Data Centering + CNN[31]	DL	97.63
CNN[8]	DL	94.79
Frequency features + CNN[8]	DL	95.75
Bidirectional LSTM[10]	DL	93.79
Dilated TCN[21]	DL	93.80
Encoder-Decoder TCN[21]	DL	94.60
Statistical features + MLP[32]	DL	95.00
Frequency and Power features + Multichannel CNN[33]	DL	95.25
Statistical features + InnoHAR[25]	DL	94.50
Stacked LSTM[11]	DL	93.13
MSTCN (Proposed Method)	DL	97.42

Table 7. **Accuracy for user independent WISDM dataset.**

Methods	Type	Accuracy (%)
Statistical features + RF[30]	HCF	83.46
Statistical features + RF[13]	HCF	83.35
Statistical features + Dropout Classifiers[30]	DL	85.36
Statistical features + CNN[31]	DL	93.32
Dilated and Strided CNN[9]	DL	88.27
Data Augmentation + Two Stage End-to-End CNN[19]	DL	84.60
Statistical features + CNN[34]	DL	94.18
MSTCN (Proposed Method)	DL	96.09

Discussion

MSTCN prevails over HCF methods on both datasets because the proposed model can better capture discriminating features from the motion data. Unlike handcrafted features, these deep features are less biased as they are not dependent on prior knowledge. This is crucial, especially for a subject independent solution. Furthermore, MSTCN outperforms most CNN-based approaches, with accuracy scores of ~97% in UCI and ~96% in WISDM. This performance exhibits that the competence of MSTCN in extracting features from the data at assorted scales via the application of different convolutional filter sizes. Besides, GAP in MSTCN not only performing feature pooling, but also minimizes overfitting since there is no parameter to be learned in the GAP.[36] This is relatively suitable for subject independent HAR solution since testing data is new/unseen data. Moreover, MSTCN dominates the recurrent model[10,11] due to its ability in modelling longer-term dependencies via dilated convolution. Further, residual connections and ReLU activations in MSTCN allow the model to be less susceptible to gradient vanishing and exploding. MSTCN is a TCN-variant model. The obtained empirical results demonstrate that MSTCN outperforms the ordinary TCNs (Dilated TCN and Encoder-Decoder TCN).[21] MSTCN learns features at multiple scales via different convolutions with differently sized filters. These multiscale features provide richer information for data analysis.

Conclusions

A new deep analytic model, known as MSTCN, is proposed for subject independent HAR. MSTCN is based on the architectures of the Inception network and temporal convolutional network. In MSTCN, different-sized filters are adopted in dilated separable convolutions to extract multiscale features with the enlarged receptive field of each kernel for longer-term dependencies modelling. Besides, average pooling is performed for dimensionality reduction and locality preservation. The inclusion of residual connections in MSTCN prevents information leakage throughout the network. The efficiency of MSTCN is evaluated using UCI and WISDM datasets. The empirical results demonstrate the superiority of MSTCN over other state-of-the-art solutions by achieving ~97% and ~96% accuracy scores, respectively, in UCI and WISDM.

Data availability

All data underlying the results are available as part of the article and no additional source data are required.

References

1. Li H, Trocan M: **Deep learning of smartphone sensor data for personal health assistance.** *Microelectronics J.* 2019; **88**(January 2018): 164–172.
Publisher Full Text
2. Yang S, *et al.*: **IoT structured long-term wearable social sensing for mental wellbeing.** *IEEE Internet Things J.* Apr. 2019; **6**(2): 3652–3662.
Publisher Full Text
3. Chen X, Xue H, Kim M, *et al.*: **Detection of Falls with Smartphone Using Machine Learning Technique.** *Proceedings - 2019 8th International Congress on Advanced Applied Informatics, IIAI-AAI 2019.* 2019; pp. 611–616.
4. Wan J, Li M, O'Grady MJ, *et al.*: **Time-Bounded Activity Recognition for Ambient Assisted Living.** *IEEE Trans. Emerg. Top. Comput.* Jan. 2021; **9**(1): 471–483.
Publisher Full Text
5. Anguita D, Ghio A, Oneto L, *et al.*: **A public domain dataset for human activity recognition using smartphones.** *ESANN 2013 proceedings, 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning.* 2013. [Accessed: 17-Sep-2021].
Reference Source
6. Seto S, Zhang W, Zhou Y: **Multivariate time series classification using dynamic time warping template selection for human activity recognition.** *Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015.* 2015; 1399–1406.
7. Kumar A, Gupta S: **Human Activity Recognition through Smartphone's Tri-Axial Accelerometer using Time Domain Wave Analysis and Machine Learning Simulation and Application performance evaluation using GPU through CUDA C & Deep Learning in TensorFlow View project Human Activi.** *Artic. Int. J. Comput. Appl.* 2015; **127**(18): 22–26.
Publisher Full Text
8. Ronao CA, Cho SB: **Human activity recognition with smartphone sensors using deep learning neural networks.** *Expert Syst. Appl.* Oct. 2016; **59**: 235–244.
Publisher Full Text
9. Yazdanbakhsh O, Dick S: **Multivariate Time Series Classification using Dilated Convolutional Neural Network.** *arXiv.* 2019.
10. Yu S, Qin L: **Human activity recognition with smartphone inertial sensors using bidir-LSTM networks.** *Proc. - 2018 3rd Int. Conf. Mech. Control Comput. Eng. ICMCCE 2018.* 2018; pp. 219–224.
11. Ullah M, Ullah H, Khan SD, *et al.*: **Stacked Lstm Network for Human Activity Recognition Using Smartphone Data.** *Proc. - Eur. Work. Vis. Inf. Process. EUVIP.* Oct. 2019; **2019-October**: 175–180.
12. Kwapisz JR, Weiss GM, Moore SA: **Activity recognition using cell phone accelerometers.** *ACM SIGKDD Explor. Newsl.* 2011; **12**(2): 74–82.
Publisher Full Text
13. Kee YJ, Shah Zainudin MN, Idris MI, *et al.*: **Activity recognition on subject independent using machine learning.** *Cybern. Inf. Technol.* Sep. 2020; **20**(3): 64–74.
14. Anjum A, Ilyas MU: **Activity recognition using smartphone sensors.** *2013 IEEE 10th Consumer Communications and Networking Conference, CCNC 2013.* 2013; pp. 914–919.
15. He Z, Jin L: **Activity recognition from acceleration data based on discrete consine transform and SVM.** *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics.* 2009; pp. 5041–5044.
16. Lara ÓD, Prez AJ, Labrador MA, *et al.*: **Centinela: A human activity recognition system based on acceleration and vital sign data.** *Pervasive and Mobile Computing..* 2012; **8**(5): 717–729.
Publisher Full Text
17. Ronao CA, Cho SB: **Human activity recognition using smartphone sensors with two-stage continuous hidden markov models.** *2014 10th International Conference on Natural Computation, ICNC 2014.* 2014; pp. 681–686.
18. Ronao CA, Cho SB: **Recognizing human activities from smartphone sensors using hierarchical continuous hidden Markov models.** *Int. J. Distrib. Sens. Networks.* 2017; **13**(1):

155014771668368.
Publisher Full Text

19. Huang J, Lin S, Wang N, *et al.*: **TSE-CNN: A Two-Stage End-to-End CNN for Human Activity Recognition.** *IEEE J. Biomed. Heal. Informatics.* Jan. 2020; **24**(1): 292–299.
Publisher Full Text

20. Pienaar SW, Malekian R: **Human Activity Recognition using LSTM-RNN Deep Neural Network Architecture.** *2019 IEEE 2nd Wireless Africa Conference, WAC 2019 - Proceedings.* 2019.

21. Nair N, Thomas C, Jayagopi DB: **Human activity recognition using temporal convolutional network.** *ACM Int. Conf. Proceeding Ser.* 2018.

22. Garcia FA, Ranieri CM, Romero RAF: **Temporal approaches for human activity recognition using inertial sensors.** *Proc. - 2019 Lat. Am. Robot. Symp. 2019 Brazilian Symp. Robot. 2019 Work. Robot. Educ. LARS/SBR/WRE 2019.* 2019; pp. 121–125.

23. Szegedy C, *et al.*: **Going deeper with convolutions.** *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2015; **07-12-June**: 1–9.

24. Ismail Fawaz H, *et al.*: **InceptionTime: Finding AlexNet for time series classification.** *Data Min. Knowl. Discov.* Nov. 2020; **34**(6): 1936–1962.
Publisher Full Text

25. Xu C, Chai D, He J, *et al.*: **InnoHAR: A deep neural network for complex human activity recognition.** *IEEE Access.* 2019; **7**: 9893–9902.
Publisher Full Text

26. Li Z, Jiang T, Yu J, *et al.*: **A lightweight mobile temporal convolution network for multi-location human activity recognition based on wi-fi.** *2021 IEEE/CIC Int. Conf. Commun. China, ICCC Work. 2021.* Jul. 2021; pp. 143–148.

27. Lin Y, Wu J: **A Novel Multichannel Dilated Convolution Neural Network for Human Activity Recognition.** *Math. Probl. Eng.* 2020; **2020**: 1–10.
Publisher Full Text

28. Ronald M, Poulose A, Han DS: **ISPLInception: An Inception-ResNet Deep Learning Architecture for Human Activity Recognition.** *IEEE Access.* 2021; **9**: 68985–69001.
Publisher Full Text

29. Kim YJ, Kang BN, Kim D: **Hidden Markov Model Ensemble for Activity Recognition Using Tri-Axis Accelerometer.** *Proceedings - 2015 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015.* 2016; 3036–3041.

30. Kolosnjaji B, Eckert C: **Neural network-based user-independent physical activity recognition for mobile devices.** *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* 2015; vol. **9375 LNCS**: pp. 378–386.

31. Ignatov A: **Real-time human activity recognition from accelerometer data using Convolutional Neural Networks.** *Appl. Soft Comput. J.* 2018; **62**: 915–922.
Publisher Full Text

32. Ogbuabor G, La R: **Human activity recognition for healthcare using smartphones.** *ACM Int. Conf. Proceeding Ser.* 2018; 41–46.

33. Sikder N, Chowdhury MS, Arif ASM, *et al.*: **Human activity recognition using multichannel convolutional neural network.** *2019 5th International Conference on Advances in Electrical Engineering, ICAEE 2019.* 2019; pp. 560–565.

34. Peppas K, Tsolakis AC, Krinidis S, *et al.*: **Real-time physical activity recognition on smart mobile devices using convolutional neural networks.** *Appl. Sci.* 2020; **10**(23): 1–25.
Publisher Full Text

35. Ding C, Jiang B: **L1-norm Error Function Robustness and Outlier Regularization.** *arXiv* May 2017.

36. Lin M, Chen Q, Yan S: **Network in network.** *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings.* 2014.

Open Peer Review

Current Peer Review Status: ✓ ? ✓

Version 2

Reviewer Report 06 March 2023

https://doi.org/10.5256/f1000research.133563.r165304

✓

Xinghua Li

School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China

This work proposed a multiscale temporal convolutional network for human activity recognition. The contribution and innovation are satisfactory. The experiment's result is pleased. This work is meaningful in this field. The authors attempted to polish this manuscript after the first revision.

It can be indexed as the current version in my opinion.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Deep learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 13 June 2022

https://doi.org/10.5256/f1000research.133563.r138230

✓

Cheng-Yaw Low

Institute for Basic Science, Seoul, South Korea

The manuscript has been revised accordingly. Particularly, the experimental section has been included with the experimental setup and the important hyper-parameter configurations for reproducibility.

I have no more comment, and I am happy to recommend this manuscript for indexing.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Deep Learning, Computer Vision, Pattern Recognition, Biometric Recognition.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 02 February 2022

https://doi.org/10.5256/f1000research.76807.r102276

?

Sultan Daud Khan

Department of Computer Science, National University of Technology, Islamabad, Pakistan

In this work, the authors proposed a framework for human activities recognition. The authors proposed a multi-scale temporal convolutional network that constituted multi-scale dilations block to capture multi-scale information. Overall, the paper is not well-written and organized and I have the following concerns the authors need to consider:

1. The contribution of the work is not clear. As there is a lot of literature on human activity recognition systems. How is the proposed framework different from its counterparts?

2. What are the gaps the authors are trying to fill which are left behind by the previous approaches?

3. Figure-1 should be improved, and more details should be incorporated.

4. Discussion section should not be in the bullets form. Please write detail in paragraph.

5. The authors should perform comparison with the following reference:
   - "Stacked lstm network for human activity recognition using smartphone data." In 2019 8th European workshop on visual information processing (EUVIP), pp. 175-180. IEEE, 2019.

References

1. Ullah m, Ullah H, Khan SD, Cheikh FA: Stacked Lstm Network for Human Activity Recognition Using Smartphone Data. *2019 8th European Workshop on Visual Information Processing (EUVIP)*. 2019. 175-180 Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

No

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

No

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computer Vision

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 07 May 2022

Sarmela Raja Sekaran

First of all, we would like to convey our heartfelt thanks to the Editors and Reviewers who have provided us with constructive comments which allowed us to improve our work.

1. (a) The contribution of the work is not clear. As there is a lot of literature on human activity recognition systems.

Response: Thanks for the feedback. The authors have revised the contribution of the work for better clarification.

(b)How is the proposed framework different from its counterparts?

Response: Thanks for the comment. For better clarification, the authors have revised the part of Contribution in the section Introduction to include the difference between the proposed model and the existing methods:

Unlike the existing methods, the proposed method does not require either complex signal pre-processing or manual feature engineering by experts. Besides, the proposed MSTCN is capable of extracting features at multiple scales and concatenating them for a better representation of the overall features. Additionally, the adoption of the dilated convolutions enables the proposed model to preserve longer-term dependencies.

2. What are the gaps the authors are trying to fill which are left behind by the previous approaches?

Response: Authors have revised the Introduction section and included the research gap in HAR.

3. Figure-1 should be improved, and more details should be incorporated.

Response: Authors have revised Figure 1 and included more details for better clarification.

4. Discussion section should not be in the bullets form. Please write detail in paragraph.

Response: Authors have revised the Discussion section. The discussion was written in paragraph form.

5. The authors should perform comparison with the following reference:"Stacked lstm network for human activity recognition using smartphone data." In 2019 8th European workshop on visual information processing (EUVIP), pp. 175-180. IEEE, 2019.

Response: Thanks for the suggestion. Authors have included the comparison with the suggested reference in Related Work section and the Experiments section.

Competing Interests: No competing interests were disclosed.

Reviewer Report 10 January 2022

https://doi.org/10.5256/f1000research.76807.r102273

? **Cheng-Yaw Low**
Institute for Basic Science, Seoul, South Korea

This is a poorly written manuscript, as many sections are unclear and most of the important information is missing.

1. The Inception model is recruited as the network backbone without any justifications. To be specific, why is Inception model useful for HAR?

2. The methodology section is difficult to read as it is incomplete. For example, the input dimension is unknown? and therefore I do not know why 1x1 conv is demanded for dimensionality reduction?

   "First, the input channels are processed via one-by-one causal convolution for dimensionality reduction."

   For clarity, the basic mathematical representation elaborating each operation should be included. On the other hand, I think the methods and results section should be separated into two.

3. Figure captions contain no detail at all, and this makes the reading very difficult. For example, the authors do not provide in Fig. 3 (and the entire manuscript) the definition for d=[1, 2, 4, 5]? What are represented by 8x8, 16x16, 20x20? How was the feature concatenation is performed? By an arithmetic operation? Or stacking over different feature tensors?

4. There are a number of ambiguous or misleading statements throughout the manuscript.

   (a) In MSTCN, GAP replaces the traditional fully connected layers because GAP is more suitable. This operation generates one feature map according to each activity from multi-dimensional feature inputs.

   >> GAP generates one feature map according to each activity?

   (b) First, the input channels are processed via one-by-one causal convolution for dimensionality reduction. This layer, known as bottleneck layer, adopts fewer filters to reduce the number of features maps while the salient features are retained.

   >> The causal convolutional layer is not a bottleneck layer.

   (c) Subject independent protocol is implemented where the training and testing sets do not share the data from the same users.

   >> The training and the testing sets do not share the data from the same users always. I think the authors are claiming that the training and testing identities (instead of data) are disjoint. State also why this training protocol is important in HAR.

   (d) L2 learns complex patterns from the dataset and prevents overfitting.

   >> L2 is only a normalization technique, and L2 does not learn.

   (e) The pooling's output is fed into a one-by-one convolution. A residual connection is formed by passing the input into a one-by-one convolution, followed by a batch normalisation.

   >> There is no residual connection found from both Fig. 2 or 3?

   >> Dilated convolution captures global features? What are these global features? and by how the global features are captured?

5. This manuscript contains only ONE mathematical equation, but it is problematic. In the meantime, the definitions for each variable should also be provided, e.g., what is z_i? what is meant by "simple" softmax classifier? Cross-entropy?

6. The dataset information, the training procedures and the empirical hyperparameters are not disclosed?

(a) What is the input dimension for each dataset? What is the data captured by accelerometer and gyroscope? What are the class number for each dataset?

(b) What is feature dimension rendered by the MSTCN? and from which layer the feature representation is extracted for inference purposes?

(c) The optimizer, learning rate, weight decay, batch size, etc., are unknown.

(d) The measurement unit for segment size and segment interval in Table 1 should be indicated.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Deep Learning, Computer Vision, Pattern Recognition, Biometric Recognition.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 07 May 2022

Sarmela Raja Sekaran

First of all, we would like to convey our heartfelt thanks to the Editors and Reviewers who have provided us with the constructive comments which allowed us to improve our work.

1. The Inception model is recruited as the network backbone without any justifications. To be specific, why is the Inception model useful for HAR?

Response: Thanks for the feedback. The main reason for recruiting the Inception model as the network backbone is because the Inception model allows multiscale feature extraction. The authors have revised the Methods section to include the justification of recruiting the Inception model for better clarification.

2. (a) The methodology section is difficult to read as it is incomplete. For example, the input dimension is unknown? and therefore I do not know why 1x1 Conv is demanded dimensionality reduction?

Response: The Methods section has been revised for better clarification. In the revised version, the authors have included the input dimension in the Model configuration and experimental setup section. The reason for employing 1x1 Conv has also been included in the Method section.

(b) I think the methods and results section should be separated into two.

Response: Authors have separated the Methods and Results sections as suggested by the reviewer.

3. (a) Figure captions contain no detail at all, and this makes the reading very difficult. For example, the authors do not provide in Fig. 3 (and the entire manuscript) the definition for d=[1, 2, 4, 5]? What is represented by 8x8, 16x16, and 20x20?.

Response: Authors have modified Figure 3 and provided a clearer explanation of the figure in the Methods section.

(b) How was the feature concatenation is performed? By an arithmetic operation? Or stacking over different feature tensors?

Response: Feature concatenation is performed by stacking over different feature maps. The Methods section has been revised to include the information for better clarification.

4. There are a number of ambiguous or misleading statements throughout the manuscript
(a) GAP generates one feature map according to each activity?

Response: Authors have revised the explanation in the Methods section for better clarification.

(b) The causal convolutional layer is not a bottleneck layer.

Response: Authors have revised the explanation in the Methods section for better clarification.

(c) The training and the testing sets do not share the data from the same users always. I think the authors are claiming that the training and testing identities (instead of data) are disjoint. State also why this training protocol is important in HAR

Response: For better clarification, the section on Model Configuration and Experimental Setup has been revised. The subject independent protocol is adopted in this work. In other words, there are no overlapping users between training and testing sets. This protocol is preferable for real-time HAR applications.

(d) L2 learns complex patterns from the dataset and prevents overfitting.

Response: Authors have revised the explanation in the Experiments section for better clarification.

(e) There is no residual connection found from both Fig. 2 or 3?

Response: Authors have revised Figure 2 and included more information regarding residual connection in the Methods section.

(f) Dilated convolution captures global features? What are these global features? and by how the global features are captured?

Response: Authors have revised the Methods section for better clarification. Confusing sentences have been revised. Dilated convolution enables longer-term time dependency of the proposed model by enlarging the convolution's receptive fields.

5. This manuscript contains only ONE mathematical equation, but it is problematic. In the meantime, the definitions for each variable should also be provided, e.g., what is z_i? what is meant by "simple" softmax classifier? Cross-entropy?

Response: Authors have revised and included equations (softmax activation and cross-entropy functions) as well as the definition for each variable of each equation in the Methods section for better clarification. Softmax classifier is implemented in the proposed model for classification purposes. Further, the categorical cross-entropy loss has been implemented to the softmax function output.

6. The dataset information, the training procedures and the empirical hyperparameters are not disclosed??
(a) What is the input dimension for each dataset? What is the data captured by the accelerometer and gyroscope? What is the class number for each dataset?

Response: Authors have included detailed information about the datasets including input dimension, class number and data type in the Experiments and Results section.

(b) What is the feature dimension rendered by the MSTCN? and from which layer the feature representation is extracted for inference purposes?

Response: Softmax classifier is implemented for inference purposes. The authors have revised the Methods section for better clarification.

(c) The optimizer, learning rate, weight decay, batch size, etc., are unknown

Response: Authors have included information regarding optimizer, learning rate, batch size, etc., in the Model configuration and experimental setup section for better clarification.

(d) The measurement unit for segment size and segment interval in Table 1 should be indicated.

Response: Authors have provided the measurement unit for segment size and segment interval in the Model configuration and experimental setup section.

Competing Interests: No competing interests were disclosed.