electronics

Article

Human Activity Recognition Using an Ensemble Learning Algorithm with Smartphone Sensor Data

Tan-Hsu Tan [1], Jie-Ying Wu [1], Shing-Hong Liu [2,*] and Munkhjargal Gochoo [3]

[1] Department of Electrical Engineering, National Taipei University of Technology, Taipei 10608, Taiwan; thtan@ntut.edu.tw (T.-H.T.); jeremydadmom@gmail.com (J.-Y.W.)
[2] Department of Computer Science and Information Engineering, Chaoyang University of Technology, Taichung 413310, Taiwan
[3] Department of Computer Science and Software Engineering, United Arab Emirates University, Al-Ain 15551, United Arab Emirates; mgochoo@uaeu.ac.ae
* Correspondence: shliu@cyut.edu.tw; Tel.: +886-4-233230000-7811

Abstract: Human activity recognition (HAR) can monitor persons at risk of COVID-19 virus infection to manage their activity status. Currently, many people are isolated at home or quarantined in some specified places due to the spread of COVID-19 virus all over the world. This situation raises the requirement of using the HAR to observe physical activity levels to assess physical and mental health. This study proposes an ensemble learning algorithm (ELA) to perform activity recognition using the signals recorded by smartphone sensors. The proposed ELA combines a gated recurrent unit (GRU), a convolutional neural network (CNN) stacked on the GRU and a deep neural network (DNN). The input samples of DNN were an extra feature vector consisting of 561 time-domain and frequency-domain parameters. The full connected DNN was used to fuse three models for the activity classification. The experimental results show that the precision, recall, F1-score and accuracy achieved by the ELA are 96.8%, 96.8%, 96.8%, and 96.7%, respectively, which are superior to the existing schemes.

Keywords: ensemble learning algorithm; human activity recognition; gated recurrent units; convolutional neural network

Citation: Tan, T.-H.; Wu, J.-Y.; Liu, S.-H.; Gochoo, M. Human Activity Recognition Using an Ensemble Learning Algorithm with Smartphone Sensor Data. *Electronics* **2022**, *11*, 322. https://doi.org/10.3390/electronics11030322

Academic Editor: Giovanni Dimauro

Received: 14 December 2021
Accepted: 19 January 2022
Published: 20 January 2022

1. Introduction

The COVID-19 virus has been spreading all over the world for more than one and a half years, which has led many people to be isolated at home, or quarantined in some specified spaces. Therefore, people's physical activity is restricted. However, as reported by a previous study, physical inactivity causes more than 5 million deaths worldwide, which does great harm to the finances of public health systems [1]. López-Bueno et al. investigated changes in physical activity (PA) levels during the first week of confinement in Spain where participants reduced their weekly PA levels by 20% [2]. The study of Matos et al. shows that body weights increased, and the weekly energy expenditure and quality of life were reduced for Brazilians during the pandemic [3]. Thus, people should maintain their levels of physical activity to stay healthy when they are isolated at home or quarantined. A human activity recognition (HAR) system can be applied to monitor the persons at risk of COVID-19 virus infection to manage their activity status. In addition, the HAR can also be used in the telecare and/or health management by observing the fitness of healthy people or patients infected with COVID-19 in daily life, such as time spent exercising and resting [4,5]. Therefore, the research on HAR has received much attention in recent years.

To perform HAR, various sensors are used to extract the human activity data [6,7]. Image-based and sensor-based methods are two commonly used data sensing methods [8]. The image-based methods usually use visual sensing devices, such as video cameras and photo cameras [9,10], to monitor human activities. However, their major disadvantages

include the invasion of privacy, large size, and the limitation of indoor installation (without mobility). On the other hand, in sensor-based method the user needs to wear various sensors, such as an accelerometer, gyroscope, and strain gauge, on their wrist or limbs [11,12]. Though it has advantage of conducting ubiquitous HAR, wearing sensors on the body for a long time makes the users uncomfortable. As reported by the National Development Commission, the number of smartphone users in Taiwan reached 29.25 million in June 2019, indicating 1.24 smartphones per person [13]. Since smartphones have already penetrated into people's daily life, many studies on HAR using the accelerometers and gyroscopes embedded in smartphones have been conducted.

In the work of [14], the logistic model trees (LMT) machine learning method was employed to recognize human activity. The experimental results indicated that the accuracy of the LMT method reaches 94.0%. Bao et al. [15] firstly segmented the action signals with 128 samples and 50% overlapping, then a geometric template matching algorithm was used to classify each space into a corresponding action. In the last stage, the Bayesian principle and voting rule were combined to fuse the results of a k-nearest neighbor classifier. Cruciani et al. [16] performed HAR utilizing the gold standard human crafted features and one-dimension (1D) convolutional neural network (CNN) which achieved an F1-Score of 93.4%. Wu and Zhang [17] employed an CNN model to execute HAR, attaining an accuracy of 95%. Wang et al. [18] proposed an attention-based CNN for the HAR using the weakly labeled activity data. In this work, to save the manpower and computing resources during the process of strictly data labeling, a weakly supervised model based on recurrent attention learning (RAL) was presented [19]. Taking the advantages of CNNs in learning complex activities and long short-term memory (LSTM) networks in capturing temporal information from time series data, He et al. [20] suggested a combination of CNN and LSTM networks for the HAR. Recently, a deep neural network that combined convolutional layers with LSTM was also proposed [21], which achieved an accuracy of 95.8%. Yu et al. [22] proposed a multilayer parallel LSTM network for the HAR. Sikder et al. [23] presented a two-channel CNN for the HAR, which employed the frequency and power features of the activity data. Intisar and Zhao [24] proposed a selective modular neural network (SMNN) that was a stacking architecture, consisting of a routing module and expert module, to enhance the accuracy of HAR. However, this model spent much time for the model training. The previous studies have suggested that the signals extracted by the accelerometer and gyroscope corresponding to different activities should be considered as temporal features. Since the recurrent neural network (RNN) can effectively describe the time dependency between different samples and the memory function, many studies have applied the LSTM to perform HAR [19–23].

The signals of accelerometer, gyroscope, and strain gauges obtained from human activities could be considered as time series data. In recent years, various ensemble deep learning models have been proposed to solve the problem of time series classification. Fawaz et al. [25] have proposed an ensemble of CNN models, named InceptionTime, to deal with the issue of time series classification. Karim et al. [26] have transformed the LSTM fully convolutional network (LSTM-FCN) and attention LSTM-FCN (ALSTM-FCN) into a multivariate time series classification model by augmenting the fully convolutional block with a squeeze-and-excitation block. Xiao et al. [27] have proposed a robust temporal feature network (RTFN) which consists of a temporal feature network (TFN) and an attention LSTM network for feature extraction in the problem of time series classification.

Some stacking deep neural networks also have been used to improve the activity recognition rate. The stacking architecture integrates various neural networks to gain the advantages for specific tasks. Li et al. [28] used the Dempster-Shafer (DS) evidence theory to build the ensemble DS-CNN model for the event sound recognition. Batchuluun et al. [29] employed the CNN stacked with LSTM and deep CNN followed by score fusion to capture more spatial and temporal features for the gait-based human identification. Du et al. [30] applied two-dimension (2D) CNN which stacked up a gated recurrent unit (GRU) to

obtain the features of micro-Doppler spectrograms. The features with the time-steps were recognized by the GRU for HAR.

The ensemble learning algorithm (ELA) is a technique that combines the predictions of multiple classifiers to form a single classifier, which generally results in a higher accuracy than that of any of the individual classifiers [31,32]. Its theoretical and practical studies have demonstrated that a good ELA was the individual classifiers in the ELA which accuracies are close and errors are distributed on the different parts [33,34]. In general, the ELA consists of two parts: how to generate differentiated individual classifiers and how to fuse them. In the generation of individual classifiers, two kinds of generation strategies, namely, the heterogeneous type and the homogeneous type are commonly employed. The former is that individual classifiers are generated using various learning algorithms. The latter uses the same learning algorithm, so different settings are necessary. Thus, Deng et al. [35] adopted linear and log-linear stacking methods to fuse convolutional, recurrent and the fully connected deep neural networks (DNNs). Xie et al. [36] proposed three DNN-based ensemble methods, which fused a series of classifiers whose inputs are the representation of intermediate layers.

This study aims to recognize the human activities with the data extracted from sensors embedded in the smart phone. An ELA combining GRU, stacking CNN+GRU and DNN was proposed to perform the HAR. The sensor data are the input samples of GRU and stacking CNN + GRU. The 561 parameters obtained from those raw sensor data are utilized as the input samples of DNN. Then, the outputs corresponding to stacking CNN + GRU, GRU, and DNN were combined to classify the six activities using the fully connected DNNs. The HAR dataset employed in this work is an open source provided by the UCI, School of Information and Computer Sciences [28]. This dataset collects six sets of activity data via the accelerometer and gyroscope built into two smartphones. An extra feature vector consisting of 561 parameters is generated from time-domain and frequency-domain based on the raw sensor data. The experimental results showed that the proposed ELA scheme outperforms the existing studies.

2. Materials and Methods

The structure of the proposed ELA for HAR consists of two parts, the feature extraction unit and classification unit. The features of sensor data were extracted by the GRU and stacking CNN + GRU, respectively. The extracted features together with the extra 561 parameters were inputted to the classification unit for activity recognition.

2.1. UCI-HAR Dataset

The UCI-HAR dataset [37] was built via recording 30 subjects aged 19–48 who wore a smartphone (Samsung Galaxy S II) with embedded inertial sensors around their waist. During the recording, all subjects followed each activity protocol. In this work, six activities to be recognized are sitting, standing, lying, walking, walking downstairs and walking upstairs, because they are the most common activities performed in the daily life. The activity signals were collected via the three-axial acceleration and three-axial angular speed with a sampling rate of 50 Hz. The gravitational force is assumed to have only low frequency components, therefore signals of three-axial acceleration were filtered by a lowpass filter with a cutoff frequency of 0.3 Hz to generate gravitational signals. The body-motion signals were extracted from the raw signals minus the gravitational signals. The nine signals were sampled in a fix-width sliding windows of 2.56 s with 50% overlapping between them. Thus, a sample contained 9-channel signals (nine signals), and each channel had 128 points. All samples were supported by the UCI-HAR dataset. The number of time-domain and frequency-domain parameters of a sample was 561 [29,38]. The number of training and testing samples was 7352 and 2947, respectively. In the training samples, 2206 samples were used for model validation. Table 1 illustrates sample numbers of the six activities.

Table 1. Sample number of six activities for training and testing of model with UCI-HAR dataset.

Activity	Training Number	Testing Number
Sitting	1286	491
Standing	1374	532
Lying	1407	537
Walking	1226	496
Walking upstairs	1073	471
Walking downstairs	986	420
Sum	7352	2947

Extra signals were obtained from the nine signals, which were the Euclidean magnitude (mag) and time differentiation (jerk). Thirteen signals were transformed to the frequency domain via the discrete Fourier transform (DFT). Table 2 shows the detail of sensor signals including the channel number of each sensor signal and the signals which were transformed to the frequency domain. A total of 561 parameters were then derived from the twenty signals based on their mean, standard deviation, median absolute value, maximum value in window, minimum value in window, signal magnitude area (SMA), energy, interquartile range, entropy, auto-regression coefficient (AR), correlation coefficient (R), maximum frequency component, mean frequency, skewness, kurtosis, energy band, and angular [38]. The 561 parameters were supported by the UCI-HAR dataset.

Table 2. Detail of sensor signals.

Signal	Channel Number	Applying DFT
Body Acc.	3	yes
Gravity Acc.	3	no
Body Acc jerk	3	yes
Body A.S.	3	yes
Body A.S. jerk	3	no
Body Acc. mag.	1	yes
Gravity Acc. mag.	1	no
Body Acc. Jerk mag.	1	yes
Body A.S. mag.	1	yes
Body A.S. jerk mag.	1	yes

Abbreviation: Acc. stands for accelerometer, A.S. stands for angular speed, Mag. stands for magnitude.

2.2. UCI-WIDSM Dataset

The UCI-WISDM dataset [39] consists of tri-axial accelerometer and gyroscope data samples obtained from 51 volunteer subjects carrying an Android phone (Google Nexus 5/5x or Samsung Galaxy S5) in the front pockets of pants and an Android watch (LG G Watch) at their wrist while performing eighteen activities. The sampling rate was 20 Hz. The 12 signals were sampled in a fix-width sliding windows of 6.4 s with 50% overlapping between them. Thus, a sample contained 12-channel signals, and each channel had 128 points. Table 3 illustrates sample numbers of the eighteen activities. The numbers of training and testing samples were 34,316 and 14,707, respectively.

Table 3. Sample number of eighteen activities for model training and testing with UCI-WISDM dataset.

Activity	Training Number	Testing Number
Walking	1921	807
Jogging	1901	827
Stairs	1920	808
Sitting	1895	833
Standing	1891	837

Table 3. *Cont.*

Activity	Training Number	Testing Number
Kicking (Soccer ball)	1932	797
Dribbling (Basketball)	1906	822
Catching (Tennis ball)	1893	835
Typing	1885	843
Writing	1880	766
Clapping	1945	783
Brushing teeth	1876	852
Folding Clothes	1919	809
Eating Pasta	1915	814
Eating Soup	1928	800
Eating sandwich	1950	778
Eating Chips	1898	830
Drinking from Cup	1861	866

2.3. UCI-OPPORTUNITY Dataset

The body-worn sensors of UCI-OPPORTUNITY dataset [40] consists of date collected by five inertial measurement units, 12 tri-axial acceleration sensors, and one IntertiaCube3 sensor from 12 volunteer subjects while performing five activities. The sampling rate was 30 Hz. The five inertial measurement units were placed on the sports jacket, the InertiaCube3 sensor was mounted on the left shoe, and the 12 tri-axial acceleration sensors were mounted on the upper body, hips, and legs. The inertial measurement unit date includes the results from a tri-axial acceleration sensor, gyroscope, tri-axial magnetic field sensor, and the orientation of the sensor with respect to a world coordinate system in quaternions. The InertiaCube3 sensor data includes tri-axial global Euler angles (deg), acceleration in the navigation coordinate frame (m/s^2), acceleration in sensor body coordinate frame (m/s^2), and angular rotation speed in body coordinate frame (rad/s). The 113 signals were sampled in a fix-width sliding windows of 4.26 s with 50% overlapping between them. Thus, a sample contained 113-channel signals, and each channel had 128 points. Table 4 illustrates sample numbers of the eighteen activities. The numbers of training and testing samples were 8717 and 1854, respectively.

Table 4. Sample number of five activities for model training and testing with UCI-OPPORTUNITY dataset.

Activity	Training Number	Testing Number
Standing	1448	378
Walking	3613	585
Sitting	2041	424
Lying	1386	379
Null	229	88

2.4. GRU Model

Figure 1 shows the structure of the GRU model, where the GRU used to extract the features of sensor signal is with unit number of 128 and batch size of 32. The control reset gate and update gate use sigmoid function, and hidden state uses tanh function. The fully connected layer consists of three layers which is used to classify the activities. The three layers have dimension of 128, 64, and 6, respectively. The activation functions are respectively ReLU in hidden layers and sofmax in the output layer. The Adam optimizer is used with a learning rate of 0.0001. The objective function is Categorical Cross-Entropy function:

$$CE = -\log\left(\frac{\exp(a_k)}{\sum_{i=1}^{N}\exp(a_i)}\right) \quad (1)$$

where N is 6, a_k is the score of sofmax for the positive class, and a_i is the scores inferred by the net for each class. Table 5 shows the settings of the GRU model.

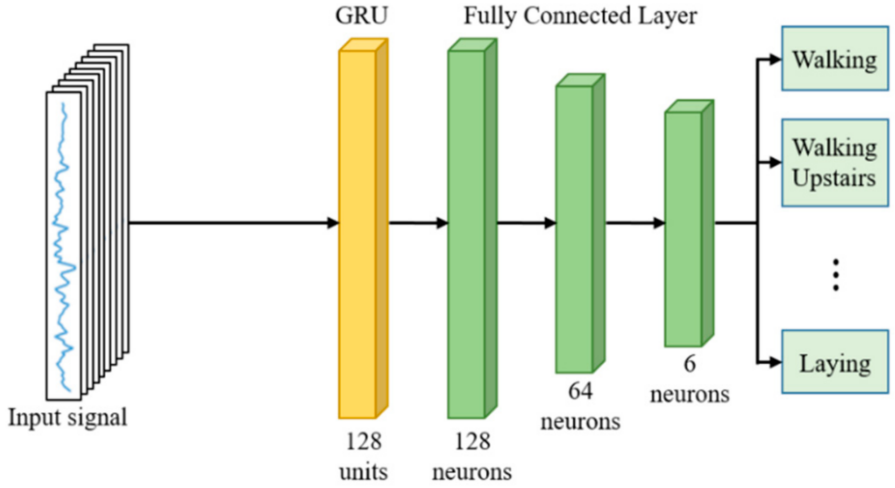


Figure 1. Structure of GRU model.

Table 5. The setting of GRU model.

Type	Channel Number	Input Size	Output Size
GRU		128 × 9	128
Flatten	1	128	128
Fc	1	128	64
Out	1	64	6

2.5. Sacking CNN + GRU Model

Because human activities are chronologically ordered, the sensor signals are time-series data. A time-distributed layer consisting of four CNN, i.e., two pairs of 1D convolutional network and maximal pool layer, is stacked on the GRU. Thus, a sample was separated into four segments and each segment contained 32 points. Figure 2 shows the structure of stacking CNN + GRU model. In the convolutional layer, the number of filters is 64, kernel size is 5, stride is 1, and padding is 4. In the polling layer, the kernel size is 2 and strike is 2. The activation function is ReLU. The unit number of GRU is 128. Batch size is 32. The control reset gate and update gate use sigmoid function, and hidden state uses tanh function. The stacking CNN + GRU is used to extract the features of sensor signals. The fully connected layer comprised of three layers, is employed to classify the activities. The numbers of neurons of the three layers are 128, 64, and 6, respectively. The activation functions are ReLU in hidden layers and softmax in output layer. The objective function is Categorical Cross-Entropy function, the Adam optimizer is used, and the learning rate is 0.0001. Table 6 shows the setting of the stacking CNN + GRU model.

2.6. Ensemble Learning Algorithm

Figure 3 shows the full structure of the proposed ELA model, which performs the recognition task with three branches. In the top branch, the features of sensor data are extracted by the stacking CNN + GRU. Then the extracted features are inputted to the fully connected layer. In the middle branch, the data features extracted by the GRU are sent to the fully connected layer. In the bottom branch the 561 time-domain and frequency-domain parameters (feature vector) obtained from raw sensor data are directly sent to the three layers of DNN in which the neuron numbers are 128, 64, and 6, respectively. These layers

could be considered as the multilayer neural network which has two hidden layers and an output layer to classify the six activities. Then, the outputs of 3 branches are fused and sent to the full connected DNN with three layers for the activity classification. The number of neurons in each layer of the fully connected layer are 18, 10, and 6 respectively. The details of GRU and stacking CNN + GRU are illustrated in Sections 2.4 and 2.5. The activation functions are ReLU in hidden layers and sofmax in output layer. The objective function is Categorical Cross-Entropy function, the Adam optimizer is used, and the learning rate is 0.0001.

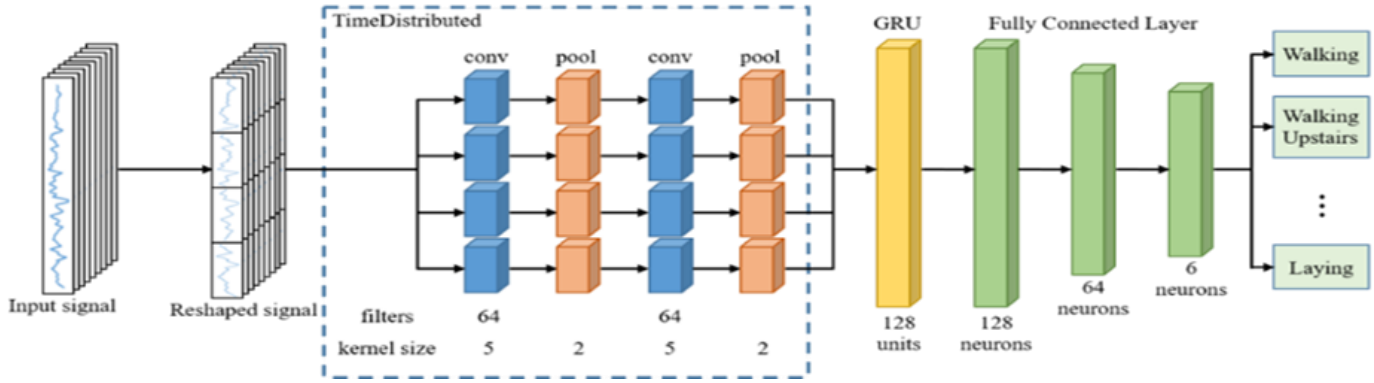


Figure 2. Structure of stacking CNN + GRU model.

Table 6. The setting of stacking CNN + GRU model.

Type	Filter Size	Channel Nummer	Input Size	Output Size
Conv 1	5	64	32 × 9	32 × 64
Max pool	2		32 × 64	16 × 64
Conv2	5	64	16 × 64	16 × 64
Max pool	2		16 × 64	8 × 64
GRU		128	32 × 64	128
Flatten		1	128	128
Fc		1	128	64
Out		1	64	6

2.7. Statistical Analysis

According to the proposed method, a sample is considered as true positive (TP) when the classification activity is correctly recognized; false positive (FP) when the classification activity is incorrectly recognized; true negative (TN) when the activity classification is correctly rejected, and false-negative (FN) when the activity classification is incorrectly rejected. In this work, the performance of the proposed method was evaluated using the measures taken in Equations (2)–(5):

$$Precision\ (\%) = \frac{TP}{TP + FP} \times 100\% \tag{2}$$

$$Recall\ (\%) = \frac{TP}{TP + FN} \times 100\% \tag{3}$$

$$F_1\text{-}score\ (\%) = \frac{2 \times precision \times Recall}{Precision + Reacll} \times 100\% \tag{4}$$

$$Accuracy\ (\%) = \frac{TP + TN}{TP \mp TN + FP + FN} \times 100\% \tag{5}$$

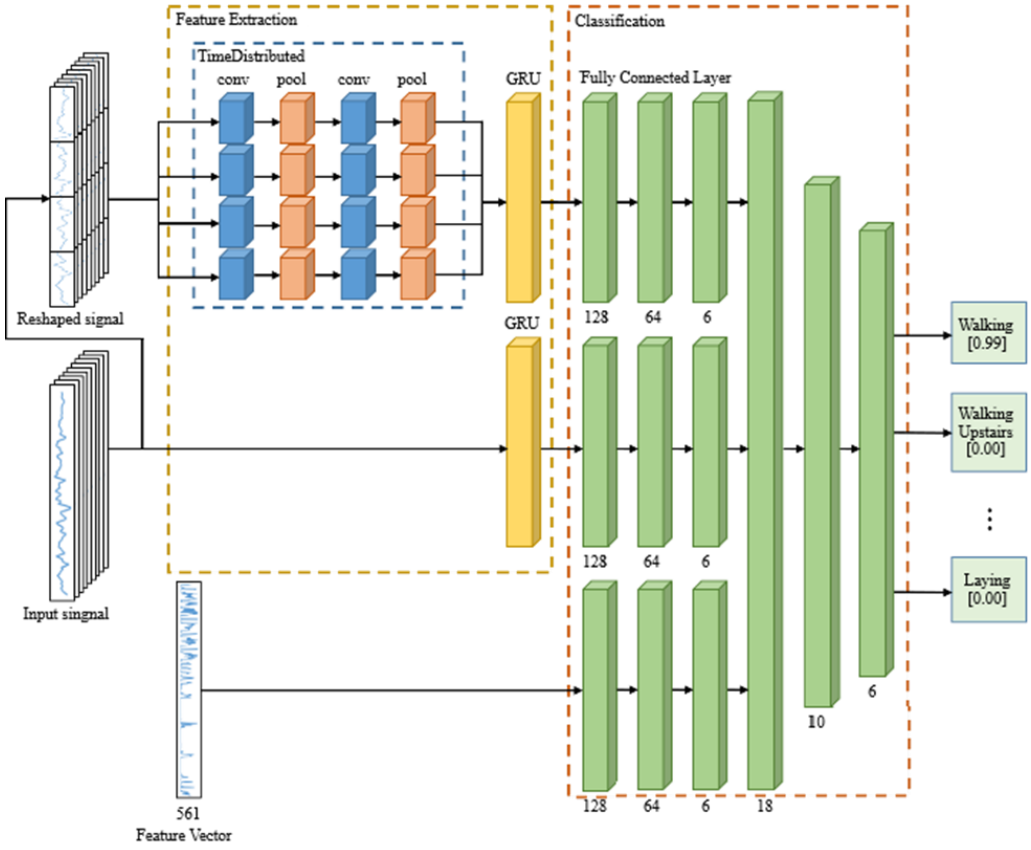
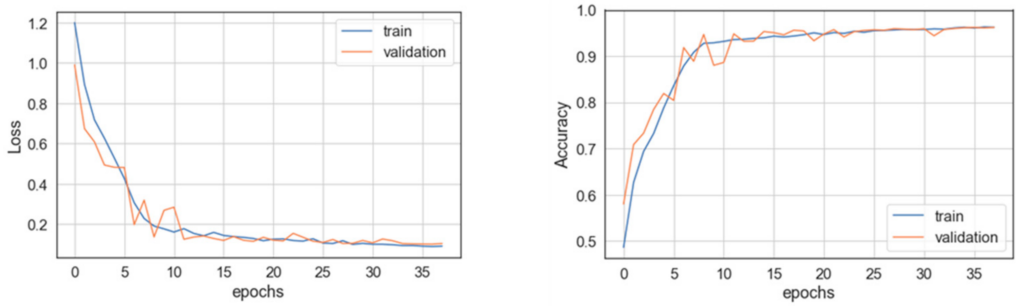


Figure 3. Full structure of proposed ELA model.

3. Results

In this study, the hardware employed was an Intel Core i7-8700 CPU and a GeForce GTX1080 GPU. The operating system was the Ubuntu 16.04LTS software, the development system was Anaconda 3 for Python 3.7 version, the deep learning tool was Pytorch 1.10, and the compiler was a Jupyter Notebook. A series of experiments is conducted to evaluate performance of the GRU model, stacking CNN + GRU model, and proposed ELA model. Figure 4 shows the training and validation curves attained by the GRU model. The loss functions obtained in training (blue line) and validation (orange line) are exhibited in Figure 4a, and the accuracies attained in training (blue line) and validation (orange line) are illustrated in Figure 4b. The accuracy achieved the best when epoch equals 37. Table 7 shows the performance of the GRU model for the six activities. As shown, the average precision, recall, F1-score, and accuracy are 92.7%, 92.6%, 92.5%, and 92.5%, respectively. While all average measures of the six activities are higher than 90%, the GRU model demonstrated an unsatisfied performance on recognizing the sitting and standing activities, because their F1-scores are less than 90%.

Figure 5 shows the training and validation curves obtained by the stacking CNN + GRU model. The resultant loss function in training (blue line) and validation (orange line) are presented in Figure 5a, and the obtained accuracies in training (blue line) and validation (orange line) are shown in Figure 5b. The accuracy reached the best when epoch is equal to 39. Table 8 illustrates the performance of the stacking CNN + GRU model for the six activities. As shown, the average precision, recall, F1-score, and accuracy are 93.0%, 92.9%, 92.9%, and 92.7% respectively. Though all average measures for six activities are higher than 90%, the stacking CNN + GRU model exhibited an unsatisfied performance on recognizing the sitting and standing activities, because their F1-scores are less than 90%.

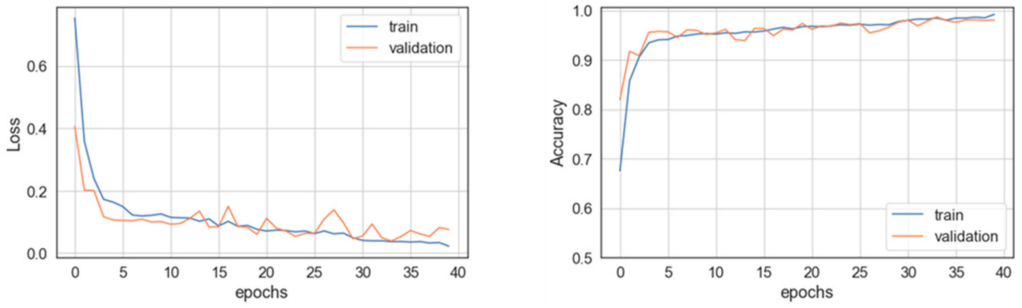


(**a**) (**b**)

Figure 4. Training and validation curves of GRU model, (**a**) Loss functions in training (blue line) and validation (orange line), (**b**) accuracies obtained in training (blue line) and validation (orange line).

Table 7. Performance of GRU model for six activities.

	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Walking	96.4	91.7	94.0	92.5
Walking Upstairs	95.6	93.0	94.3	
Walking Downstairs	91.1	99.5	95.1	
Sitting	89.9	79.4	84.3	
Standing	83.1	91.7	87.2	
Lying	100	100	100	
Average	92.7	92.6	92.5	



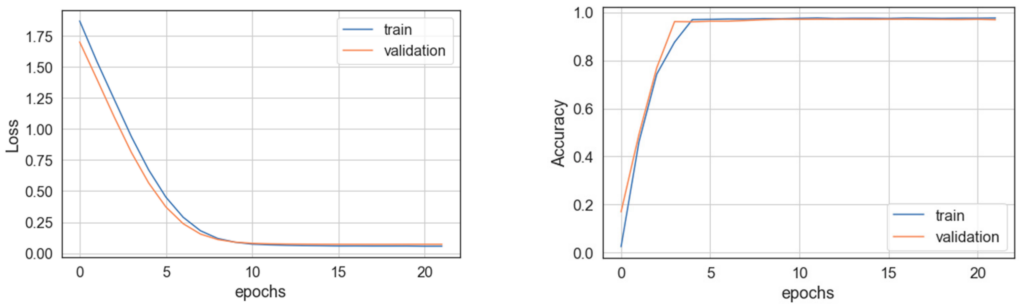
(**a**) (**b**)

Figure 5. Training and validation curves of stacking CNN + GRU model, (**a**) Loss functions in training (blue line) and validation (orange line), (**b**) accuracies obtained in training (blue line) and validation (orange line).

Table 8. Performance of stacking CNN + GRU model for six activities.

	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Walking	99.6	97.6	98.6	92.7
Walking Upstairs	92.6	97.9	95.1	
Walking Downstairs	97.2	99.5	98.4	
Sitting	87.6	78.0	82.5	
Standing	82.8	89.7	86.1	
Lying	98.1	95.0	96.5	
Average	93.0	92.9	92.9	

Figure 6 shows the training and validation curves of the proposed ELA model which only combines GRU and stacking CNN + GRU. The obtained loss functions in training (blue line) and validation (orange line) are exhibited in Figure 6a, and the attained accuracies in training (blue line) and validation (orange line) are illustrated in Figure 6b. The accuracy achieved the best when epoch equals 10. Table 9 shows performance of the ELA model without the 561 parameters which fused the outputs of three branches for activity classification. The average precision, recall, F1-score, and accuracy are 93.5%, 93.6%, 93.5%, and 93.4%, respectively. However, the performances of ELA without 561 parameters for recognizing the sitting and standing activities do not have the significant raise.



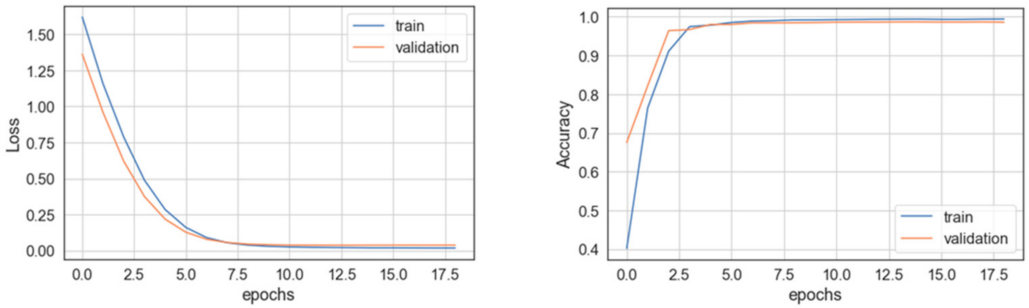
(**a**) (**b**)

Figure 6. Training and validation curves of proposed ELA model which only combines GRU and stacking CNN + GRU, (**a**) Loss functions in training (blue line) and validation (orange line), (**b**) accuracies in training (blue line) and validation (orange line).

Table 9. Performance of ELA model which only combines GRU and stacking CNN + GRU for six activities.

	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Walking	99.3	92.3	95.7	93.4
Walking Upstairs	95.9	98.1	97.0	
Walking Downstairs	93.9	99.8	96.8	
Sitting	87.1	82.7	84.8	
Standing	85.3	88.5	86.9	
Lying	99.4	100	99.7	
Average	93.5	93.6	93.5	

Figure 7 shows the training and validation curves of the proposed ELA model. The obtained loss functions in training (blue line) and validation (orange line) are exhibited in Figure 7a, and the attained accuracies in training (blue line) and validation (orange line) are illustrated in Figure 7b. The accuracy achieved the best when epoch equals 18. Table 10 shows performance of the ELA model which fused the outputs of three branches for activity classification. The average precision, recall, F1-score, and accuracy are 96.8%, 96.8%, 96.8%, and 96.7%, respectively. Notably, the F1-scores of six activities are all higher than 90%. In addition, the F1-scores obtained for recognizing the sitting and standing activities are 91.7% and 92.9%, respectively, which achieved the significant improvement as compared to the GRU and stacking CNN + GRU models.



(**a**) (**b**)

Figure 7. Training and validation curves of proposed ELA model, (**a**) Loss functions in training (blue line) and validation (orange line), (**b**) accuracies in training (blue line) and validation (orange line).

Table 10. Performance of ELA model for six activities.

	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Walking	99.6	98.2	98.9	96.7
Walking Upstairs	98.3	98.7	98.5	
Walking Downstairs	98.1	99.5	98.8	
Sitting	93.3	90.2	91.7	
Standing	91.7	94.0	92.9	
Lying	99.6	100	99.8	
Average	96.8	96.8	96.8	

In order to further verify the effectiveness of the models employed in this study, the WISDM and OPPORTUNITY datasets were also employed. Since the WISDM and OPPORTUNITY datasets do not support the 561 time-domain and frequency-domain parameters, the proposed ELA model was not applied in the experiment, and therefore only the GRU and stacking CNN + GRU were used for performance evaluation. Figure 8 illustrates the F1-score of the GUR and stacking CNN + GRU models employing UCI-HAR, WISDM and OPPORTUNITY datasets. The F1-score of the GRU and stacking CNN + GRU models based on the WISDM and OPPTUNITY datasets are 83.8% and 86.2%, and 91.7% and 87.4%, respectively, which are lower than those with HCI-HAR dataset.

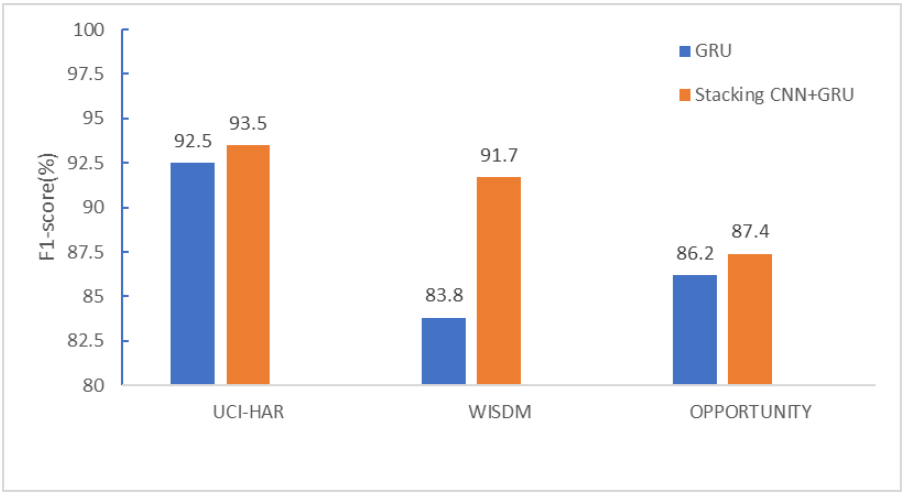


Figure 8. F1-score values of GUR and stacking CNN + GRU models for using UCI-HAR, WISDM and OPPORTUNITY datasets.

Table 11 presents the computation time required for testing each activity sample based on the GRU, stacking CNN + GRU, and ELA models. The result indicates that the ELA spent the longest time (1.681 ms), and the GRU model spent the shortest time (0.031 ms).

Table 11. Computation time required for testing each activity sample based on three models.

Model	Time (ms)
GRU	0.031
Stacking CNN + GRU	0.817
ELA	1.681

4. Discussion

In pattern recognition, the procedure of the traditional methods is that feature vectors are firstly extracted from the raw data. Then, a suitable model based on the feature vectors is employed for classification [41]. In recent years, a great success in complicated fields of pattern recognition is the DNN with more than three layers, which combines feature extraction and classification into a signal learning structure and directly constructs a decision function [42]. The major core of generation strategies is to make individual classifiers that depend on errors and diversity to enhance the performance of classification, such as the commonly used Simple Average and Weighted Average scheme [43]. In addition, some other schemes combining multiple classifiers are suggested, such as Dempster-Shafer Combination Rules [44], Stacking Method [32], and Second-Level Trainable Combiners [45]. Ensemble learning has been proved to be able to improve the generalization ability effectively in both theory and practice [46]. In this study, we have proposed the ELA model to classify the six activities. The specific point of the samples employed for model training is that they are the combination of feature vector extracted from the raw senor data. The feature vector are the time-domain and frequency-domain parameters generated from the raw senor data. In Table 9, the performance of the ELA model which only combining GRU and stacking CNN + GRU is better than the individual GRU and stacking CNN + GRU models. However, the results of recognizing sitting and standing activities do not exhibit a satisfied performance, because their F1-scores are less than 90%. These results are the same as the results of individual GRU and stacking CNN+GRU models illustrated in Tables 7 and 8.

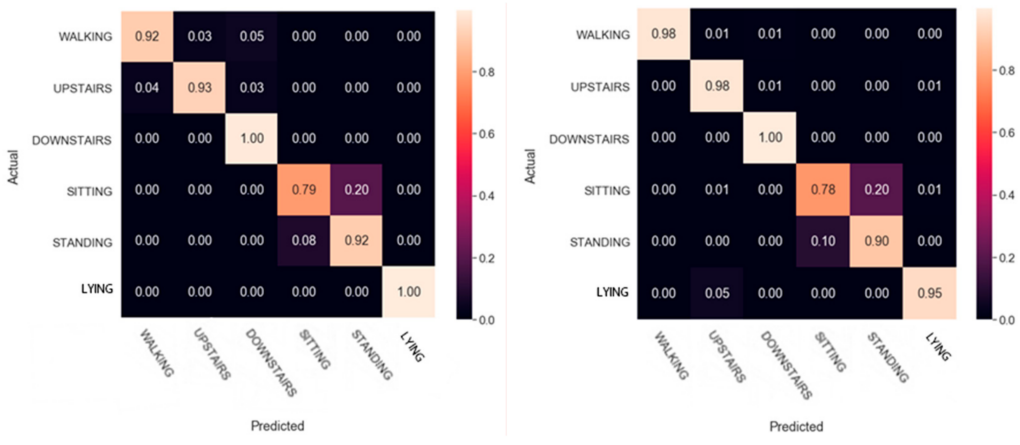
According to Tables 7 and 8, the performance of the stacking CNN + GRU model is slightly better than that of the GRU model. The comparative results indicate that, the averages of precision, recall, F1-score, and accuracy are 93.0% vs. 92.7%, 92.9% vs. 92.6%, 92.9% vs. 92.5%, and 92.7% vs. 92.5%. The major problems of the individual GRU and stacking CNN + GRU models are that two activities, the sitting and standing, cannot be classified well enough. However, in Table 10, the proposed ELA model shows a significantly improved performance. Especially, the F1-scores of sitting and standing activities are higher than 90%. The averages of precision, recall, F1-score, and accuracy achieved by the proposed ELA are 96.8%, 96.8%, 96.8%, and 96.7%, respectively. Thus, extracting the useful features as the input patterns could effectively improve the performance in the practice for the ELA model.

Table 12 shows the comparative result of our method with other studies using the UCI-HAR dataset. Notably, the previous studies usually only used sensor signals to perform activity recognition with deep learning methods [17–23], or used 561 parameters with machine learning methods [14–16]. As shown, the proposed ELA model attains performance of F1-score and accuracy of 96.8% and 96.7%, respectively, which is among the best.

Table 12. Comparative result of various methods using UCI-HAR dataset.

Ref.	Classification Method	F1-Score (%)	Accuracy (%)
[14]	Logistic Model Tree	N/A	94.0
[15]	GTM-Bayes-Voting	92.4	92.5
[16]	HCF-NN	95.5	N/A
[17]	CNN	N/A	95
[18]	Attention-CNN	N/A	93.4
[19]	RAL	N/A	94.8
[20]	CNN-LSTM	93.4	93.4
[21]	LSTM-CNN	95.8	95.8
[20]	Multilayer Parallel LSTM	N/A	94.3
[22]	Multichannel CNN	95.3	95.3
[23]	SMNN	N/A	96.0
Proposed method	ELA Model	96.8	96.7

We have analyzed the confusion matrices of the GRU model and stacking CNN + GRU model as shown in Figure 9. The result exhibits that the misclassification of standing and sitting activities occurs frequently. Therefore, in this study, the 561 time-domain and frequency-domain parameters were applied to enhance the HAR performance. Figure 10 illustrates the feature differences of mean and standard division (SD) between the training and testing data of standing and sitting activities introduced by 561 parameters. In Figure 10a, the blue line is the mean differences of 561 parameters between the training set of standing activity (x4_mean) and the testing set of sitting activity (x5_mean), and the orange line is the mean differences of 561 parameters between the training set (x4_mean) and testing set of the standing activity (x6_mean). We can find that the values in the blue line are much higher than the values in the orange line. In Figure 10b, the blue line is the SD differences of 561 parameters between the training set of standing activity (x4_SD) and the testing set of sitting activity (x5_SD), and orange line is the SD differences of 561 parameters between the training set (x4_SD) and testing set of standing activity (x6_SD). We can find that the values in the blue line are also much higher than the values in the orange line. The results indicate that introducing the 561 parameters broadens the feature difference between training data of the standing activity and the testing data of the sitting activity, while decreasing the feature difference between training and testing data of the standing activity. In Table 9, the results are distinct from those mentioned above if the 561 parameters are not included in the ELA scheme.



(**a**) (**b**)

Figure 9. (**a**) Confusion matrix of GRU model, (**b**) confusion matrix of stacking CNN + GRU model.

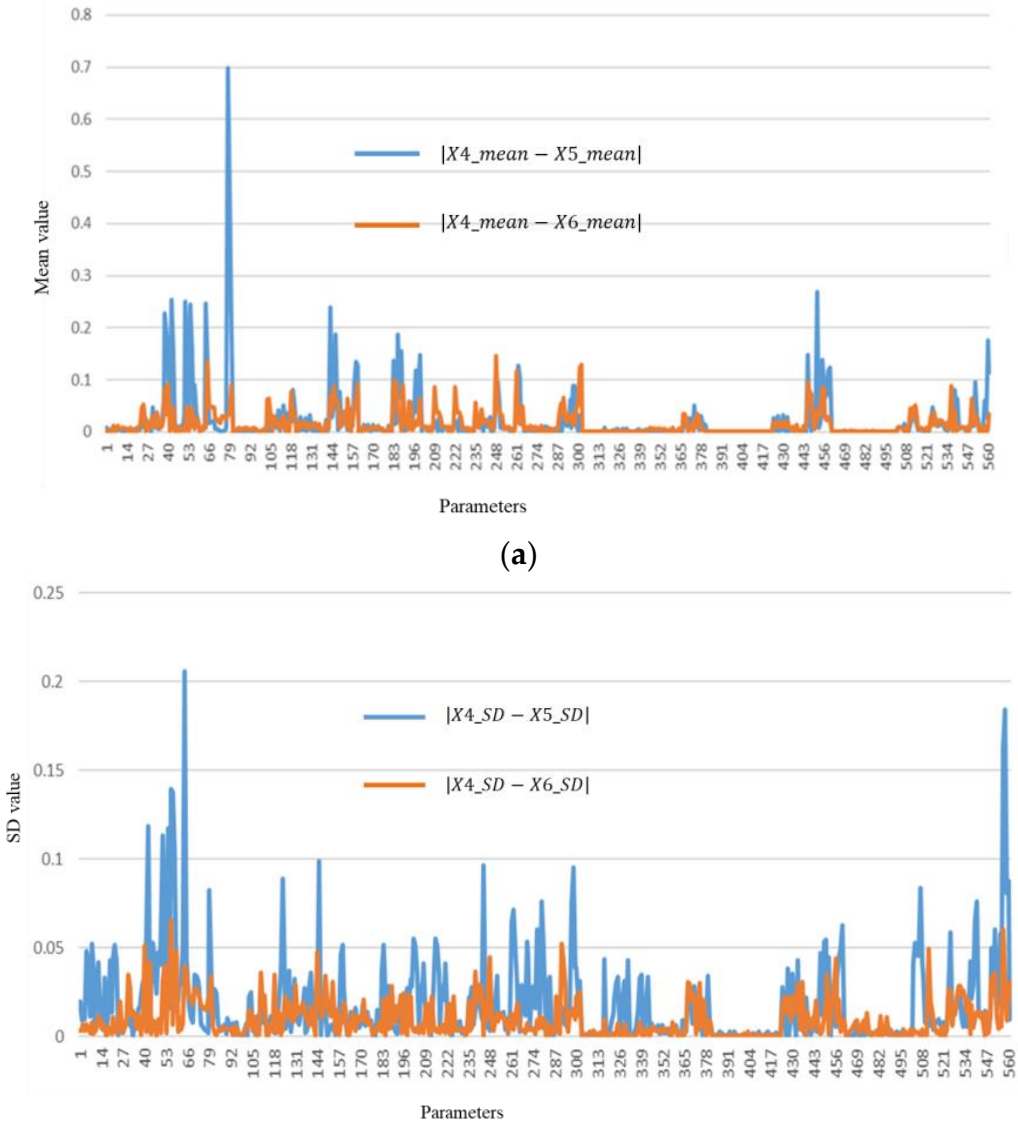


Figure 10. The feature differences of mean and standard division (SD) between training and testing data of standing and sitting activities introduced by 561 parameters, (**a**) mean differences of 561 parameters between the training set of standing activity (x4_mean) and the testing set of sitting activity (x5_mean) (blue line), and the mean differences of 561 parameters between the training set (x4_mean) and testing set (x6_mean) of the standing activity (orange line), (**b**) the SD differences of 561 parameters between the training set of standing activity (x4_SD) and the testing set of sitting activity (x5_SD) (blue line), and the SD differences of 561 parameters between the training set (x4_SD) and testing set (x6_SD) of standing activity (orange line).

The proposed ELA scheme fused with deep learning and machine learning methods. In order to compare with the previous studies, we did not study the generalization of ELA model with k-fold cross validation. In UCI-HAR dataset, the training and testing samples have been separated. All previous studies used the same training and testing samples to validate the performance of their proposed methods. Moreover, since a long time is required to implement a system for real time application, it is difficult to see how well the proposed model works in actual (real life) testing in the current stage. Moreover, when the smartphone is charging or not placed at the waist, the HAR would not be done. This is also the limitation of this approach. In the near future, we will design a wearable device that has the accelerometer and gyroscope. The parameters of the proposed ELA model are

embedded in a Movidius neural compute stick, like the Intel® Neural Compute Stick 2, to verify the HAR performance in the real scenario.

5. Conclusions

An ELA model consisting of the GRU, stacking CNN + GRU, and DNN was presented in this study. The input samples included the sensor data extracted from smartphones, and 561 parameters obtained from the sensor data. The outputs of the three models are fused for activity classification. The experimental results showed that the performance of the proposed ELA model was superior to the other existing schemes using deep learning methods in terms of the precision, recall, F1-score, and accuracy. Notably, we found that the standing and siting were two activities easily confused in the classification process. This investigation demonstrated that the use of 561 time-domain and frequency-domain parameters could significantly broaden the feature difference between training data of the standing activity and the testing data of the sitting activity, while decreasing the feature difference between training and testing data of the standing activity, thus effectively decreasing the recognition error rate for these two activities. In addition, since the data is recorded from a smartphone, our scheme could have the potential to be used for monitoring the daily activities of isolated people with the risk of COVID-19 in real time at home or any specified place.

Author Contributions: Conceptualization, T.-H.T. and S.-H.L.; Data curation, J.-Y.W.; Investigation, T.-H.T.; Methodology, J.-Y.W.; Project administration, T.-H.T.; Software, J.-Y.W.; Supervision, M.G.; Validation, M.G.; Writing original draft, S.-H.L.; Writing review and editing, S.-H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Technology, Taiwan, under grants MOST 109-2221-E-324-002-MY2 and MOST 109-2221-E-027-97.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lee, I.M.; Shiroma, E.J.; Lobelo, F.; Puska, P.; Blair, S.N.; Katzmarzyk, P.T. Effect of physical inactivity on major non-communicable diseases worldwide: An analysis of burden of disease and life expectancy. *Lancet* **2012**, *380*, 219–229. [CrossRef]
2. López-Bueno, R.; Calatayud, J.; Andersen, L.L.; Balsalobre-Fernandez, C.; Casana, J.; Casajus, J.A.; Sith, L.; Lopez-Sanchez, G.F. Immediate impact of the COVID-19 confinement on physical activity levels in Spanish adults. *Sustainability* **2020**, *12*, 5708. [CrossRef]
3. De Matos, D.G.; Aidar, F.J.; de Almeida-Neto, P.F.; Moreira, O.S.; de Souza, R.F.; Marcal, A.C.; Marcucci-Barbosa, L.S.; Martins Júnior, F.D.A.; Lobo, L.F.; dos Santos, J.L.; et al. The impact of measures recommended by the government to limit the spread of coronavirus (COVID-19) on physical activity Levels, quality of life, and mental health of Brazilians. *Sustainability* **2020**, *12*, 9072. [CrossRef]
4. Nyboe, L.; Lund, H. Low levels of physical activity in patients with severe mental illness. *Nord. J. Psychiatry* **2013**, *67*, 43–46. [CrossRef]
5. Oliveira, J.; Ribeiro, F.; Gomes, H. Effects of a home-based cardiac rehabilitation program on the physical activity levels of patients with coronary artery disease. *J. Cardiopulm. Rehabil. Prev.* **2008**, *28*, 392–396. [CrossRef]
6. Bulling, A.; Blanke, U.; Schiele, B. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput. Surv.* **2014**, *46*, 33. [CrossRef]
7. Liu, S.H.; Chang, Y.J. Using accelerometers for physical actions recognition by a neural fuzzy network. *Telemed. e-Health* **2009**, *15*, 867–876. [CrossRef] [PubMed]
8. Chen, L.; Hoey, J.; Nugent, C.D.; Cook, D.; Yu, Z. Sensor-based activity recognition. *IEEE Trans. Syst. Man. Cybern. C* **2012**, *42*, 790–808. [CrossRef]
9. Tan, T.H.; Hus, J.H.; Liu, S.H.; Huang, Y.F.; Gochoo, M. Using direct acyclic graphs to enhance skeleton-based action recognition with a linear-map convolution neural network. *Sensors* **2021**, *21*, 3112. [CrossRef] [PubMed]
10. Fernando, B.; Gavves, E.; Oramas, M.J.; Ghodrati, A.; Tuytelaars, T. Modeling video evolution for action recognition. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5378–5387.
11. Tufek, N.; Yalcin, M.; Altintas, M.; Kalaoglu, F.; Li, Y.; Bahadir, S.K. Human action recognition using deep learning methods on limited sensory data. *IEEE Sens. J.* **2020**, *20*, 3101–3112. [CrossRef]
12. Wong, W.Y.; Wong, M.S.; Lo, K.H. Clinical applications of sensors for human posture and movement analysis: A review. *Prosthet. Orthot. Int.* **2007**, *31*, 62–75. [CrossRef]

13. National Development Commission, 108 Survey Report on Digital Opportunities of People with Mobile Phones, Taiwan. 2019. Available online: https://ws.ndc.gov.tw/Download.ashx?u=LzAwMS9hZG1pbmlzdHJhdG9yLzEwL2NrZmlsZS9hZjg2Nzg1Ny01YWE0LTRjZTYtODQ3OS00NzVhMWY5NTkyOGMucGRm&n=6ZmE5Lu2OS0xMDjlubTmiYvmqZ%2Fml4%2FmlbjkvY3mqZ%2FmnIPoqr%2Fmn6XloLHlkYot5YWs5ZGK54mILnBkZg%3D%3D&icon=.pdf (accessed on 1 August 2020).
14. Nematallah, H.; Rajan, S.; Cretu, A. Logistic Model Tree for Human Activity Recognition Using Smartphone-Based Inertial Sensors. In Proceedings of the IEEE Sensors Conference, Montreal, QC, Canada, 27–30 October 2019; pp. 1–4.
15. Bao, J.; Ye, M.; Dou, Y. Mobile Phone-Based Internet of Things Human Action Recognition for E-Health. In Proceedings of the IEEE 13th International Conference on Signal Processing, Chengdu, China, 6–10 November 2016; pp. 957–962.
16. Cruciani, F.; Vafeiadis, A.; Nugent, C.; Cleland, I.; Mcullagh, P.; Votis, K.; Giakoumis, D.; Tzovaras, D.; Chen, L.; Hamzaoui, R. Comparing CNN and Human Crafted Features for Human Activity Recognition. In Proceedings of the IEEE Smart World, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, Leicester, UK, 19–23 August 2019; pp. 960–967.
17. Wu, W.; Zhang, Y. Activity Recognition from Mobile Phone Using Deep CNN. In Proceedings of the Chinese Control Conference, Guangzhou, China, 27–30 July 2019; pp. 7786–7790.
18. Wang, K.; He, J.; Zhang, L. Attention-based convolutional neural network for weakly labeled human activities' recognition with wearable sensors. *IEEE Sens. J.* **2019**, *19*, 7598–7604. [CrossRef]
19. He, J.; Zhang, Q.; Wang, L.; Pei, L. Weakly supervised human activity recognition from wearable sensors by recurrent attention learning. *IEEE Sens. J.* **2019**, *19*, 2287–2297. [CrossRef]
20. Deep, S.; Zheng, X. Hybrid Model Featuring CNN and LSTM Architecture for Human Activity Recognition on Smartphone Sensor Data. In Proceedings of the 20th International Conference on Parallel and Distributed Computing, Applications and Technologies, Gold Coast, Australia, 5–7 December 2019; pp. 259–264.
21. Xia, K.; Huang, J.; Wang, H. LSTM-CNN architecture for human activity recognition. *IEEE Access* **2020**, *8*, 56855–56866. [CrossRef]
22. Yu, T.; Chen, J.; Yan, N.; Liu, X. A Multi-Layer Parallel LSTM Network for Human Activity Recognition with Smartphone Sensors. In Proceedings of the 10th International Conference on Wireless Communications and Signal Processing, Hangzhou, China, 18–20 October 2018; pp. 1–6.
23. Sikder, N.; Chowdhury, M.S.; Arif, A.S.M.; Nahid, A. Human Activity Recognition Using Multichannel Convolutional Neural Network. In Proceedings of the 5th International Conference on Advances in Electrical Engineering, Dhaka, Bangladesh, 26–28 September 2019; pp. 560–565.
24. Ntisar, C.M.I.; Zhao, Q. A selective modular neural network framework. In Proceedings of the IEEE 10th International Conference on Awareness Science and Technology, Morioka, Japan, 23–25 October 2019; pp. 1–6.
25. Fawaz, H.I.; Lucas, B.; Forestier, G.; Pelletier, C.; Schmidt, D.F.; Weber, J.; Webb, G.I.; Idoumghar, L.; Muller, P.-A.; Petitjean, F. InceptionTime: Finding AlexNet for Time Series Classication. *Data Min. Knowl. Discov.* **2020**, *34*, 1936–1962. [CrossRef]
26. Karim, F.; Majumdar, S.; Darabi, H.; Harford, S. Multivariate LSTM-FCNs for time series classification. *Neural Netw.* **2019**, *116*, 237–245. [CrossRef]
27. Xiao, Z.; Xu, X.; Xing, H.; Luo, S.; Dai, P.; Zhan, D. RTFN: A robust temporal feature network for time series classification. *Inf. Sci.* **2021**, *571*, 65–86. [CrossRef]
28. Li, S.; Yao, Y.; Hu, J.; Liu, G.; Yao, X.; Hu, J. An ensemble stacked convolutional neural network model for environmental event sound recognition. *Appl. Sci.* **2018**, *8*, 1152. [CrossRef]
29. Batchuluun, G.; Yoon, H.S.; Kang, J.K.; Park, K.R. Gait-based human identification by combining shallow convolutional neural network-stacked long short-term memory and deep convolutional neural network. *IEEE Access* **2018**, *6*, 63164–63186. [CrossRef]
30. Du, H.; Jin, T.; He, Y.; Song, Y.; Dai, Y. Segmented convolutional gated recurrent neural networks for human activity recognition in ultra-wideband radar. *Neurocomputing* **2020**, *396*, 451–464. [CrossRef]
31. Breiman, L. Stacking. *Mach. Learn.* **1996**, *24*, 49–64. [CrossRef]
32. Wolpert, D. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
33. Opitz, D.; Shavlik, J. Actively searching for an effective neural-network ensemble. *Connect. Sci.* **1996**, *8*, 337–353. [CrossRef]
34. Hashem, S. Optimal linear combinations of neural networks. *Neural Netw.* **1997**, *10*, 599–614. [CrossRef]
35. Deng, L.; Platt, J.C. Ensemble deep learning for speech recognition. In Proceedings of the 15th Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014; pp. 1915–1919.
36. Xie, J.J.; Xu, B.; Chuang, Z. Horizontal and vertical ensemble with deep representation for classification. *arXiv* **2013**, arXiv:1306.2759. Available online: https://arxiv.org/abs/1306.2759 (accessed on 1 August 2020).
37. Dua, D.; Graff, C. *UCI Machine Learning Repository*; University of California, School of Information and Computer Sciences: Irvine, CA, USA, 2017; Available online: http://archive.ics.uci.edu/ml (accessed on 1 August 2020).
38. Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; Reyes-Ortiz, J.L. A public domain dataset for human activity recognition using smartphones. In Proceedings of the 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 24–26 April 2013; pp. 437–442.
39. Weiss, G. *UCI Machine Learning Repository*; University of California, School of Information and Computer Sciences: Irvine, CA, USA, 2019; Available online: https://archive.ics.uci.edu/ml/datasets/WISDM+Smartphone+and+Smartwatch+Activity+and+Biometrics+Dataset+# (accessed on 1 August 2020).

40. Roggen, D.; Calatroni, A.; Nguyen-Dinh, L.-V.; Chavarriaga, R.; Sagha, H.; Digumarti, S.T. (Eds.) *UCI Machine Learning Repository*; University of California, School of Information and Computer Sciences: Irvine, CA, USA, 2012. Available online: https://archive.ics.uci.edu/ml/datasets/opportunity+activity+recognition (accessed on 1 August 2020).
41. Jin, L.P.; Dong, J. Ensemble deep learning for Biomedical Time Series Classification. *Comput. Intell. Neurosci.* **2016**, *2016*, 6212684. [CrossRef]
42. Bengio, Y. Learning deep architectures for AI. *Found. Trends Mach. Learn.* **2009**, *2*, 1–55. [CrossRef]
43. Fumera, G.; Roli, F. A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *9*, 942–956. [CrossRef]
44. Rogova, G. Combining the results of several neural network classifiers. *Neural Netw.* **1994**, *7*, 777–781. [CrossRef]
45. Duin, R.P.W.; Tax, D.M.J. Experiments with classifier combining rules. In *Multiple Classifier Systems*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2000; Volume 1857, pp. 16–29.
46. Opitz, D.; Maclin, R. Popular ensemble methods: An empirical study. *J. Artif. Intell. Res.* **1999**, *11*, 169–198. [CrossRef]