Journal of Engineering Science and Technology
Vol. 17, No. 5 (2022) 3235 - 3253

STACKING SPATIAL-TEMPORAL DEEP LEARNING ON INERTIAL DATA FOR HUMAN ACTIVITY RECOGNITION

CHEW YONG SHAN[1], PANG YING HAN[1,*],
OOI SHIH YIN[1], POH QUAN WEI[2]

[1]Faculty of Information Science and Technology, Multimedia University,
Jalan Ayer Keroh Lama, 75450 Ayer Keroh, Melaka Malaysia
[2]Winnefy Enterprise, Taman Sri Duyong 2, 75460 Melaka, Malaysia
*Corresponding Author: yhpang@mmu.edu.my

Abstract

Insufficient physical activity has negative effects on quality of life and mental health. Further, physical inactivity is one of the top ten risk factors for mortality. Regular recognition and self-monitoring of physical activity are in the hope to encourage users to stay active. One such application is through intelligent human activity recognition which is usually embedded in ambient assisted living systems. A spatial-temporal deep learning is proposed in this paper for smartphone-based intelligent human physical activity recognition. In this work, a stacking spatial-temporal deep model is devised to extract deep spatial and temporal features of inertial data. In the proposed system, a convolutional architecture is pipelined with Bidirectional Long Short Term Memory to encapsulate the spatial and temporal state dependencies of the motion data. Support Vector Machine is adopted as the classifier to distinguish human activities. Empirical results demonstrate that the proposed system exhibits promising performances on two public datasets (UC Irvine dataset and Wireless Sensor Data Mining database) with 92% and 87% accuracy, respectively.

Keywords: Bidirectional long short term memory, Convolutional neural network, Human activity recognition, Inertial data, Smartphone.

1. Introduction

Surveys data promulgated that the global prevalence of insufficient physical activity was 27.5% in 2016 [1]. The statistics showed that prevalence in 2016 was more than twofold as high in high-income countries as in low-income countries. Furthermore, insufficient activity had increased in high-income countries over time from 31.6% in 2001 to 36.8% in 2006. Insufficient physical activity has negative effects on quality of life and mental health. Further, physical inactivity is one of the top ten risk factors for mortality. Overwhelming testimonies substantiate that the insufficiency of physical activity may be a contributing factor to chronic diseases such as ischemic heart disease, high blood pressure, diabetes, stroke, hypertension, depression, and cancers [2, 3]. The upsurge of chronic diseases will hamper social and economic development, i.e., unemployment, household financial burden, poverty, etc. [4]. Hence, communities must proactively advocate increasing individuals' physical activity.

Regular recognition and self-monitoring of physical activity can potentially cultivate the habits of adopting a healthy lifestyle, i.e., one may do exercises regularly in order to have a better body shape [5, 6]. The aforementioned application can be widely found in various ambient assisted living systems. Generally, computational intelligence, including machine learning, artificial intelligence etc., can be made use to enhance ambient assisted living systems. It is a significant science application in the context of IoT environments as well as ambient assisted living systems. The invented science and technology can be advocated as a technological solution to improve human living. Utilizing computational intelligence in monitoring human activity is one of the solutions for smart life.

Motivation and contribution

An intelligent human activity recognition system is one of the innovations adopting computational intelligence to improve human living. There are three kinds of activity recognition systems: vision-based, wearable sensor-based, and smartphone-based [7-10]. Vision-based and wearable sensor-based methods are the two most common approaches and excelled in human activity recognition. However, the usage of them is inconvenient and not easy to implement. For instance, there is a privacy issue evolving in the vision-based approach. Placing a surveillance camera in public places may violate the land law and require extensive justifications to obtain permit. Whereas, in wearable sensor-based approaches, it is not easy and almost irrational to force every user to wear the sensor device(s).

Henceforth, physical activity recognition using a smartphone is a contemporary research in the human activity recognition domain. Smartphone is a sensor-based ubiquitous piece of technology that is far more than just a communication device. With the great technology development, smartphones are packed with high-end hardware and features. Several sensors are embedded in smartphones, including accelerometer, gyroscope sensor, ambient light sensor, fingerprint sensor etc. The potential of smartphone-based human activity recognition is uplifted due to the mobility and simplicity of smartphone usage.

In literature, spatial and temporal features are acclaimed as effective parameters in characterizing human motion. These features are exploited for better human activity recognition. However, there is room for improvement in the recognition

performance, especially for the subject-independent solution. This subject-independent protocol is challenging since the motion patterns in the human gait of an individual are different due to external factors such as walking surface, clothing, carrying conditions, etc. But the subject-independent solution is always preferable and practical for real-world applications. In viewing this, a subject-independent solution of a smartphone-based human activity recognition system is presented. In this work, a stacking spatial-temporal deep model for human activity recognition is proposed. To be specific, an amalgamation of one-dimensional Convolutional Neural Network (CNN) and Bidirectional Long Short Term Memory (BLSTM) is proposed to classify human activities based on the inertial data captured by a smartphone, as illustrated in Fig. 1.

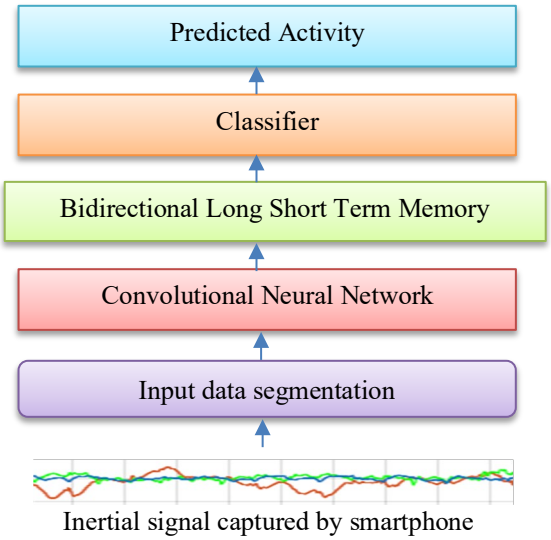


Fig. 1. Overview of the proposed system.

Upholding the hypotheses of (1) the spatial and temporal information embedded in the inertial signal is crucial to represent activity, and (2) there is a profound high-level knowledge about human activity, we propose a dynamics deep learner to extract deep features from spatial and temporal domains. The main contributions of this work are twofold:

1) A stacking spatial-temporal deep model is developed to extract deep spatial and temporal features of inertial data for human activity recognition. Piling a convolutional structure to BLSTM enables the encapsulation of both spatial and temporal state dependencies for human activity recognition.

2) An extensive experimental analysis using various performance measures, such as true positive rate, false positive rate, precision, recall, the area under the curve, confusion matrix, etc., is conducted on two publicly available datasets, namely WISDM and UCI HAR.

2. The Related Work

As aforementioned, human activity recognition can be categorized into three spheres: (1) vision-based, (2) wearable sensor-based and (3) smartphone-based.

Vision-based approach is a process of categorizing a sequence of image recording with activity class labels [7]. These systems are extensively employed in various applications, especially for public area surveillance, healthcare monitoring and human-computer interaction. The proposed approaches include but are not limited to Discrete Fourier transform-based system to extract the global representation of activity data [11], stacked Fisher vectors to capture more statistical information from frame images [12], extraction of multi-features from body silhouettes and joints information [13], etc. However, the privacy issue is a major concern in the vision-based approach. Hence, the wearable sensor-based approach comes in as an alternative. Wearable sensors in this context are referring to any devices which may include the accelerometer, gyroscope, and magnetometer. Some prior arts include the usage of artificial neural network and smartwatch, CNN for *k*-nearest neighbourhood-based wearable sensor [14], adopting deep learning and Extreme Learning Machine for activity recognition based on wearable sensors [15], etc. Inconveniences of wearing, technological barriers (i.e., the battery lifetime), and culture barriers (i.e., the association of a stigma with the use of medical sensing devices for monitoring) limit the potential of wearable-based approach usage.

Owing to the fact that most smartphones are equipped with a built-in gyroscope and accelerometer, it becomes a seemly alternative for collecting motion inertial signals. In recent years, there are extensive research works on smartphone-based activity recognition [16-20]. For instance, Kwapisz et al. [8] utilized triaxial accelerometer data to recognize human activities. In this work, accelerometer data from six activities were collected, forming a database termed as Wireless Sensor Data Mining (WISDM) dataset. The collected data was preprocessed into segments with a duration of 10 seconds. Further, forty-three statistical features were computed for each segment. WISDM dataset was tested with various machine learning classifiers, including J48 Binary Tree, Logistic Regression, Multilayer Perceptron and Straw Man. An encouraging performance was reported.

Anguita et al. [17] constructed another database, UC Irvine (UCI) Human Activity Recognition (HAR) dataset from a group of volunteers with a smartphone on their waist. UCI HAR dataset contains both acceleration data and angular velocity data. These signals were pre-processed with median and low-pass filters for noise removal. Then, they were sampled in fixed-width sliding windows of 2.56 seconds with a 50% overlap between them. To obtain rich features, the signal was processed for 561 feature variables. Support Vector Machine was used to classify the activities.

In recent years, deep learning is a dominant research in human activity recognition. For instances, the works of [21-25] presented the potential of deep learning approaches in exploiting the characteristics of time series motion data for human activity recognition. Among the literature, the CNN-based method is widely exploited. CNN was applied in mobile sensor-based activity recognition to extract local dependency and scale invariant characteristics of the acceleration signal [26]. The authors incorporated a more relaxed weight-sharing strategy (partial weight sharing) in CNN to enhance its performance. The results exhibited the superiority of the proposed CNN-based approach.

Besides that, a temporal deep learner, called Long Short Term Memory (LSTM), was employed on triaxial accelerometer data for human activity prediction [27]. This model adopts past information to predict the outcome of activity recognition. It allows the network to learn when to "forget" the previous

hidden states and when to update the hidden states whenever new information is discovered. Owing to the superiority of LSTM in analysing human behaviour, LSTM had been enhanced by integrating assorted LSTM learners into a collective classifier [28]. The empirical results substantiated the superior performance of the proposed integrated system of LSTM learners.

However, some information may not be captured completely through LSTM since human motion data is continuously oriented. Hence, BLSTM was proposed for smartphone-based human activity recognition [29]. This sequential network tackles both past and future information, extracting a richer feature set. BLSTM cells are stacked into layers. For each layer, the cell takes in the information from the horizontal direction (both past and future information), as well as the information from the vertical direction from the lower layer. Hernández et al. also utilized BLSTM for activity recognition [30]. There are not many differences in terms of algorithm implementation between these works. Both works demonstrate the feasibility of BLSTM for activity recognition.

Furthermore, Ignatov presented a user-independent deep learning-based method with global and local features for real-time human activity recognition [20]. In this method, global statistical features are manually computed, whereas the deep local features are automatically extracted by the CNN network. Then, these both features are concatenated and fed into Softmax for classification. The performance of the proposed method was assessed based on open data sets and the reported findings showed that the model had the advantages of good classification performance and light computation.

3. Proposed System

A spatial-temporal deep learner that comprehends both spatial and dynamic patterns of the motion data is proposed. Inertial data acquired from the smartphone is transformed into deep features via the feature extraction structure of CNN and BLSTM. These deep features are further analysed to recognize human activity by using a machine learning classifier. Specifically, in the proposed architecture, the feature extraction structure comprises three convolutional layers with rectified linear unit (RELU) activation function, one max-pooling layer, one flattening layer, and one BLSTM layer, as illustrated in Fig. 2. The neural processors of the lower layers attain local features of the inertial signal to signify the elementary motion in the human activity; whilst higher layer neural processors extract a better abstraction of the motion with deep features and temporal analysis.

To better preserve the temporal attribute for BLSTM analysis, each data sequence is sub-segmented into sub-windows and each sub-window is fed into feature extraction structure in the chronological order they are in, see Fig. 2. Each sub-sequence (appearing in the sub-window) has its own feature extraction flow. In each flow, convolutional layers analyse each sub-sequence using a kernel/neuron that reads in a small block at a time and strides across the entire data, see Fig. 3. Each read results the input being projected onto a feature map, representing the internal interpretation of the input. Since each convolutional layer contains multiple kernels, multiple feature maps will be constructed after every layer and then the maps are concatenated. To avoid a very long training time for each convolution flow for different weights, the same weights are shared in each layer. Through the convolution operations, the spatial local dependencies in the inertial data are

apprehended. The correlation between nearby signal points is pictured, revealing the structure of the signal pattern. The max-pooling layer is to downsample each feature map independently for a summarized version. Besides, the max-pooling layer helps with the model's invariance of the local translations. With this property, a slight translational variance of the data will not affect the values of the pooled output.

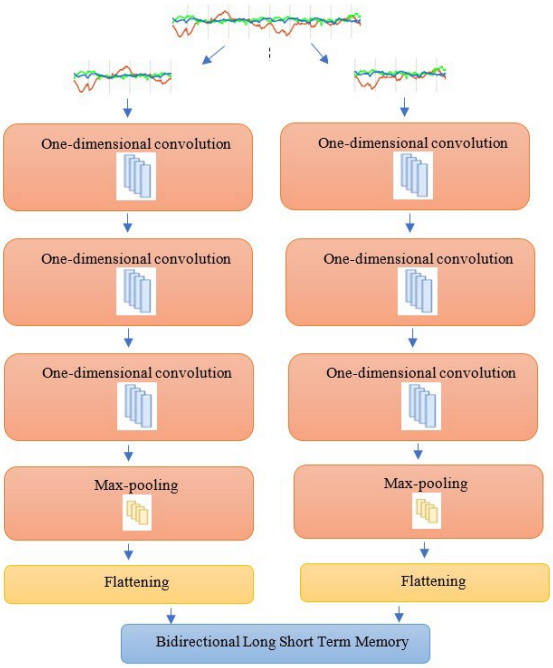


Fig. 2. The proposed architecture.

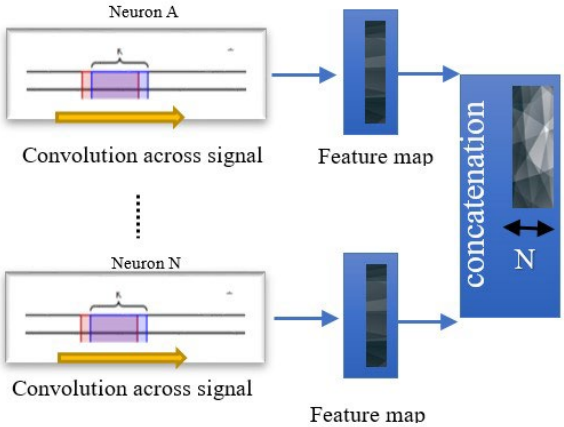


Fig. 3. One-dimensional convolution on sub-sequence signal.

Since convolutional layers unearth the underlying patterns of the signal, this proposed model is able to encapsulate the tiny changes in the motion signal. The

changes in sequential form are substantial to characterize activity motion. Hence, BLSTM is included to capture the dynamics features by analysing the underlying sequential pattern in the spatial-temporal feature map. BLSTM has a better prediction compared with LSTM since BLSTM is using both past and future information [31]. BLSTM is a stack of two LSTM units on top of each other, illustrated in Fig. 4. One unit moves in the forward direction, while the another moves oppositely. LSTM's outputs are fused and computed as BLSTM's output. Then, the deep features are extracted and fed into a classifier for recognition.

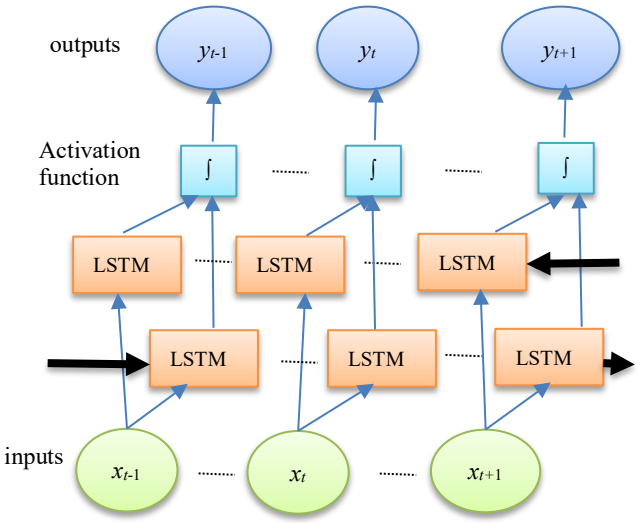


Fig. 4. BLSTM with both forward → and backward ← layers

Formulation

In this work, both CNN and BLSTM are implemented. The process of the convolutional layer is defined as below

$$C_n = A((B + W_{n-1}V) \quad (1)$$

where C_n denotes convolutional output, called feature map, at nth layer, A is the activation function, i.e., Rectified Linear Unit in this case, B refers to the bias term, W_{n-1} is the weight from the previous layer, V is the input vector of the inertial signal. The pooling layer performs downsampling to the generated feature maps by summarizing the presence of features in patches of the feature map. Maximum pooling is used in this work

$$P = MAX((C^i{}_n) \quad (2)$$

where P is the output of the pooling operation, and $C^i{}_n$ is the ith patch of a feature map from nth convolutional layer. Next, P is flattened and fed for temporal dynamics analysis via LSTM nodes. In the LSTM node, there are memory cells and four gates (i.e., forget gate, input gate, input modulation gate, and output gate). Forget gate at any given timestep t is formulated as below

$$F_{n_t} = \sigma\left(W_f[L_{n_{t-1}}, V_{n_t}] + B_f\right) \quad (3)$$

where F_{n_t} is the forget gate output at n layer at timestep t and σ is the sigmoid function, W_f denotes the weight of the connection at forget gate, $L_{n_{t-1}}$ is the LSTM output from the previous layer, $\mathrm{V_{n_t}}$ is the input vector, and B_f is the bias term for the forget gate. The next gate is the input gate

$$I_{n_t} = \sigma\left(W_i[L_{n_{t-1}}, V_{n_t}] + B_i\right) \quad (4)$$

where $\mathrm{I_{n_t}}$ is the input gate output at n layer at timestep t and σ is the sigmoid function, W_i denotes the weight of the connection at the input gate, $L_{n_{t-1}}$ is the LSTM output from the previous layer, V_{n_t} is the input vector, and B_i is the bias term for the input gate. Similar to the input gate and forget gate, the output gate exhibits similar formulation

$$O_{n_t} = \sigma\left(W_o[L_{n_{t-1}}, V_{n_t}] + B_o\right) \quad (5)$$

O_{n_t} denotes the output gate output at n layer at timestep t and σ is the sigmoid function, W_o denotes the weight of the connection at the output gate, $\mathrm{L_{n_{t-1}}}$ is the LSTM output from the previous layer, V_{n_t} is the input vector, and B_o is the bias term for the output gate. The input modulation gate is a function of the input vector and the previous state output

$$G_{n_t} = tanh\left(W_s[L_{n_{t-1}}, V_{n_t}] + B_s\right) \quad (6)$$

where G_{n_t} denotes the input modulation gate output at n layer at timestep t with $tanh$ function, W_s denotes the weight of the connection at the input modulation gate which is based on the previous state, $L_{n_{t-1}}$ is the LSTM output from the previous layer, V_{n_t} is the input vector, and B_s is the bias term for the input modulation gate based on the previous state gate. The state gate or memory cell consists of two terms: the previous memory cell state $S_{n_{t-1}}$ which is modulated by forget gate F_{n_t}, and input modulation gate G_{n_t} which is modulated by input gate I_{n_t} at timestep t

$$S_{n_t} = F_{n_t} S_{n_{t-1}} + I_{n_t} G_{n_t} \quad (7)$$

The LSTM outputs the combination of the output gate and state gate, with the following equation

$$L_{n_t} = O_{n_t} * tanh\left(S_{n_t}\right) \quad (8)$$

where O_{n_t} is the output gate's output at n layer with $tanh$ function and S_{n_t} is the state gate's output at n layer. Each LSTM node generates two values for the next LSTM nodes. The subsequent node will then use both information accordingly to update their states and eventually the entire network. This allows LSTM network takes the ability to consider past information. BLSTM is alike stacking two LSTM units on top of each other. In other words, BLSTM has both forward sequences $\vec{L}$ and backward sequences $\overleftarrow{L}$ in the hidden layer. At time t, the hidden layer and the input layer can be defined as follows

$$\vec{L} = A((W_{\vec{L}}\, \vec{L}_{n-1} + B_{\vec{L}}) \quad (9)$$

$$\overleftarrow{L} = A((W_{\overleftarrow{L}}\,\overleftarrow{L}_{n-1} + B_{\overleftarrow{L}}) \tag{10}$$

$$\overleftrightarrow{L} = \mathrm{A}(\vec{L} + \overleftarrow{L} + B_{\overleftarrow{L}}) \tag{11}$$

where $\vec{L}$ denotes the forward sequence, $\overleftarrow{L}$ denotes the backward sequence of LSTM operation, $\overleftrightarrow{L}$ refers to the BLSTM output, A is the activation function used in the network, W is the weight of the connection and B is the bias term. These forward and backward sequences allow LSTM nodes to take in previous and subsequent information to update its state, this will subsequently update the state of the whole network. This gives BLSTM's properties to use past and future information to effectively project a deeper representation of a set of data inputs.

4. Experiments and Discussions

The effectiveness of the proposed architecture is evaluated by using two public databases collected using inertial sensors embedded in smartphones. These databases are: (1) WISDM and (2) UCI HAR. In this work, Support Vector Machine (SVM) is adopted as the machine learning algorithm to classify the extracted deep features for human activity classification. Table 1 shows the details of these datasets.

Table 1. Database details.

Remark	UCI	WISDM
Activities	Walking, Walking upstairs, Walking downstairs, Sitting, Standing, Laying	Walking, Jogging, Stairs-Up, Stairs-Down, Sitting, Standing
Location of smartphone	Waist	Front leg pocket
Data used in this work	Triaxial gravity acceleration data Triaxial body acceleration Triaxial angular velocity	Triaxial gravity acceleration data
Time steps	2.56 seconds (128 timesteps)	10 seconds (200 timesteps)

4.1. Experimental setup and protocol

In this work, we explore user-independent activity recognition. The training-testing split is implemented based on individuals where samples of a panel of users are used for training, whilst samples of the other group of users are used to test the system performance. There are no overlapping users between training and testing subsets. This user-independent protocol is quite challenging because the motion patterns in human gait are different among people. However, it is more practical and fit for real-world application; the model can be directly applied to any new users without a need for model retraining and regeneration. In UCI HAR dataset, 21 users are selected as training users and the remaining 9 users are as testing users. On the other hand, in WISDM dataset, samples of the first 28 users are used for model training, whereas samples of the remaining users are used for system performance testing.

Numerous hyperparameters are tested in the proposed stacked spatio-temporal deep learner with softmax function used in the final layer. In the validation process, various numbers of kernels in the convolutional layers are examined from 32 kernels to 256 kernels; different hidden nodes in BLSTM layer are studied from 32 units to 400 units. Through extensive trials, the optimal parameters are determined to build an efficient

model to extract the desired deep features. In UCI HAR dataset, there are 64 kernels in each convolutional layer and 256 nodes in the BLSTM layer; whereas in WISDM dataset, 100 kernels and 240 nodes are employed. In this model, we adopt Adam optimization algorithm - an extension to stochastic gradient descent, to update network weights for cost minimization. The reasons of choosing Adam are it is computationally efficient, and it is well-suited for problems with large data. A small learning rate is good for reliable training, but it requires many updates towards the minimum of the loss function.

In contrast, a high learning rate may cause drastic updates which lead to divergent behaviour. Hence, the proposed model is trained with a learning rate of 0.001 which swiftly reaches the minimum point. Weight decay of 0.0001 is adopted for regularization to minimize the overfitting issue. Batch size is always limited by the computing hardware's memory. Due to the constraint of the hardware, a batch size of 32 is used in UCI HAR and 80 in WISDM with 30 epochs. Next, the extracted deep features are fed into machine learning classifiers for system performance testing.

Classification accuracy is usually used to measure the overall accuracy of a classification model. However, it is not substantial to decide whether the model is well enough to make robust predictions, i.e., it is not able to discriminate kinds of misclassifications. So, performance metrics such as precision, recall, F1-score, and the Area under the Receiver Operating Characteristic Curve (AUC) are topped up for performance measurement. Precision is the number of positive predictions divided by the total number of positive class values predicted.

A low precision indicates a large number of False Positive. Recall is a measure of a model's completeness/ accuracy. F1 score/ measure is a harmonic mean of precision and recall, conveying the balance between the two measures. The value of AUC represents the robustness of a classification model. A higher value of AUC indicates the probability of a classifier to rank a randomly chosen positive instance higher than a randomly chosen negative instance. In this work, the empirical results are testified based on performance evaluation metrics of (i) true positive (TP) rate, (ii) false positive (FP) rate, (iii) precision, (iv) recall (v) AUC, (vi) F1 score and (vii) classification accuracy.

4.2. Evaluation and discussion

Temporal feature is one of the key features in motion signals. In order to unfold the temporal feature, BLSTM is employed in the proposed architecture. The efficacy of our architecture in analysing the underlying temporal features is investigated by comparing the performances of the full architecture and the architecture without BLSTM structure. Table 2 records the performances of the architectures. It is observed that the inclusion of BLSTM is able to improve performance, reducing the false positives and negatives. The temporal state dependency of motion signal captured by BLSTM is beneficial and prominent in recognizing human activity.

Table 2. Performance of full architecture and architecture without BLSTM.

	Accuracy	Precision	Recall	F1 score
UCI HAR database				
Full architecture	91.9919	0.922	0.920	0.920
Architecture without BLSTM	90.6685	0.908	0.907	0.907
WISDM database				
Full architecture	88.5176	0.894	0.885	0.888
Architecture without BLSTM	87.9557	0.890	0.880	0.883

4.2.1. Performance analysis with Support Vector Machines

In this work, Support Vector Machine (SVM) is employed to classify human activities based on the extracted features. The adoption of SVM is due to its characteristic of flexibility which could solve a variety of problems with minimal tuning. Moreover, SVM implements automatic complexity control to overcome overfitting issues. Kernel trick is a key component in SVM, and it bridges linearity and non-linearity. Real-world data is randomly distributed, and it is hard to separate linearly. Utilizing a kernel function, these data samples can be projected onto a nonlinear embedding subspace where the projected samples are easily separable. The influence of kernel tricks of SVM is examined in this section. Specifically, the extracted deep features from the proposed architecture are classified using SVMs with different kernels. Tables 3 and 4 record the performances of different SVM's kernels for UCI HAR and WISDM, respectively. Figures 5 and 6 show the confusion matrices of SVMs for UCI HAR and WISDM, respectively.

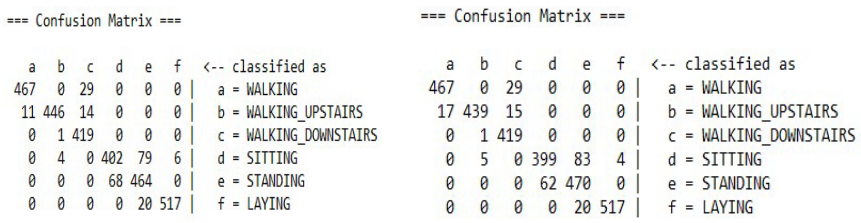
From the results, we can observe that the proposed classification model with the appropriate kernel exhibits good performance in UCI HAR and WISDM, i.e., approximately 92% and 88% accuracy, respectively. The adoption of SVM enables a decision hyperplane creation with the maximum margin. The maximization of the margin distance of inter-class data provides a certain degree of reinforcement so that future unknown data can be classified with more confidence. This feature allows a good generalization performance. The proposed classification model shows slight inferiority in classifying sitting and standing actions in UCI HAR database, see Table 3. Lower precision and recall are observed in these activities, denoting more false positives and negatives in predicting these activities. This may be due to the proposed model is not competent enough to distinguish sitting and standing inertial data of UCI HAR, as illustrated in the confusion matrix in Fig. 5.

Table 3. Performances of support vector machines in UCI HAR.

SVM's Kernel	Class	TP rate	FP rate	Precision	Recall	F1 Score	AUC
Linear	Walking	0.942	0.004	0.977	0.942	0.959	0.969
	Upstairs	0.947	0.002	0.989	0.947	0.967	0.972
	Downstairs	0.998	0.017	0.907	0.998	0.950	0.990
	Sitting	0.819	0.028	0.855	0.819	0.837	0.896
	Standing	0.872	0.041	0.824	0.872	0.847	0.916
	Laying	0.963	0.002	0.989	0.963	0.975	0.980
	Average	**0.921**	**0.016**	**0.923**	**0.921**	**0.922**	**0.953**
	Accuracy			**92.1276**			
Polynomial	Walking	0.942	0.007	0.965	0.942	0.953	0.994
	Upstairs	0.932	0.002	0.987	0.932	0.959	0.997
	Downstairs	0.998	0.017	0.905	0.998	0.949	0.991
	Sitting	0.813	0.025	0.866	0.813	0.838	0.944
	Standing	0.883	0.043	0.820	0.883	0.851	0.956
	Laying	0.963	0.002	0.992	0.963	0.977	0.982
	Average	**0.820**	**0.016**	**0.922**	**0.920**	**0.920**	**0.977**
	Accuracy			**91.9919**			
Radial	Walking	0.942	0.003	0.983	0.942	0.962	0.969
	Upstairs	0.943	0.003	0.984	0.943	0.963	0.970
	Downstairs	0.998	0.020	0.893	0.998	0.943	0.989
	Sitting	0.845	0.031	0.845	0.845	0.845	0.907
	Standing	0.852	0.037	0.836	0.852	0.844	0.907
	Laying	0.957	0.002	0.990	0.957	0.973	0.978
	Average	**0.920**	**0.016**	**0.922**	**0.920**	**0.921**	**0.952**
	Accuracy			**92.0258**			

Table 4. Performances of support vector machines in WISDM.

SVM's Kernel	Class	TP rate	FP rate	Precision	Recall	F1 Score	AUC
Linear	Downstairs	0.540	0.028	0.678	0.540	0.601	0.756
	Jogging	0.955	0.001	0.997	0.955	0.976	0.977
	Sitting	0.996	0.011	0.870	0.996	0.929	0.992
	Standing	0.808	0.000	0.990	0.808	0.890	0.904
	Upstairs	0.687	0.046	0.649	0.687	0.668	0.820
	Walking	0.928	0.083	0.864	0.928	0.895	0.922
	Average	0.869	0.039	0.870	0.869	0.867	0.915
	Accuracy			**86.938**			
Polynomial	Downstairs	0.732	0.051	0.611	0.732	0.666	0.940
	Jogging	0.944	0.004	0.990	0.944	0.966	0.986
	Sitting	0.820	0.008	0.891	0.920	0.905	0.991
	Standing	0.854	0	0.997	0.854	0.920	0.995
	Upstairs	0.803	0.030	0.767	0.803	0.784	0.898
	Walking	0.901	0.049	0.913	0.901	0.907	0.952
	Average	0.885	0.028	0.894	0.885	0.888	0.960
	Accuracy			**88.5176**			
Radial	Downstairs	0.542	0.019	0.754	0.542	0.630	0.761
	Jogging	0.966	0.003	0.993	0.966	0.979	0.981
	Sitting	0.966	0.010	0.882	0.966	0.936	0.993
	Standing	0.819	0.001	0.984	0.819	0.894	0.909
	Upstairs	0.690	0.048	0.642	0.690	0.665	0.821
	Walking	0.938	0.080	0.870	0.938	0.903	0.929
	Average	0.877	0.038	0.878	0.877	0.875	0.920
	Accuracy			**87.7278**			



```
=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 467   0  29   0   0   0 |   a = WALKING
  11 446  14   0   0   0 |   b = WALKING_UPSTAIRS
   0   1 419   0   0   0 |   c = WALKING_DOWNSTAIRS
   0   4   0 402  79   6 |   d = SITTING
   0   0   0  68 464   0 |   e = STANDING
   0   0   0   0  20 517 |   f = LAYING
```

(a)

```
=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 467   0  29   0   0   0 |   a = WALKING
  17 439  15   0   0   0 |   b = WALKING_UPSTAIRS
   0   1 419   0   0   0 |   c = WALKING_DOWNSTAIRS
   0   5   0 399  83   4 |   d = SITTING
   0   0   0  62 470   0 |   e = STANDING
   0   0   0   0  20 517 |   f = LAYING
```

(b)

```
=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 467   0  29   0   0   0 |   a = WALKING
   6 444  21   0   0   0 |   b = WALKING_UPSTAIRS
   0   1 419   0   0   0 |   c = WALKING_DOWNSTAIRS
   0   5   0 415  66   5 |   d = SITTING
   2   1   0  76 453   0 |   e = STANDING
   0   0   0   0  23 514 |   f = LAYING
```

(c)

Fig. 5. Confusion matrices of support vector machine with the kernel (a) linear, (b) polynomial and (c) radial for UCI HAR.

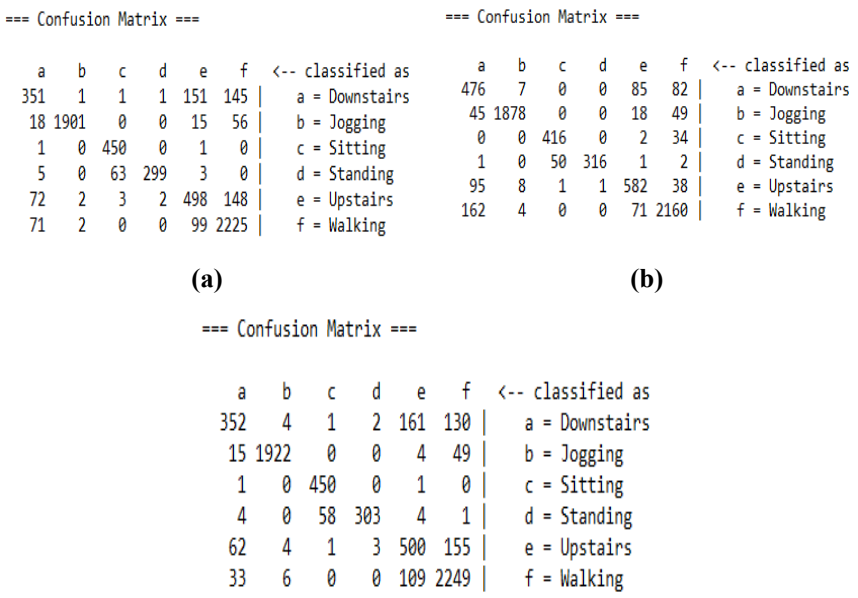


Fig. 6. Confusion matrices of support vector machine with the kernel (a) linear, (b) polynomial and (c) radial for WISDM.

On the other hand, the proposed classification model suffers inferiority when dealing with inertial data of downstairs and upstairs activities in WISDM database. The model has a high possibility of falsely predicting a downstairs activity as upstairs and walking activities, leading to low recall in the downstairs class. The high false negatives are also observed in the confusion matrix in Fig. 6. Besides, it is also noticed that a low precision is obtained in the upstairs activity. The model falsely predicts another activity as the upstairs class, causing more false positives. The proposed model confuses downstairs and walking data as the upstairs activity. The position of the smartphone could be one of the reasons for the performance discrepancy between these two databases. In UCI HAR, the smartphone is placed at the body waist; whereas in WISDM, the smartphone is placed in the front leg pocket. Undeniably, the position of the smartphone is one of the factors that could greatly influence the quality of data and subsequently affect the accuracy of the classification model [32]. Overall, the F1 score is pretty good for the proposed model, with score of 0.922, and 0.888 for UCI HAR and WISDM databases, respectively.

4.2.2. Computational time

In this section, we examine the computational efficiency of the proposed system. The experiments were conducted in the environment of an AMD Ryzen 7 4800H with Radeon Graphics 2.90 GHz processor laptop with 16.0 GB running on Windows 10 operating system. Both the training and testing time of the deep learner, as well as the SVM classifier, are computed in Table 5. The experiments were implemented based on the UCI HAR dataset with 7352 training samples and 2947 testing samples. In the proposed deep learner, the data captured from the inertial sensors are fed into

convolutional layers and then the LSTM layer for feature extraction. During the LSTM layer training, the calculation of each time step depends on the output of the previous time step. Therefore, the process cannot be computed in parallel, which causes the training of the deep learner to take longer time. However, system training is usually conducted in offline mode, so it is still acceptable to spend 5 to 6 minutes training the system for model upgrading purposes. The testing time shown in the table is for the whole testing set, comprising 2947 samples. On average, it just needs merely 4.123 milliseconds to test and classify one sample.

Table 5. Computational time of the proposed system on UCI HAR dataset.

Process	Computational time (s)
A: Training of deep learner (build and train the deep learner)	282.39
B: Testing of deep learner (extract deep features from the model)	4.80
C: Training of classifier (build and train the SVM classifier)	53.4
D: Testing of classifier (classify the deep features)	7.35
Total training time (A+C)	335.79
Total testing time (B+D)	12.15

4.2.3. Performance comparison with other approaches

In this section, performance comparisons between the proposed model and state-of-the-art approaches are addressed. Classification accuracies of the approaches are summarized in Tables 6 and 7. From the empirical results, we can observe that the proposed model generally exhibits superior or comparable performance with the other state-of-the-art approaches, including artificial neural networks [9, 21, 33], recurrent neural networks [27, 29, 30] and Autoencoder [34], as well as Impersonal Smartphone-based Activity Recognition model [35]. This observation substantiates that the deep spatial and temporal features of inertial data, depicted in the proposed model, are able to signify the traits of human activities. Besides that, the discriminant decision hyperplane of SVM facilitates higher confidence to classify unseen data. This is significant especially in the subject-independent solution where the training and testing data are from different individuals.

From the results, we can notice that Hierarchical Multi-View Aggregation Network [36] exhibits slightly higher accuracy in both UCI HAR and WISDM, with about 95% accuracy rate. The model is an aggregation network that integrates black-box features with white-box features in a hierarchical multi-view structure. Those features are then aggregated into a unified representation. Our proposed method is more simplified with concatenating spatio-temporal deep features in layers, whereas Hierarchical Multi-View Aggregation Network is multifaceted and involves more steps in handling multi-view features. In the aggregation network, a non-local operation is applied to aggregate the feature level features. Then, the correlation of sensor positions is taken into consideration for position level aggregation. Lastly, a soft attention mechanism is conducted for modality level aggregation.

Table 6. Performances of various approaches in UCI HAR.

Algorithms/ Architecture	Accuracy (%)
Hierarchical Multi-View Aggregation Network (2-layer)** [36]	95.5
Simplified Hierarchical Multi-View Aggregation Network (1-layer)** [36]	94.7
Hierarchical Continuous Hidden Markov Model** [37]	93.18
Dynamic Time Warping** [38]	89.00
CNN* [21]	89.41
CNN[33]	90.89
CNN* [9]	89.24
LSTM * [27]	89.79
BLSTM* [29]	89.07
BLSTM* [30]	87.41
Autoencoder + Random Forest** [34]	77.81
Proposed model + SVM*	92.13

* Denotes models are trained using 30 epochs based on the original architecture
** denotes the empirical results reported in the respective papers

Table 7. Performances of various Approaches in WISDM.

Model	Accuracy (%)
Impersonal Smartphone-based Activity Recognition** [35]	75.22
CNN* [21]	74.79
CNN* [9]	85.98
LSTM* [27]	79.62
BLSTM* [29]	72.98
BLSTM* [30]	82.09
Autoencoder + Dropout** [34]	83.45
Proposed model + SVM*	88.52

* Denotes models are trained using 30 epochs based on the original architecture
** denotes the empirical results reported in the respective papers

5. Conclusion

In this paper, a spatial-temporal deep feature extraction architecture is proposed by stacking one-dimensional convolutional layers with a BLSTM model. This stacking deep model is capable of extracting deep features of inertial data for human activity recognition while encapsulating both spatial and temporal state dependencies of inertial data. The employment of Support Vector Machine enables good classification accuracy in both UCI HAR and WISDM databases. Support Vector Machine creates a discriminant decision hyperplane with the maximum margin which supports a certain extent of reinforcement to facilitate higher confidence to classify future unknown data. Empirical results demonstrate that the proposed model is able to achieve 92% and 88% accuracy in UCI HAR and WISDM, respectively. Though the proposed model is effective to learn spatial and temporal features of inertial data, there is no analysis of multi-view data. It is believed that different feature spaces from different viewpoints are able to provide rich information to classify activities. Thus

in the future work, we plan to explore a deep hierarchical multi-view aggregative analysis for smartphone-based human activity recognition.

Acknowledgement

This research is supported by Fundamental Research Grant Scheme (FRGS), FRGS/1/2020/ICT02/MMU/02/7, and Multimedia University Internal Fund.

Nomenclatures

A	Activation function
B	Bias
C_n	Convolutional output/ feature map at nth layer
F_{n_t}	Forget gate output at n layer at timestep t
G	Input modulation gate output
I	Input gate
L	LSTM output
$\vec{L}$	Forward sequence
$\overleftarrow{L}$	Backward sequence
O	Output gate output
P	Pooling operation output
S	Memory cell state
V	Input vector
W	Weight
σ	Sigmoid function

Abbreviations

AUC	The Area under the Receiver Operating Characteristic Curve
BLSTM	Bidirectional Long Short Term Memory
CNN	Convolutional Neural Network
FP	False Positive
LSTM	Long Short Term Memory
SVM	Support Vector Machine
TP	True Positive
UCI HAR	UC Irvine Human Activity Recognition
WISDM	Wireless Sensor Data Mining

References

1. Guthold, R.; Stevens, G.A. .; Riley, L.M.; and Bull, F.C. (2018). Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. *The Lancet Global Health*, 6(10), e1077-e1086.
2. Booth, F.W.; Roberts, C.K.; and Laye, M.J. (2012). Lack of exercise is a major cause of chronic diseases. *Comprehensive Physiology*, 2(2), 1143-1211.
3. Cattadori, G.; Segurini, C.; Picozzi, A.; Padeletti, L.; and Anzà, C. (2018). Exercise and heart failure: an update. *ESC Heart Failure*: Wiley-Blackwell, 5(2), 222-232.

4. Voicu, R.A.; Dobre, C.; Bajenaru, L.; and Ciobanu, R.I. (2019). Human physical activity recognition using smartphone sensors. *Sensors* ((*Switzerland*), 19(3).
5. Carels, R.A.; Darby, L.A.; Rydin, S.; Douglass, O.M.; Cacciapaglia, H.M.; and O'Brien, W.H. (2005). The relationship between self-monitoring, outcome expectancies, difficulties with eating and exercise, and physical activity and weight loss treatment outcomes. *Annals of Behavioral Medicine*, 30(3), 182-190.
6. MacPherson, M.M.; Merry, K.J.; Locke, S.R.; and Jung, M.E. (2019). Effects of mobile health prompts on self-monitoring and exercise behaviors following a diabetes prevention program: secondary analysis from a randomized controlled trial. *JMIR mHealth and uHealth*, 7(9), e12956.
7. Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6), 976-990.
8. Kwapisz, J.R.; Weiss, G.M.; and Moore, S.A. (2011). Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter*, 12(2), 74-82.
9. Lee, S.M.; Cho, H.; and Yoon, S.M. (2017). Human activity recognition from accelerometer data using convolutional neural network. *IEEE International Conference on Big Data and Smart Computing* ((*BigComp*)., 62, 131-134.
10. Lee, K.; and Kwan, M.P. (2018). Physical activity classification in free-living conditions using smartphone accelerometer data and exploration of predicted results. *Computers*, *Environment and Urban Systems*, 67(Jan 2018), 124-131.
11. Kumari, S.; and Mitra, S.K. (2011). Human action recognition using DFT. *Proceedings* - 2011 3*rd National Conference on Computer Vision*, *Pattern Recognition*, *Image Processing and Graphics*, *NCVPRIPG* 2011, 239-242.
12. Peng, X.; Zou, C.; Qiao, Y.; and Peng, Q. (2014). Action recognition with stacked fisher vectors. In *Lecture Notes in Computer Science* ((*including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), 581-595.
13. Jalal, A.; Kamal, S.; and Kim, D. (2017). A depth video-based human detection and activity recognition using multi-features and embedded hidden Markov models for health care monitoring systems. *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(4), 54.
14. Sani, S.; Wiratunga, N.; and Massie, S. (2017). Learning deep features for knn-based human activity recognition. *International Conference on Case-based Reasoning*, 95-103.
15. Sun, J.; Fu, Y.; Li, S.; He, J.; Xu, C.; and Tan, L. (2018). Sequential human activity recognition based on deep convolutional network and extreme learning machine using wearable sensors. *Journal of Sensors*, 2018, 8580959.
16. Brezmes, T.; Gorricho, J.L.; and Cotrina, J. (2009). Activity recognition from accelerometer data on a mobile phone. *Lecture Notes in Computer Science* ((*including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), 5518 *LNCS*(PART 2), 796-799.
17. Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; and Reyes-Ortiz, J.L. (2013). A public domain dataset for human activity recognition using smartphones. In *ESANN* 2013 *proceedings*, 21*st European Symposium on Artificial Neural Networks*, *Computational Intelligence and Machine Learning*, 437-442.

18. Bayat, A.; Pomplun, M.; and Tran, D.A. (2014). A study on human activity recognition using accelerometer data from smartphones. *Procedia Computer Science*, 34, 450-457.
19. Lockhart, J.W. (2014). The benefits of personalized data mining approaches to human activity recognition with smartphone sensor data. Retrieved May 2014, from https://research.library.fordham.edu/dissertations/AAI1568376/
20. Ignatov, A. (2018). Real-time human activity recognition from accelerometer data using convolutional neural networks. *Applied Soft Computing Journal*, 62, 915-922.
21. Ronao, C.A.; and Cho, S.B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*, 59, 235-244.
22. Murad, A.; and Pyun, J.Y. (2017). Deep recurrent neural networks for human activity recognition. *Sensors* ((*Switzerland*), 17(11), 2556.
23. Ali, G.Q.; and Al-Libawy, H. (2021). Time-series deep-learning classifier for human activity recognition based on smartphone build-in sensors. *Journal of Physics: Conference Series*, 1973, 012127.
24. Shi, X.; Li, Y.; Zhou, F.; and Liu, L. (2018). Human activity recognition based on deep learning method. 2018 *International Conference on Radar*, *RADAR* 2018, 1-5.
25. Wang, J.; Chen, Y.; Hao, S.; Peng, X.; and Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119(1), 3-11.
26. Zeng, M.; Nguyen, L.T.; Yu, B.; Mengshoel, O.J.; Zhu, J.; Wu, P.; and Zhang, J. (2014). Convolutional neural networks for human activity recognition using mobile sensors. 6*th International Conference on Mobile Computing*, *Applications and Services*, 197-205.
27. Chen, Y.; Zhong, K.; Zhang, J.; Sun, Q.; and Zhao, X. (2016). LSTM networks for mobile human activity recognition. 2016 *International Conference on Artificial Intelligence: Technologies and Applications*, 50-53.
28. Nweke, H.F.; Teh, Y.W.; Al-garadi, M.A.; and Alo, U.R. (2018). Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Systems with Applications*. Elsevier Ltd, 105, 233-261.
29. Yu, S.; and Qin, L. (2018). Human activity recognition with smartphone inertial sensors using bidir-LSTM networks. *Proceedings* - 2018 3*rd International Conference on Mechanical*, *Control and Computer Engineering*, *ICMCCE* 2018, 219-224.
30. Hernández, F.; Suárez, L.F.; Villamizar, J.; and Altuve, M. (2019). Human activity recognition on smartphones using a bidirectional LSTM network. 2019 22*nd Symposium on Image*, *Signal Processing and Artificial Vision*, *STSIVA* 2019 - *Conference Proceedings*, 1-5.
31. Ogawa, A.; and Hori, T. (2017). Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks. *Speech Communication*, 89, 70-83.

32. Lima, W.S.; Souto, E.; El-Khatib, K.; Jalali, R.; and Gama, J. (2019). Human activity recognition using inertial sensors in a smartphone: An overview. *Sensors* ((*Switzerland*), 19, 3213.
33. Ronao, C.A.; and Cho, S.B. (2015). Evaluation of deep convolutional neural network architectures for human activity recognition with smartphone sensors. Retrieved 2015, from http://www.iro.umontreal.ca/~bengioy/dlbook,
34. Kolosnjaji, B.; and Eckert, C. (2015). Neural network-based user-independent physical activity recognition for mobile devices. *International Conference on Intelligent Data Engineering and Automated Learning*, In *Lecture Notes in Computer Science* ((*including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), 378-386.
35. Dungkaew, T.; Suksawatchon, J.; and Suksawatchon, U. (2017). Impersonal smartphone-based activity recognition using the accelerometer sensory data. *Proceeding of* 2017 2*nd International Conference on Information Technology*, *INCIT* 2017, 1-6.
36. Zhang, X.; Wong, Y.; Kankanhalli, M.S.; and Geng, W. (2019). Hierarchical multi-view aggregation network for sensor-based human activity recognition. *PLoS ONE*, 14(9), e0221390.
37. Ronao, C.A.; and Cho, S.B. (2017). Recognizing human activities from smartphone sensors using hierarchical continuous hidden Markov models. *International Journal of Distributed Sensor Networks*, 13(1).
38. Seto, S.; Zhang, W.; and Zhou, Y. (2015). Multivariate time series classification using dynamic time warping template selection for human activity recognition. In *Proceedings* - 2015 *IEEE Symposium Series on Computational Intelligence*, *SSCI* 2015, 1399-1406.