Journal of Computers Vol. 34 No. 3, June 2023, pp. 221-235
doi: 10.53106/199115992023063403016

# Human Activity Recognition Based on CNN and LSTM

Xu-Nan Tan*

School of Exercise and Health, Shanghai University of Sport,
Shanghai 200438, China
tanxunan_bsu@163.com

*Received 30 June 2022; Revised 3 October 2022; Accepted 11 December 2022*

**Abstract.** Human activity recognition (HAR) based on wearable devices is an emerging field of great interest. HAR can provide additional information on a human subject's physical status. Utilising new technologies for HAR will become very meaningful with the development of deep learning. This study aims to mine deep learning models for HAR prediction with the highest accuracy on the basis of time-series data collected by mobile wearable devices. To this end, convolutional neural networks (CNN) and long short-term memory neural networks (LSTM) are combined in a deep network model to extract behavioural facts. The proposed CNN model contains two convolutional layers and a maximum pooling layer, and batch normalisation is added after each convolutional layer to improve convergence speed and avoid overfitting. This structure yields significant results in terms of performance. The model is evaluated on the MHEALTH dataset with a test set accuracy of 99.61% and can be used for the intelligent recognition of human activity. The results of this study show that the proposed model has better robustness and motion pattern detection capability compared to other models.

**Keywords:** human activity recognition, CNN, LSTM, deep learning, model integration

## 1 Introduction

Human behaviour analysis or activity recognition is an important part of the biometric field. Wearable device-based human activity recognition (HAR) plays an important role in people's daily life. Researchers can obtain a large amount of biological information about human activities from the raw data of mobile sensors. Thus, HAR has emerged as a major research area with the potential to assist multiple applications [1-2]. Society is increasingly becoming committed to improving the well-being of individuals through the use of various technologies. HAR facilitates the monitoring of living conditions and has the potential to improve the quality of life and health care of people with limited mobility or elderly people. Other features and uses of HAR continue to be developed within the field. From the perspective of daily life, the goals of HAR are to (1) create a predictive model that can distinguish between normal and abnormal behaviours and (2) provide health care providers with the necessary patient information feedback to identify specific behaviours. This approach can help monitors take the necessary action promptly to prevent and mitigate dangerous conditions. Currently, HAR is widely used in healthcare, behavioural judgment, gait analysis and motor status recognition [3].

HAR can be broadly classified into two categories: vision-based and motion sensor-based methods. Between these methods, vision-based is the more popular [4]. Although this activity monitoring method can provide high recognition accuracy, it is inapplicable in many special environments. In addition, the results of vision-based methods are easily affected by light variations and environmental occlusions. These factors also greatly limit the application of vision-based methods. An alternative approach to activity recognition is the use of motion sensors [5]. Information from different types of behaviours is typically collected from a set of dedicated wearable motion sensors, such as accelerometers and gyroscopes. Acceleration and angular velocity data vary in accordance with human motion. Thus, they can be used to infer human activity. These wearable sensors are inexpensive, miniaturised, portable and non-invasive and have low dependence on the surrounding environment. Activity recognition by mobile sensors has attracted considerable attention due to the portability and high acceptance in daily life of mobile sensors [6-8]. Zhu et al. [9] developed an intelligent remote assisted health care system that collects the daily activity characteristics of elderly people through multi-sensor fusion. The system can infer human intentions and health status from daily activity characteristics. Thirteen common daily activities were recognised. This system provides a new idea for remote assisted health care. Can et al. [10] collected data on the physical activity

* Corresponding Author

and physiological signals of subjects in different mental states through a wearable device and used several different machine learning algorithms to differentiate between the subjects' physiological states and stress situations. The models presented a high accuracy rate, and the study showed that the intelligent determination of human physiological states based on wearable devices is feasible. Wearable devices that measure gait in daily life are also one of the directions for the intelligent recognition of activity patterns. Given that human gait patterns may contain a portion of bodily information, including disease or other characteristics, monitoring daily gait parameters can provide additional information to clinicians or rehabilitation practitioners. Chen et al. [11] proposed a novel gait analysis method based on a wearable smart insole system that extracted 26 gait parameters on the basis of daily activities for the intelligent recognition of 'standing', 'walking', 'running' and other daily activities. The model can be analysed to infer whether the change in gait was caused by activity or by disease. Martinez-Hernandezet al. [12] proposed a Bayesian recognition method using wearable sensors to identify motion and gait phases simultaneously. The mobile sensors were attached to the thighs, calves and feet of the subjects, and data were collected and processed during uphill, downhill and flat motions. The results showed that the method can identify walking activity and gait phases with 99.87% and 99.20% accuracy, respectively. Therefore, this method can provide information on human walking and reliable help for the intelligent recognition of human walking activities. Human behaviour judgments based on wearable devices are also of great importance. Hashiguchi et al [13] developed a prediction model for judging the physical workload of construction workers. The study collected data on the workers' heart rate and physical activity from biosensors and accelerometer devices, relied on sensor data to predict workers' workload and judged workers' health risks with high practical value. The accuracy of the model judgment was 89.2%. The above study illustrated that activity pattern recognition based on wearable devices is feasible and that HAR has high practical value in many fields.

Two needs for developing smart HAR drive the work in this study: developing algorithms that can recognise multiple patterns and improving recognition accuracy. Thus, our algorithm is expected to address the need for multi-sensor activity recognition mainly by integrating existing algorithms. The contributions that we have made in this work can be summarised as follows:

(1) We propose a combined convolutional neural network (CNN)–long short-term memory neural network (LSTM) structure that can automatically extract spatial features and process human motion time series.

(2) On the basis of the periodic characteristics of human motion patterns, we introduce a CNN with moderate network depth as the basis for improvement. This CNN can capture local dependencies and extract effective spatial features.

(3) CNN followed by LSTM is introduced to capture the time sequence relationship.

(4) The proposed framework may be applied to human activities with different sensor patterns in different domains.

The remainder of this paper is organised as follows: Section 2 describes the related research in the field of HAR intelligent recognition. Section 3 describes the methodology, in which the CNN–LSTM model for predicting HAR is described in detail. Section 4 describes the dataset, data preprocessing and evaluation metrics. Section 5 presents the experimental results and their discussion. This paper is concluded in Section 6.

## 2 Related Works

In HAR, the main task is the quantitative analysis and automatic predictive classification of human motion, which confers machine learning and deep learning methods with a wide range of applications in this research area. Researchers have conducted extensive research on different sensor pattern recognition methods and have proposed numerous models and methods to study HAR [14]. The previous phase of research mainly used traditional machine learning methods for modelling predictions. KNN, HMM, SVM, RF and XGBoost are some of the most commonly used traditional algorithms [15]. Traditional methods extract a large number of features after pre-processing the raw data and selecting some key features that represent the essential differences between different activities. Iqbal et al. [16] used a traditional supervised machine learning algorithm to predict 13 outdoor sports collected by wearable devices, and the overall accuracy of the model was 90%. The study did not tune the parameters of the algorithm, which may have contributed to the relatively low accuracy. Lee and Cho [17] used a tri-axial accelerometer on a handheld smartphone to identify five activities with a hierarchical hidden Markov model. Motion data were collected from four participants. The results showed difficulty in distinguishing between upstairs and downstairs movements. Chereshnev et al. [18] proposed RapidHARe, which uses Bayesian networks to model distributions of raw data in a contextual window for the analysis of real-time human activity, and the

model achieved a predictive classification accuracy of 98.94%. Sun et al. [19] proposed an activity recognition method that uses accelerometers to identify seven physical activities based on six pocket locations. They extract features from data collected from seven subjects, including both time-domain and frequency-domain features. The overall F-score of the SVM classifier could reach 94.8% with a priori knowledge of known pocket locations. In most of the related work, filtering techniques such as Gaussian filters, low-pass filters, and Kalman filters are used to mitigate the effect of noise on the data. This is because the raw sensor data is always affected by noise, which makes it difficult to accurately measure and reflect the true motion variation of the sensor. Traditional methods extract a large number of features and select some main features that represent the essential differences between different activities after preprocessing the raw data [20]. Features extracted from the frequency domain, time domain, and quadrature distance are widely used. PCA or LDA is widely implemented to select the main features. In addition, normalization of the feature vector is a more common preprocessing method.

Traditional machine learning methods rely on heuristic manual feature extraction, which is often limited by the empirical knowledge of the researcher. In addition, it is difficult to distinguish between very similar activities due to different experimental bases. For these reasons, the performance of traditional pattern recognition algorithms is greatly limited in terms of classification accuracy and model generalization. Unlike traditional methods, deep learning can greatly reduce the workload of designing features and easily learn more meaningful high-level features by training end-to-end neural networks. Therefore, we believe that deep learning can perform HAR, which has been widely demonstrated in existing work [21]. In recent years, researchers have developed predictive models based on wearable sensor data mainly as deep learning models [21-22]. Hassan et al. [23] proposed a deep belief network, which firstly extracts features from the raw sensor data and performs principal component analysis (KPCA) and linear discriminant analysis (LDA), and finally feeds the processed features into a deep network model for training, achieving better results. Convolutional Neural Network (CNN) is a very widely used model in deep learning, CNN was proposed by Lecun of New York University in 1998 [24]. Much research has focused on the use of two-dimensional convolutional neural networks (2D CNNs), especially in image recognition. 1D CNNs are well suited for time-series analysis of sensor data (such as a gyroscope or accelerometer data); they are also suitable for signal data with fixed-length periods (such as audio signals), and can effectively capture the local dependencies and scale differences of the signal. transsexual. Wan et al. [25] proposed a CNN architecture to predict human activity patterns in real-time, and this study embedded a deep learning model architecture into smartphones to provide a basis for intelligent health monitoring. However, the data of this study came from the accelerometer of the smartphone, which has some limitations. Compared with wearable devices, smartphones acquire less motion information, which makes it difficult to achieve better classification and monitoring effects. Zhu et al. [26] proposed a CNN activity pattern recognition architecture consisting of two convolutional layers with input data from accelerometers, gyroscopes, and magnetometers. 235977 sensory samples from 100 subjects were collected, and the results showed that the model could achieve an accuracy of 96.1%. Tang et al. [27] developed a lightweight CNN for identifying HAR, and this study used CNN to process the temporal data, embedded in mobile sensors. A good result was obtained. Labati et al. [28] used CNN for ECG recognition. Both the ECG monitoring data and the accelerometer samples used in this study are time series data, indicating that CNN has better processing in 1-dimensional time series. Hochreiter and Schmidhuber (1997) proposed the LSTM unit, a deformation structure of RNN, and the model can handle time series data very well [29]. Thapa et al. [30] adapted the LSTM to a synchronization algorithm, enabling the model to employ multiple parallel input sequences to produce multiple parallel synchronized output sequences, achieving 97% accuracy in the public HAR dataset. Wang et al. [31] developed a hierarchical depth LSTM, which does not input the original sequence data for prediction and requires smoothing, noise reduction, and feature extraction of the data before input to the LSTM. the model proposed in this study achieved an accuracy of 99.15% in HAR. This non-end-to-end format achieves better results, but in a more complex prediction manner.

In this paper, we propose a deep neural network structure combining CNN and LSTM. The model can automatically extract features and perform activity pattern prediction classification, and the results show that the model has good accuracy, generalization ability, and convergence speed.

## 3 Proposed Model

In this paper, we propose a CNN-LSTM model architecture for the recognition of active patterns. This section describes the adopted CNN and LSTM network architecture in detail.

### 3.1 CNN

CNN is an extensively used model in deep learning and was proposed by Lecun of New York University in 1998 [24]. In 2015, CNN reached the human level of accuracy in image recognition tasks [32]. This model is essentially a multilayer perceptron, and its success is due to its use of local connectivity and weight sharing. CNN performs convolutional operations to obtain features from a large amount of input data. Similar to conventional neural networks, CNN consists of learnable weights and bias neurons. These components include the input, at least one convolutional and pooling layer and at least one fully connected layer. Numerous entities constitute the individual convolutional layers in a CNN. The convolution operation aims to extract different features from the dataset, and the pooling layer serves mainly to amplify the extracted information. CNN models have been widely used for image classification, speech recognition and time-series data processing [33].

1D CNNs are commonly used for sequential data processing [34], whereas 2D CNNs are usually used for image and text recognition [35]. Medical images and video data are mostly recognised by using three-dimensional CNNs [36]. In this work, a 1D CNN is used for the time series analysis of sensor data. A 1D CNN can well be applied to sensor data.

Fig. 1 represents the convolution and pooling of a 1D CNN. The left half of the figure indicates the input time series data, and the blue and red boxes indicate the different filters. The blue and red arrows represent the convolution along the time axis. After one layer of convolution to extract features, the feature dimension changes. This change is mainly related to the input time series data dimension, filter size, convolution step size and padding metrics. The data processed by the convolution layer are generally pooled again to further extract features and compress the data. The right half of the figure depicts the pooling operation, with yellow indicating maximum pooling.

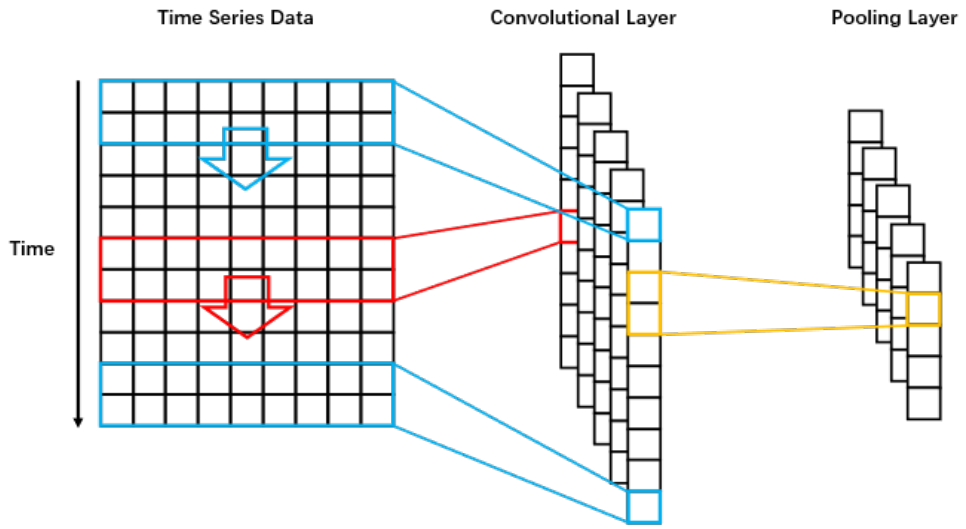


**Fig. 1.** Convolution and pooling process of 1D CNN

Fig. 2 represents the 1D convolutional process based on the MHEALTH dataset, wherein the input is a 128 × 12 matrix and the output is a 128 × 32 matrix after processing with 32 filters where kernel = 3 and padding is set to ‘same’.

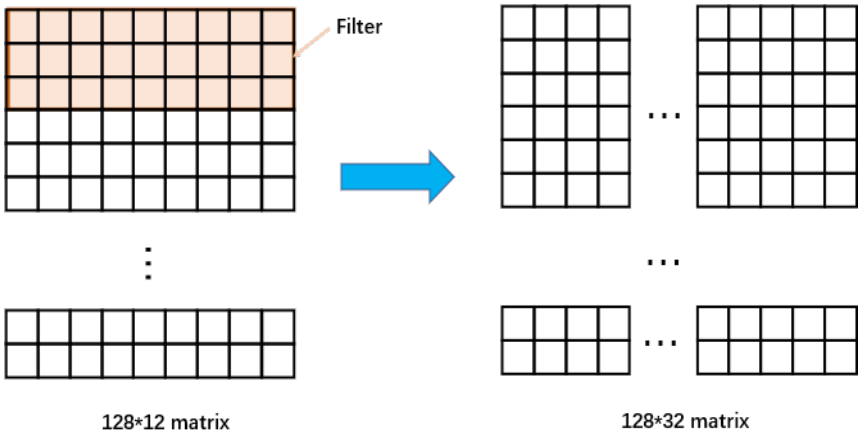


**Fig. 2.** The result of MHEALTH after a layer of convolution processing

### 3.2 LSTM

LSTM is a deformation structure of RNN [29]. To control the memory information of the time series data, a memory unit is added to the hidden layer. Messages are transferred among distinct units of the hidden layer via several controlled gates (forgetting gates, input gates, output gates), thus allowing to control the degrees of memory and forgetting of prior and current messages. As opposed to traditional RNNs, LSTMs have a long-term memory and can avoid gradient disappearance problems. In the LSTM, two gates are designed to handle the state of the memory cells, an erasing gate that determines how many cells can be stored to "remember" the prior state. There is also an input gate, which determines how many input moments can be saved to the cell state, and controls the "history". Another is the input gate, which determines how many current input moments can be stored in the cell state and controls the ratio of "history" information to "current" stimulus fusion.

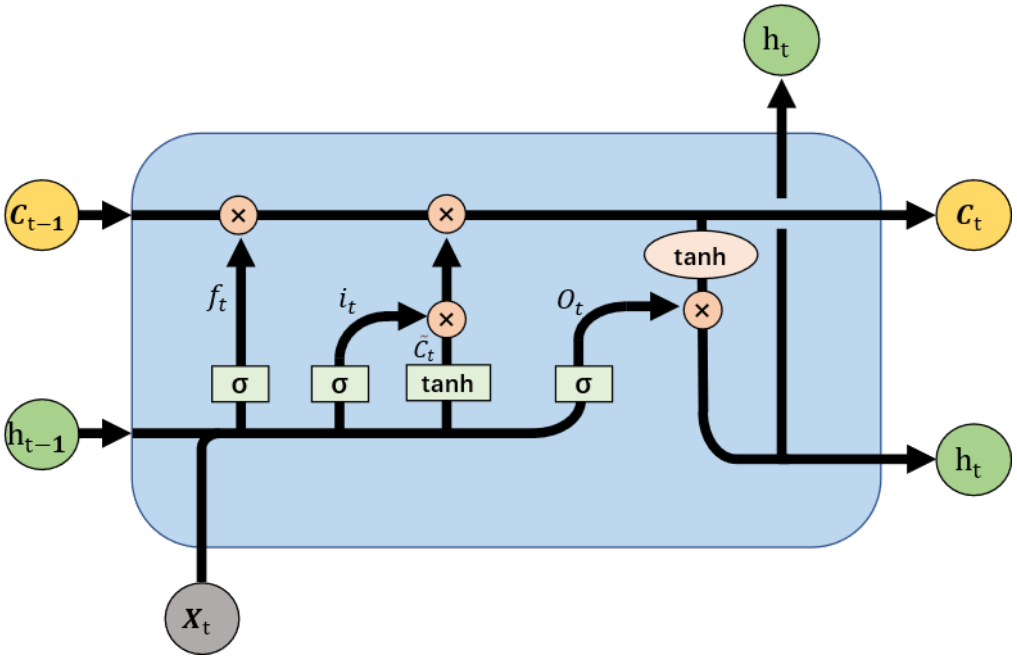


**Fig. 3.** LSTM network cell structure

The LSTM network structure is shown in Fig. 3. There are three inputs, $C_{t-1}$, the hidden state $h_{t-1}$ and the input vector $x_t$ at time $t$, while there are two outputs, $C_t$, and the hidden state $h_t$, where $h_t$ is also used as the output at time t. $f_t$ is the oblivion gate, which is implemented by the sigmoid function. The update gate has two parts, one is $\tilde{C}_t$, this part can bring new information input and the tanh function normalizes the input to between -1 and 1. The other update gate is $i_t$, which mainly keeps the new information.

The output gates are mainly $o_t$ and $h_t$, where $o_t$ uses a sigmoid function that indicates what is output, and $C_t$ is scaled by tanh and multiplied by $o_t$, which is the output at the current moment. The schematic network structure in Fig. 3 can be described by equations (1)-(6).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f). \tag{1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i). \tag{2}$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t]) + b_c). \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t. \tag{4}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o. \tag{5}$$

$$h_t = o_t * \tanh(C_t). \tag{6}$$

Where $W_f$, $W_i$, $W_c$, $W_o$, are weights, $b_f$, $b_i$, $b_c$, $b_o$ are bias terms. $t - 1$ is the previous moment, $t$ is the present moment. The input is $x$, the output is $h$, and $C$ is the cell status. $\sigma$ is a sigmoid function whose output is between 0 and 1.

### 3.3 CNN-LSTM

In this paper, we combine CNN with LSTM to construct a hybrid CNN-LSTM model to improve the accuracy of predicting HAR. Fig. 4 shows the network structure of the proposed hybrid CNN-LSTM model with multivariate time series data as input and multiclass motion patterns as output.

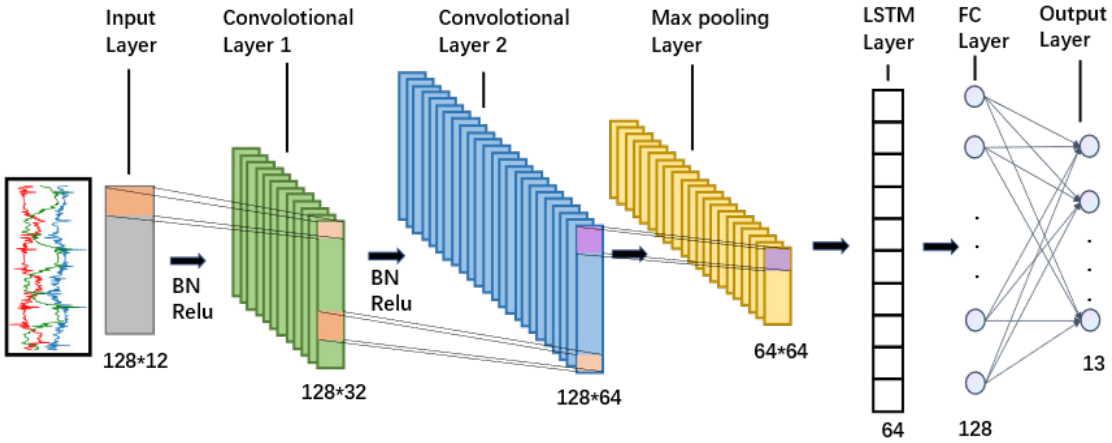


**Fig. 4.** Network structure of a CNN-LSTM hybrid model

The convolutional network is partially constructed with 2 convolutional layers and 1 pooling layer. Batch Normalization is added after the convolutional layer for processing. As the depth of the network increases, the distribution of eigenvalues in each layer gradually moves closer to the top and bottom of the output interval of the activation function (the saturation interval of the activation function). Batch-Normalization is a method to bring the layer eigenvalue distribution back to the standard normal distribution. The eigenvalues will fall in the range where the activation function is more sensitive to input. A small change in the input can lead to a large change in the loss function, making the gradient larger and avoiding gradient disappearance, and also accelerating convergence. Finally, two Dense fully connected network structures are added to the output of the LSTM, and the number of neurons in the overall model output is set to 13, corresponding to 13 different motion patterns. The activation function is the key to neural network to deal with nonlinear problems. The commonly used activation functions are the sigmoid function, tanh function (hyperbolic tangent function), Relu function, etc. In this study, Relu is selected as the activation function (Equation 7), which can well increase the nonlinear relationship between the layers of the network and reduce the interdependence of parameters, alleviating the occurrence of overfitting problems.

$$f(x) = \max(0, x). \tag{7}$$

## 4 Experimental Setup

### 4.1 Data Set Description

In the MHEALTH study, physical activity and vital sign recordings were collected from 10 different participants during 12 physical activity tasks [37] (Table 1). Data were acquired by using Shimmer2 wearable sensors attached to the participant's chest, right wrist and contralateral ankle with elastic bands (Fig. 5). Multiple sensors facilitated capturing motion in a variety of anatomical structures. In addition to acceleration, rotation rate and magnetic field direction, improved physical strength measurements were recorded. Furthermore, the chest sensor obtained two-lead ECG data, which were unrelated to the evolution of recognition paradigms. The cardiac information can be used for empirical cardiac monitoring, arrhythmia detection and exercise ECGs. A sampling rate of 50 Hz was used for all sensors because it was deemed suitable for detecting human behaviour. Data were collected outside of the laboratory, and no restrictions were imposed on how they could be collected. The only requirement was that participants perform the exercise to the best of their abilities.

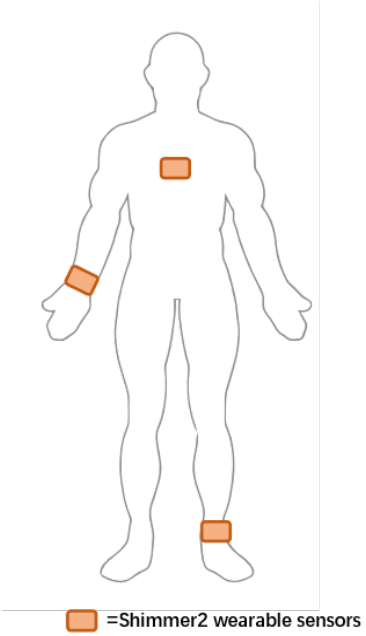


**Fig. 5.** Sensor placement on the human body for the MHEALTH dataset

**Table 1.** Description of data characteristics

| No | Activity | Number of repetitions or duration | Value counts |
|---|---|---|---|
| 0 | Nothing | - | 872550 |
| 1 | Stand still | 1min | 30720 |
| 2 | Sitting and relaxing | 1min | 30720 |
| 3 | Lying down | 1min | 30720 |
| 4 | Walking | 1min | 30720 |
| 5 | Climbing stairs | 1min | 30720 |
| 6 | Waist bends forward | 20 times | 28315 |
| 7 | Frontal elevation of arms | 20 times | 29441 |
| 8 | Knees bending (crouching) | 20 times | 29337 |
| 9 | Cycling | 1min | 30720 |
| 10 | Jogging | 1min | 30720 |
| 11 | Running | 1min | 30720 |
| 12 | Jump front and back | 20 times | 10342 |

### 4.2 Data Preprocessing

Given that MHEALTH data were obtained by sampling from wireless devices, problems, such as missing values, outliers and uneven data distribution, may exist. The preprocessing of the dataset is required to improve the model's performance.

(1) Find and process the missing values. Single-value interpolation is performed by using the K-nearest neighbour method. The 20 nearest samples with missing data are firstly determined on the basis of the Euclidean distance and weighted to average to estimate the missing data for that sample and filled in.

(2) Remove outliers. The accelerometer and gyroscope timing data are visualised, and the outliers are removed by firstly performing manual analysis and screening by observing the data trends. Data are also removed on the basis of other than 98% confidence level.

(3) Standardisation. Given that modelling using the original data may lead to bias, the dataset is normalised.

### 4.3 Implementation

The deep learning model training workstation proposed in this paper is trained using NVIDIA RTX3060 and the operating system is windows 10 professional 64-bit. Anaconda is used as the integrated learning development platform and the deep learning architecture is constructed using Keras.

### 4.4 Performance Evaluation Metrics

The CNN-LSTM model presented in this study is applied to the MHEALTH dataset, and the following parameters will be used for model evaluation for the multi-classification task. The model predicts positive samples, and TP indicates the actual positive samples. The FN indicates the actual positive samples, while the model predicts the negative samples. FP indicates actual negative samples, while the model predicts positive samples. The TN represents actual negative samples, and the model predicts negative samples (Fig. 6). TPR stands for true positive rate (Equation 8). FPR stands for false positive rate (Equation 9). Accuracy is the number of samples correctly predicted out of the total number of samples. TP and TN are the number of samples with correct predictions, and TP+FN+FP+TN is the number of samples overall. According to the definition, the formula of Accuracy can be determined (Equation 10). Accuracy is defined as the proportion of true positive samples among those predicted to be positive (TP+FP) (Equation 11). The recall rate is the percentage of all true positive classes (TP+FN) that were predicted to be positive (TP) (Equation 12). Here, the two metrics, precision, and recall are usually contrasted; Large datasets make it extremely difficult to obtain optimal results for both, and they are mutually constraining in large datasets, so they need to be taken into account together. In most cases, F1-Measure is used, which is the summed average of precision and recall, When the parameter $\alpha = 1$, this is the most common F1 metric. F1 combines the results of both precision and recall, and when F1 is high, it can indicate that the test is more effective (Equation 13).

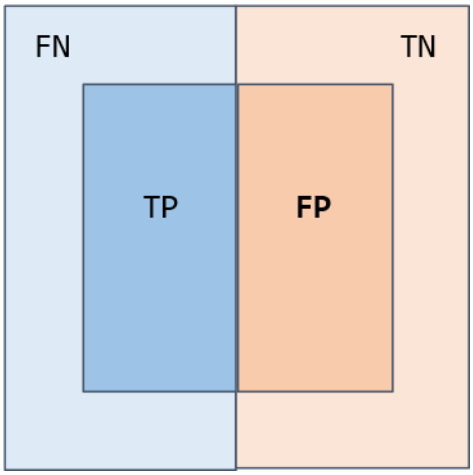


**Fig. 6.** Example of classification judgment

$$TPR = \frac{TP}{TP + FN}. \tag{8}$$

$$FPR = \frac{FP}{FP + TN}. \tag{9}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{10}$$

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{11}$$

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{12}$$

$$\text{F1} = \frac{2 * Precision * Recall}{Precision + Recall}. \tag{13}$$

Based on the confusion matrix, we can determine what classifications are correct and incorrect globally. Rows and columns are used to organize the actual labels and labels predicted by the classifier. It helps visualize how accurately the model can classify each category.

## 5 Results and Discussion

The proposed model was evaluated according to the following criteria: the MHEALTH public dataset was used in this study, and the hyperparameters selected for the model included optimizer, maximum epochs, Batch size, learning rate, and decay rate (Table 2). The number of epochs is chosen to be 200, and the model is early stopped at 72 epoch rounds to achieve convergence. The complete computational graph based on the Sequential model in Keras and the training parameters included in each layer of the network structure is shown in Table 3, with a total of 50,797 parameters, including 50,605 training parameters and 192 non-training parameters.

**Table 2.** Values used for the hyperparameters

| Hyperparameters | Selected parameters |
|---|---|
| Optimizer | SGD |
| Maximum epochs | 200 (earling stopping with 72) |
| Batch size | 128 |
| Learning rate | 0.1 |
| Decay rate | 0.002 |

**Table 3.** Number of parameters in the network structure

| Layer (type) | Output shape | Parameter |
|---|---|---|
| Conv1D | (None, 128, 32) | 1184 |
| Batch Normalization | (None, 128, 32) | 128 |
| Relu | (None, 128, 32) | 0 |
| Conv1D | (None, 128, 64) | 6208 |
| Batch Normalization | (None, 128, 64) | 256 |
| Relu | (None, 128, 64) | 0 |
| Max pooling | (None, 64, 64) | 0 |
| LSTM | (None, 64) | 33024 |
| Dense | (None, 128) | 8320 |
| Dense | (None, 13) | 1677 |

The training set accuracy of the model on the MHEALTH dataset was 99.96%, with a training set loss of 0.0044. the accuracy on the test set was 99.61%, with a test set loss of 0.0084. the corresponding accuracy, recall, and F1 index for the 12 activity models are shown in Table 4. among them, sitting and relaxing, climbing stairs, knees bending (crouching), and jump front and back had classification errors.

**Table 4.** Training results of models with different output categories

| No | Precision | Recall | F1 | Support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 92 |
| 1 | 0.99 | 1.00 | 1.00 | 122 |
| 2 | 1.00 | 0.99 | 1.00 | 124 |
| 3 | 1.00 | 1.00 | 1.00 | 122 |
| 4 | 1.00 | 1.00 | 1.00 | 120 |
| 5 | 0.99 | 0.98 | 0.98 | 84 |
| 6 | 1.00 | 1.00 | 1.00 | 105 |
| 7 | 1.00 | 1.00 | 1.00 | 112 |
| 8 | 0.98 | 0.99 | 0.98 | 117 |
| 9 | 1.00 | 1.00 | 1.00 | 120 |
| 10 | 1.00 | 1.00 | 1.00 | 90 |
| 11 | 1.00 | 1.00 | 1.00 | 52 |
| 12 | 1.00 | 0.96 | 0.98 | 26 |

During the model training process, we need to rely on the visualization of the validation set loss to determine the training effect of the model, whether there is overfitting and whether we need to stop training, etc. The training and validation set losses of the model training phase on the MHEALTH dataset are shown in Fig. 7. It can be observed that the training loss decreases as the number of epochs increases, indicating that the network is learning. When the epoch exceeds 20, it can be observed that the training loss and validation loss of the model decrease to approximately the same value and have reached convergence, and training is stopped to avoid overfitting. The accuracy of the training and validation sets is shown in Fig. 8. It can be observed that the training accuracy increases with epoch, indicating that the network continues to improve accuracy through learning. When the epoch exceeds 20, the training accuracy of the model remains the same, and the validation accuracy also reaches the same value as the training accuracy, indicating that the model reaches convergence, and finally the model stops training at 72 epoch.

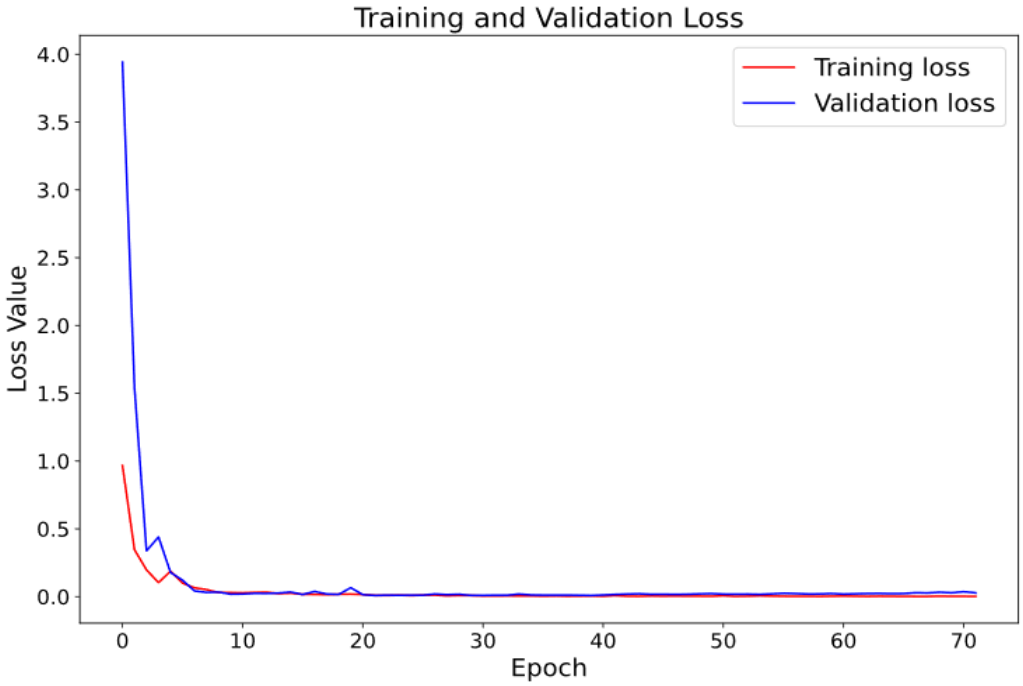


**Fig. 7.** On the MHEALTH data set, the accuracy of the training and validation sets during the model training stage

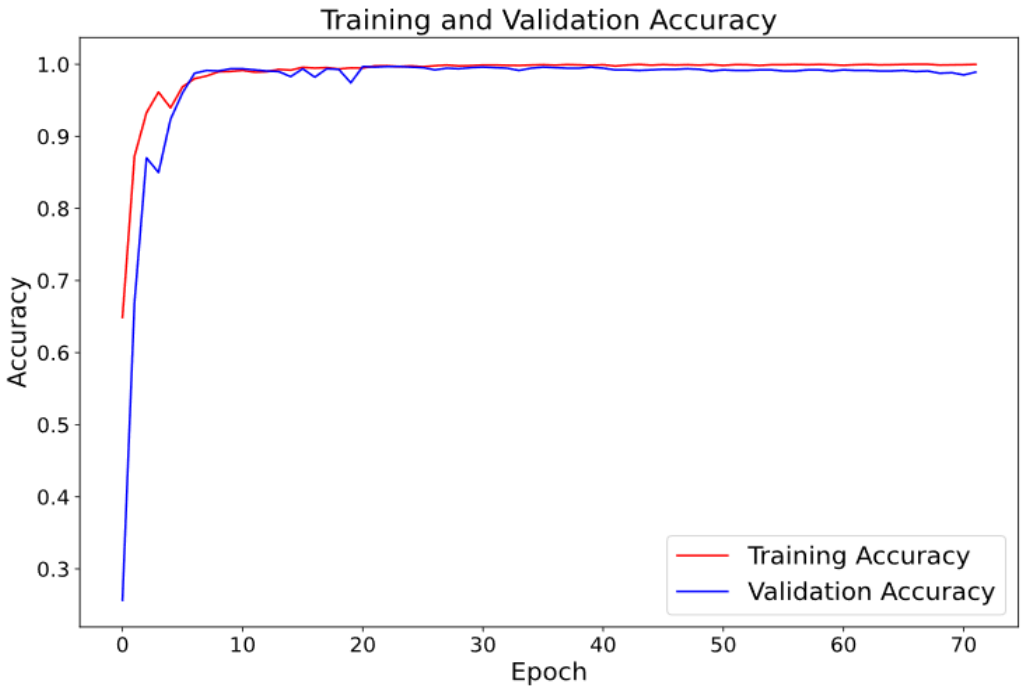


**Fig. 8.** Loss of training and validation sets during the training stage for the MHEALTH dataset

The confusion matrix is a visual measure of the accuracy of a classifier model, and the confusion matrix for the multiclassification results of this study is shown in Fig. 9. Columns in the matrix represent predicted categories, totals in columns indicate how much data each category contains. Each row represents a true attribution category, and the sum of each row represents the number of instances. The value in each column indicates the num-

ber of actual data predicted to be included in that category. Among them, sitting and relaxing, climbing stairs, knees bending (crouching), and jump front and back are misclassified. It can be found that these misclassified movement patterns have high similarities.

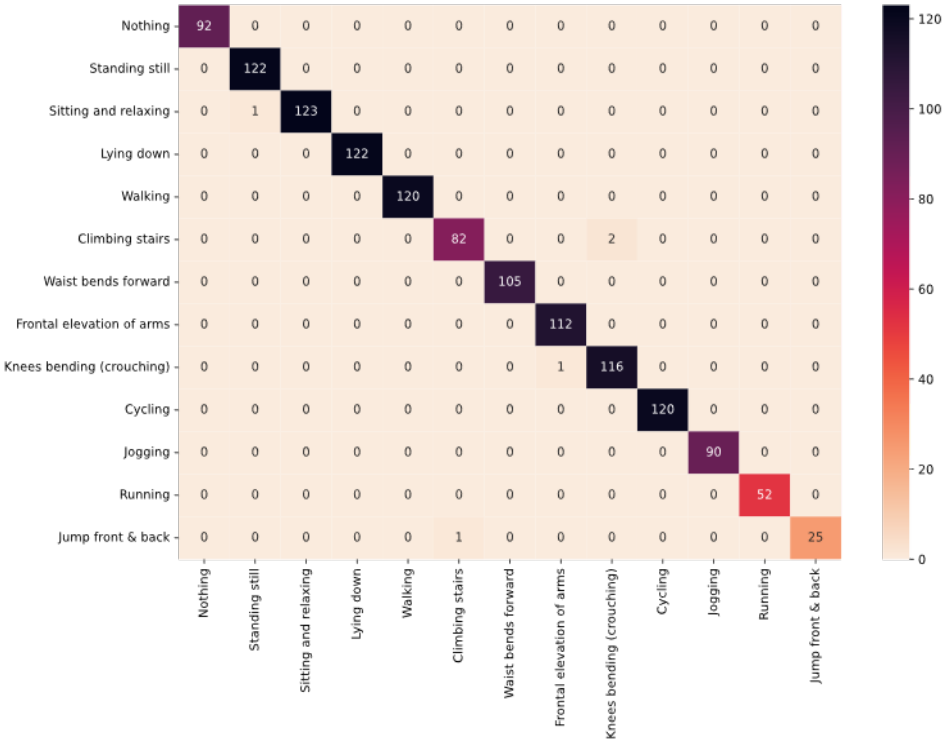


**Fig. 9.** Confusion matrix for MHEALTH data set

In Table 5, we compare our findings with those of other studies. In this study, Spatio-temporal features are extracted and processed differently from that presented in [38-40]. Three methods can be used to extract spatiotemporal features using CNN-LSTM; A CNN extracts spatial features and passes them to an LSTM that can extract temporal features, as shown in [40]. Two discrete branches are used in the second approach to extract spatial and temporal features in parallel. Branches are concatenated to facilitate the prediction of both features, as shown in [38]. In the third approach, features are input into an LSTM sub-model for extraction of temporal features before data is passed on to a CNN to extract spatial features [39]. Feature extraction principles determine each method's effectiveness. In this study, we use stepwise processing to capture spatial and temporal features separately and achieve over 99% high accuracy in the mHealth dataset.

**Table 5.** A comparison of the proposed model with other studies

| Data set | Reference | Model | Year | Accuracy (%) |
|---|---|---|---|---|
| MHEALTH | [32] | SVD-G-DNN | 2019 | 82.11 |
| | [33] | GADF | 2020 | 99.20 |
| | [34] | LSTM-CNN | 2020 | 95.78 |
| | [35] | CNN-GRU | 2021 | 99.35 |
| | Proposed | CNN-LSTM | 2022 | 99.61 |

In this work, a combination of CNN and LSTM is used to model the open data. The results show that the proposed deep learning network architecture can effectively predict the 12 motion patterns in the MHEALTH dataset. In consideration of the sampling frequency of 50 HZ and the small number of sample points, the input time-series points of the CNN architecture are set to 128 to improve the model performance because this length sequence contains one basic cycle period during the rapid repetitive motion of the human body. The training results also prove that this parameter setting is effective. Given the limitation of the sample size of the subjects

(10), the study did not acquire a large amount of time-series data. Therefore, CNN is chosen to extract the input features to reduce the number of parameters while also avoiding overfitting. The classification results are presented by using a confusion matrix to analyse the model classification results in detail. The misclassifications are mainly "sitting and relaxing" with "standing still", "climbing stairs" with "bending knees (squatting)", "bending knees (squatting)" with "frontal arm lift", and "jumping back and forth" with "climbing stairs". The above errors occurred mainly because the activity patterns of the limbs were similar, resulting in close sensor vibration signals and eventually misclassification. In the CNN part, only two layers of convolution and one layer of pooling are selected for feature extraction. In model training, the model reaches convergence at approximately 20 epochs, indicating that the parameter settings selected for the study satisfy the required conditions for training.

## 6 Conclusion

This study evaluates the proposed hybrid model by determining the most suitable model for human activity recognition. Thus, it contributes to the scientific field of machine learning related to HAR. We propose a deep learning model based on a combination of CNN and LSTM to predict activity types by analysing the MHEALTH dataset. The model has two key modules to improve performance: CNN for extracting features, and LSTM for processing time series. In the proposed architecture, the data collected by the mobile sensors are fed to CNN then to LSTM, thus enabling learning the temporal dynamics at different spatial scales based on the learning parameters of LSTM and improving accuracy. This deep learning network architecture can solve mobile sensor time series data well. The evaluation results show that the proposed model takes full advantage of CNN and LSTM to improve the accuracy of HAR and is relevant and practical. In addition, we explore the effects of some hyperparameters, such as input parameters, filter number, optimiser type and batch size, on model performance. Finally, the best hyperparameters of the final design are selected to train the model. In summary, the CNN-LSTM model exhibits consistently superior performance with better generalisation than the other methods proposed in the literature.

Our proposed hybrid model performs well in recognising the high-frequency repetitive activities of a single person. However, it may be unsuitable for multi-person activities and highly complex motion patterns. Future research plans include improving and updating the proposed method to enhance its recognition of complex activities and multi-person activities. In addition, we intend to investigate not only cutting-edge deep learning methods, including reinforcement learning, active learning and incremental learning, but also migration learning methods for such models in other areas and sectors of big data and cloud infrastructure.

## References

[1] A. Rajkomar, E. Oren, K. Chen, A.M. Dai, N. Hajaj, M. Hardt, P.J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G.E. Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S.L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N.H. Shah, A.J. Butte, M.D. Howell, C. Cui, G.S. Corrado, J. Dean, Scalable and accurate deep learning with electronic health records, NPJ Digital Medicine 1(1)(2018) 1-10.

[2] J. Wang, Y. Chen, S. Hao, X. Peng, L. Hu, Deep learning for sensor-based activity recognition: A survey, Pattern Recognition Letters 119(2019) 3-11.

[3] P. Picerno, M. Iosa, C. D'Souza, M.G. Benedetti, S. Paolucci, G. Morone, Wearable inertial sensors for human movement analysis: a five-year update, Expert Review of Medical Devices 18(sup1)(2021) 79-94.

[4] X. Yang, Y. Tian, Super normal vector for human activity recognition with depth cameras, IEEE Transactions on Pattern Analysis and Machine Intelligence 39(5)(2017) 1028-1039.

[5] M. Cornacchia, K. Ozcan, Y. Zheng, S. Velipasalar, A survey on activity detection and classification using wearable sensors, IEEE Sensors 17(2)(2017) 386-403.

[6] J. Wu, L. Chang, G. Yu, Effective data decision-making and transmission system based on mobile health for chronic disease management in the elderly, IEEE Systems Journal 15(4)(2021) 5537-5548.

[7] Y. Zhao, Q. Ni, R. Zhou, What factors influence the mobile health service adoption? A meta-analysis and the moderating role of age, International Journal of Information Management 43(2018) 342-350.

[8] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, P. Havinga, Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey, in: Proc. 23th International Conference on Architecture of Computing Systems, 2010.

[9] C. Zhu, W. Sheng, Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living, IEEE

Transactions on Systems, Man, and Cybernetics 41(3)(2011) 569-573.
[10] Y.S. Can, Stressed or just running? Differentiation of mental stress and physical activity by using machine learning, Turkish Journal of Electrical Engineering and Computer Sciences 30(1)(2022) 312-327.
[11] D. Chen, Y. Cai, X. Qian, R. Ansari, W. Xu, K.C. Chu, M.C. Huang, bring gait lab to everyday life: gait analysis in terms of activities of daily living, IEEE Internet of Things Journal 7(2)(2020) 1298-1312.
[12] U. Martinez-Hernandez, I. Mahmood, A.A. Dehghani-Sanij, Simultaneous bayesian recognition of locomotion and gait phases with wearable sensors, IEEE Sensors Journal 18(3)(2018) 1282-1290.
[13] N. Hashiguchi, K. Kodama, Y. Lim, C. Che, S. Kuroishi, Y. Miyazaki, T. Kobayashi, S. Kitahara, K. Tateyama, Practical judgment of workload based on physical activity, work conditions, and worker's age in construction site, Sensors 20(13) (2020) 3786.
[14] D. Thakur, S. Biswas, Smartphone based human activity monitoring and recognition using ML and DL: a comprehensive survey, Journal of Ambient Intelligence and Humanized Computing 11(11)(2020) 5433-5444.
[15] M.D.E. Beily, M.D. Badjowawo, D.O. Bekak, S. Dana, A sensor based on recognition activities using smartphone, in: 2016 International Seminar on Intelligent Technology and Its Applications (ISITIA), 2016.
[16] A. Iqbal, F. Ullah, H. Anwar, A. Ur Rehman, K. Shah, A. Baig, S. Ali, S. Yoo, K.S. Kwak, Wearable internet-of-things platform for human activity recognition and health care, International Journal of Distributed Sensor Networks 16(6) (2020) 1-14.
[17] Y.S. Lee, S.B. Cho, Activity recognition using hierarchical hidden markov models on a smartphone with 3D Accelerometer, in: Proc. International Conference on Hybrid Artificial Intelligence Systems, 2011.
[18] R. Chereshnev, A. Kertesz-Farkas, RapidHARe: A computationally inexpensive method for real-time human activity recognition from wearable sensors, Journal of Ambient Intelligence and Smart Environments 10(5)(2018) 377-391.
[19] L. Sun, D. Zhang, B. Li, B. Guo, S. Li, Activity recognition on an accelerometer embedded mobile phone with varying positions and orientations, in: Proc. Ubiquitous Intelligence and Computing, 2010.
[20] A. Bhavan, S. Aggarwal, Stacked generalization with wrapper-based feature selection for human activity recognition, in: Proc. 2018 IEEE Symposium Series on Computational Intelligence (SSCI), 2018.
[21] S. Zhang, Y. Li, S. Zhang, F. Shahabi, S. Xia, Y. Deng, N. Alshurafa, Deep learning in human activity recognition with wearable sensors: A review on advances, Sensors 22(4)(2022) 1476.
[22] H.F. Nweke, Y.W. Teh, M.A. Al-Garadi, U.R. Alo, Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges, Expert Systems with Applications 105 (2018) 233-261.
[23] M.M. Hassan, S. Huda, M.Z. Uddin, A. Almogren, M. Alrubaian, Human activity recognition from body sensor data using deep learning, Journal of Medical Systems 42(6)(2018) 1-8.
[24] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86(11)(1998) 2278-2324.
[25] S. Wan, L. Qi, X. Xu, C. Tong, Z. Gu, Deep learning models for real-time human activity recognition with smartphones, Mobile Networks and Applications 25(2)(2020) 743-755.
[26] R. Zhu, Z. Xiao, Y. Li, M. Yang, Y. Tan, L. Zhou, S. Lin, H. Wen, Efficient human activity recognition solving the confusing activities via deep ensemble learning, IEEE Access 7(2019) 75490-75499.
[27] Y. Tang, Q. Teng, L. Zhang, F. Min, J. He, Layer-wise training convolutional neural networks with smaller filters for human activity recognition using wearable sensors, IEEE Sensors Journal 21(1)(2021) 581-592.
[28] R.D. Labati, E. Munoz, V. Piuri, R. Sassi, F. Scotti, Deep-ECG: Convolutional neural networks for ECG biometric recognition, Pattern Recognition Letters 126(2019) 78-85.
[29] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: LSTM cells and network architectures, Neural Computation 31(7)(2019) 1235-1270.
[30] K. Thapa, Z.M.A. AI, S.-H. Yang, Adapted long short-term memory (LSTM) for concurrent human activity recognition, CMC-Computers Materials & Continua 69(2)(2021) 1653-1670.
[31] L. Wang, R. Liu, Human activity recognition based on wearable sensor using hierarchical deep LSTM networks, Circuits, Systems, and Signal Processing 39(2)(2020) 837-856.
[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[33] R. Yamashita, M. Nishio, R.K.G. Do, K. Togashi, Convolutional neural networks: An overview and application in radiology, Insights into Imaging 9(4)(2018) 611-629.
[34] S. Kiranyaz, T. Ince, M. Gabbouj, Real-time patient-specific ECG classification by 1-D convolutional neural networks, IEEE Transactions on Biomedical Engineering 63(3)(2016) 664-675.
[35] X. Yang, Y. Ye, X. Li, R.Y.K. Lau, X. Zhang, X. Huang, Hyperspectral image classification with deep learning models, IEEE Transactions on Geoscience and Remote Sensing 56(9)(2018) 5408-5423.
[36] G. Yao, T. Lei, J. Zhong, A review of convolutional-neural-network-based action recognition, Pattern Recognition Letters 118(2019) 14-22.
[37] O. Banos, R. Garcia, J.A. Holgado-Terriza, M. Damas, H. Pomares, I. Rojas, A. Saez, C. Villalonga, mHealthDroid: A novel framework for agile development of mobile health applications, in: Proc. International Workshop on Ambient Assisted Living, 2014.
[38] O. Nafea, W. Abdul, G. Muhammad, Multi-sensor human activity recognition using CNN and GRU, International

Journal of Multimedia Information Retrieval 11(2)(2022) 135-147.
[39] K. Xia, J. Huang, H. Wang, LSTM-CNN architecture for human activity recognition, IEEE Access 8(2020) 56855-56866.
[40] T. Li, M. Hua, X. Wu, A hybrid CNN-LSTM model for forecasting particulate matter (PM2.5), IEEE Access 8(2020) 26933-26940.