Computing
https://doi.org/10.1007/s00607-021-00928-8

REGULAR PAPER


# Multi-input CNN-GRU based human activity recognition using wearable sensors

**Nidhi Dua**[1] · **Shiva Nand Singh**[1] · **Vijay Bhaskar Semwal**[2]

Received: 4 September 2020 / Accepted: 17 February 2021


## Abstract

Human Activity Recognition (HAR) has attracted much attention from researchers in the recent past. The intensification of research into HAR lies in the motive to understand human behaviour and inherently anticipate human intentions. Human activity data obtained via wearable sensors like gyroscope and accelerometer is in the form of time series data, as each reading has a timestamp associated with it. For HAR, it is important to extract the relevant temporal features from raw sensor data. Most of the approaches for HAR involves a good amount of feature engineering and data pre-processing, which in turn requires domain expertise. Such approaches are time-consuming and are application-specific. In this work, a Deep Neural Network based model, which uses Convolutional Neural Network, and Gated Recurrent Unit is proposed as an end-to-end model performing automatic feature extraction and classification of the activities as well. The experiments in this work were carried out using the raw data obtained from wearable sensors with nominal pre-processing and don't involve any handcrafted feature extraction techniques. The accuracies obtained on UCI-HAR, WISDM, and PAMAP2 datasets are 96.20%, 97.21%, and 95.27% respectively. The results of the experiments establish that the proposed model achieved superior classification performance than other similar architectures.

**Keywords** Deep neural networks · Human activity recognition · CNN · Long short term memory (LSTM) · GRU

**Mathematics Subject Classification** Primary Classification: 68T · 68W · 15A · 62H

✉ Nidhi Dua
2016rsec001@nitjsr.ac.in

Shiva Nand Singh
snsingh.ece@nitjsr.ac.in

Vijay Bhaskar Semwal
vsemwal@gmail.com

[1] Department of ECE, NIT Jamshedpur, Jamshedpur, Jharkhand, India

[2] Department of CSE, MANIT Bhopal, Bhopal, M.P, India

Published online: 10 March 2021

## 1 Introduction

HAR is gaining much attention, as it can be of significant use in continuous monitoring of human behaviours in smart homes, intelligent surveillance systems, enhanced manufacturing, healthcare, rehabilitation, abnormal behaviour detection, gaming, personal fitness, etc. HAR frameworks provide ways to detect body movements, ambulatory and postural activities, and user actions by using data obtained by various sensors. These include mainly video-based sensors, wearable sensors, smartphone sensors, and wireless sensors. HAR can be categorized broadly into two categories. One is sensor-based activity recognition and the other is video-based. The video-based system uses video-based sensors like cameras, to capture video and images to recognize activities of daily life and human behaviours [1]. Sensor-based systems utilize ambient or wearable sensors to recognize human activities [2]. The data generated via smartphones and wearable sensors dominates the research of HAR because of their ubiquity, easy installation, and ease of use. Sensors are widely embedded in smart devices like mobile phones, watches, etc. They are capable of continuously logging information about human motion.

In the recent past, several Machine Learning (ML) techniques have been employed for HAR. Some of them include Support Vector Machine (SVM) [3], random forest [4], and k-Nearest Neighbours (kNN) [5]. For instance, the method for HAR proposed in [6] used features based on ensemble empirical mode decomposition, and a feature selection method based on game theory was introduced. The features and feature selection method were evaluated using SVM and kNN classifiers. In [7] an entropy-based hierarchical fusion model was proposed for HAR, which comprised of a sensor fusion layer and a classifier fusion layer. The entropy weight method was used to estimate the weights in the model. The use of ML techniques for activity recognition requires quite a lot of effort in data preparation, data pre-processing, and feature extraction which in turn requires domain expertise. Even though these techniques achieved decent recognition performances, they involve handcrafted feature learning methods. Moreover, features extracted using these techniques are application-specific and cannot be used for similar tasks [8].

Deep learning (DL), a branch of machine learning has achieved tremendous success in various areas like image segmentation [9], classification [10, 11], object detection [12], natural language processing [13], etc. DL has gained momentum in HAR. It overcomes the drawbacks associated with the handcrafted feature extraction based techniques, as it involves automatic detection of features, and hence involves lesser human intervention. Some of the deep learning approaches are CNNs [14], Deep Belief Networks (DBN) [15], Recurrent Neural Networks (RNN) [16], and deep feedforward neural networks.

DL techniques can extract and learn features from raw sensor signals and can make predictions efficiently. Many of the DL based approaches that have been applied for HAR include CNN, LSTM, CNN-LSTM, deep feed-forward networks, and more. In [17] a CNN model was proposed for HAR that used a

smartphone inertial sensor-based dataset. A deep CNN based architecture for multivariate time series data was proposed in [18]. They introduced a scheme for the transformation of input tensor and used convolution operation to capture the local interaction among the variables. MCNN [19], yet another deep learning approach that designed an end-to-end classifier for univariate time series data, made use of CNN. The author in [20] proposed a CNN based method for HAR from accelerometer data, where CNNs were useful for extracting local features, and the information related to the signal's global characteristics were derived by the use of statistical features. Another CNN based HAR classifier model was proposed in [21], where a multi-layer convolutional neural network was designed, which comprised of convolutional and pooling layers alternately, followed by a Fully Connected (FC) layer. The model performed activity recognition on the raw sensor data of the UCI-HAR [3] dataset.

Another extensively used member of the DL family is RNN. HAR is a classification problem dealing with time series data. For such a task, it becomes important to capture temporal dependency in the data. RNNs are naturally designed for this purpose. RNNs have been proposed for HAR in several research works. For instance, [22] proposed an LSTM based feature extractor for the classification of human activities and demonstrated the results on WISDM [23] dataset. Another LSTM based model was designed in [24], where the data obtained from the gyroscope and accelerometer were first normalized. The normalized data were then passed on to the stacked LSTM network, whose output was fed to a softmax layer. Another research [25] on HAR proposed a bi-directional LSTM based model for HAR using the data obtained by gyroscope and accelerometer sensors embedded in a mobile phone worn around the waist by the subjects. In [26] authors designed a bidirectional LSTM model for HAR using the time series data of the UCI-HAR dataset.

More recent works on HAR made use of a combination of CNN layers and RNN layers. For instance, the authors in [27] proposed an Activity Recognition (AR) system based on convolutional layers followed by recurrent dense layers. Authors in [28] exploited the advantages of both CNN and LSTM recurrent layers. They proposed a CNN-LSTM classifier for HAR that can take data from multimodal sensors and perform activity recognition without much data pre-processing. The experiments were performed on the data collected via gyroscopes and accelerometers individually and on their combinations as well. In [29] authors designed a deep network containing LSTM and convolutional layers and evaluated the model on three publicly available datasets. The model comprised of two LSTM layers followed by convolution layers, which were succeeded by a Global Average Pooling (GAP) layer, a Batch Normalization (BN) layer, and a softmax layer.

In recent literature, several neural network-based multi-headed models have been proposed. In [30], the authors proposed a multi-head CNN-RNN model for anomaly detection in an industrial scenario having multiple sensors. The multi-head CNN-RNN model had one CNN head dedicated to each sensor. The feature maps generated by the CNN heads were concatenated and passed to the RNN block which then finds out temporal patterns in the feature maps. Whereas a two-head model proposed in [31], used a fully convolutional block in parallel with an LSTM block. The LSTM block comprised an LSTM layer followed by a dropout layer. The features

extracted by both the blocks were concatenated and given to a softmax layer. In [32], the authors designed three types of multi-headed architectures (multi-headed LSTM, multi-headed ConvLSTM, and multi-headed CNN-LSTM) and an ensemble of all three of them. In these multi-headed models, each head handled one feature. These architectures were designed to forecast patients' average expenditure on two medications.

Some more miscellaneous approaches were also proposed for HAR. Extreme learning machines (ELM) [33, 34], another type of feed-forward network which doesn't use backpropagation, had also been used for HAR. The authors in [35] proposed an Extreme Learning Machine (ELM) based technique (TransM-RKELM) for HAR. The proposed approach enabled the classifier to be adapted for new sensor locations. The technique proposed in [36] was a non-DL based technique, where authors had proposed a genetic algorithm's variant method for feature extraction and introduced features reweighting during feature selection. In [37] a U-net based model was designed to perform HAR on sensor data. They had also used statistical features to analyze variations in signals.

Most of the state-of-the-art techniques for HAR rely on manual feature engineering [38] and hence require domain expertise. In contrast, the model proposed in this paper does not use any hand-engineered features. The DL based approaches for HAR face major challenges in feature extraction due to the issues like class imbalance in data or the noise present in the data collected via smartphone sensors. This paper presents a model for HAR using raw data obtained via wearable sensors like gyroscopes and accelerometers. In this work, a multi-input CNN-GRU classifier for HAR is proposed. The architecture of the proposed model involves the use of different sized convolutional filters (i.e. 3, 7, 11) and hence it is capable of capturing different temporal local dependencies in data. The proposed model leverages the advantages offered by CNN as well as RNN. The local features are extracted by the CNN and the long-term dependencies in data are well captured by the GRU layers. Hence, the model can capture the diversity of data. Three publicly accessible datasets viz. UCI-HAR [3], WISDM [23], and PAMAP2 [39] are applied to exhibit the superior performance of the proposed classifier.

### 1.1 The key contributions of the proposed work

1. The multi-input CNN-GRU model for HAR uses data from wearable sensors. It works on the raw data with nominal pre-processing and doesn't involve any handcrafted feature extraction technique.
2. The model is realized using CNN and GRU, and it thus exploits the advantages of both CNN and GRU. The local features are extracted by the CNN and the long-term dependencies in data are well captured by the GRU layers. Hence, the model can capture the diversity of data.
3. The model comprises a three head architecture and uses three different convolutional filter sizes. The model can capture local correlations at different lengths by making use of multiple filter sizes.

4. The model is validated through experiments performed using three different human activity datasets namely UCI-HAR, WISDM, and PAMAP2. The accuracy values obtained for UCI-HAR, WISDM, and PAMAP2 datasets are 96.20%, 97.21%, and 95.27% respectively.

### 1.2 Organization of paper

The rest of this paper is ordered as follows: Section II explains the methodology proposed for human activity recognition using multi-input CNN-GRU Model, whereas experiments and results are briefed in section III, and section IV presents the conclusions drawn.

## 2 Methodology

Human activity is characterized by the time series data [40, 41]. Wearable sensors (accelerometers, gyroscopes, etc.) worn at different positions on the human body can capture activity data that are sampled at regular intervals resulting in time series data [42, 43]. To capture the basic movements and the transitions in the activity, it is important to extract and learn the temporal features. HAR is a classification problem, where the raw time series data are the input, and the activity class is the output. It involves human walking [44] and locomotion behaviours [45]. DL based methods are capable of extracting features automatically from the raw input time series data, without any expert intervention. Therefore, using DNNs for feature extraction is useful to build end-to-end models that can handle all from feature extraction to classification. Datasets used for experiments for the proposed model, are UCI-HAR, WISDM, and PAMAP2 datasets.

### 2.1 Data segmentation

In the process of activity recognition, the first step is to create segments of the time series data. A sliding window is usually applied to segment the sensor data. For the WISDM dataset, the raw sensor data are segmented into samples (or frames) that have 128 timestamps per sample, and 3 features (corresponding to acceleration in x, y, and z-direction) associated with each timestamp. UCI-HAR dataset is already segmented where each sample contains 128 timestamps and has 9 features per timestamp. Thus, the input vector length for the UCI-HAR dataset is 128, and the number of channels is 9. For the WISDM dataset, the input vector is of length 128 and the number of channels is 3. The input vector length used for the PAMAP2 dataset is 128, and the number of channels is 52. This real value input vector represents one sample for one activity. With this input, we perform n channels (where $n=3$ for WISDM dataset, $n=9$ for UCI-HAR dataset, and $n=52$ for PAMAP2 dataset), 1dimensional (1D) convolution operation.

### 2.2 Feature extraction

The model proposed in this work uses the advantages offered by both CNN and RNN. CNNs and RNNs are capable of extracting features automatically. CNNs [46] are designed to process data that are in the form of multiple arrays. CNN comprises mainly convolution layers, pooling layers, and FC layers. A typical convolution operation on a time series data of dimensions 'n x v' is depicted in Fig. 1. Where n represents the length of the time series and 'v' represents the number of features. The input time series is convolved with filters of length 'k' and depth 'v'. Filter depth should be the same as that of the input to the convolution layer. Each unit of the convolution layer is arranged in a feature map.

CNN generates a feature map by multiplying a kernel (filter) with a local set of variables (i.e. receptive field) in the input array or feature maps of the previous convolutional layer. The size of the receptive field is equal to the size of the filter. Each unit of the receptive field is multiplied by the corresponding weight value in the filter bank. The obtained values are then summed up to get a single value in the feature map. This local weighted sum is moved through a non-linear function (generally

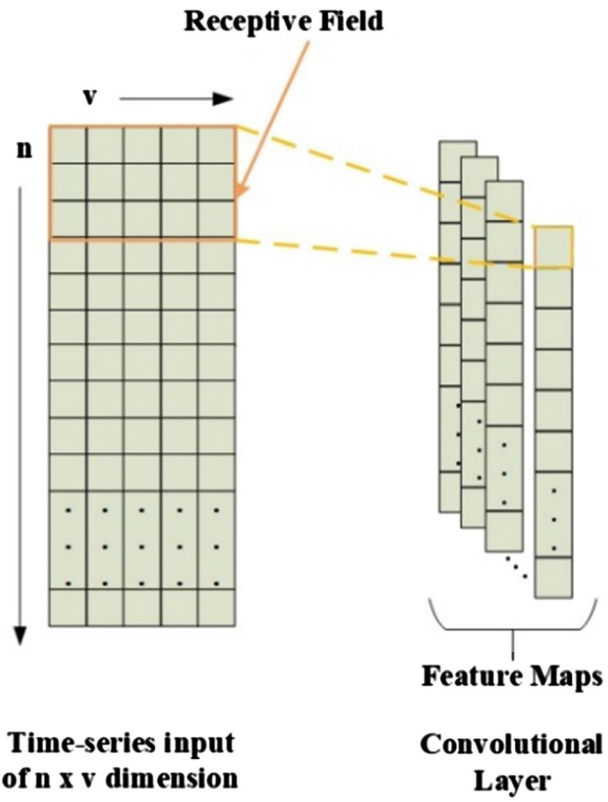


**Fig. 1** A typical convolutional operation on time series data

Rectified Linear Unit—ReLU). Each unit in the feature map shares the same kernel. The number of feature maps in a layer depends on the number of kernels used. In array data, the local set of values are highly correlated. The convolution layer is important to detect the local conjunctions in the preceding feature maps. Feature maps generated by the convolution layer are fed to the pooling layer. Each pooling unit computes the maximum of a local patch in the feature map. Pooling is performed on the local patches of the feature maps in the sliding window fashion. The pooling layer helps in dimensionality reduction and creating invariance to small distortions and shift. The output of the pooling layer is then flattened and forwarded to further layers.

Time series data have a strong 1D structure, this means the temporally nearby variables exhibit a strong correlation. These local correlations are the reason behind extracting and combining local features [46]. Hence, the extraction of local features is very important. The CNNs are capable of extracting local features by limiting the hidden units' receptive field to be local. Convolutional filters can capture short variations (for example a peak) in the time series signal. CNN considers each frame of sensor data as independent and extracts the feature for these isolated portions of data without considering the temporal context beyond the boundaries of the frame. Hence the features extracted by the CNN are short-term and local. Activity data is a long time series sequence data and to recognize the activity precisely it is very important to consider the inter-frame temporal context. Therefore, several approaches have used RNNs to learn temporal features in sequence data. Traditional RNNs were inefficient to capture long-term dependencies because the gradients are likely to vanish (or explode) when the sequences are long [47]. GRU was introduced in [48]. GRUs can overcome the vanishing and exploding gradients problem, which was seen in traditional RNNs. Therefore, GRUs allows a neural network to capture long-term dependencies in time series data [16]. It integrates gating units in the traditional recurrent unit to make it capture different time-scale dependencies adaptively. These gating units modulate the information flow inside the recurrent unit.

In the proposed model GRU layers have been used to capture the long-term dependencies in the activity data. A GRU unit consists of a reset gate and an update gate which enable the GRU units to remember the information at many earlier timestamps and predict the current state depending upon the information obtained from the previous states. These gates help GRU determine when and how much information from the past is to be sent to the future state. Hence, the GRU layers help the model to extract long-term temporal dependencies in activity data. Thus, the proposed model can capture the diversity of data in the course of training by using both CNN and GRU.

### 2.3 Model architecture

The block diagram of the proposed multi-input CNN-GRU classifier for HAR is shown in Fig. 2b. It comprises three heads, each fed with identical inputs. Each head is designed to have a Conv1D layer with 64 filters each. The filter size for head-1, head-2, and head-3 are 3, 7, and 11 respectively. Different filter sizes

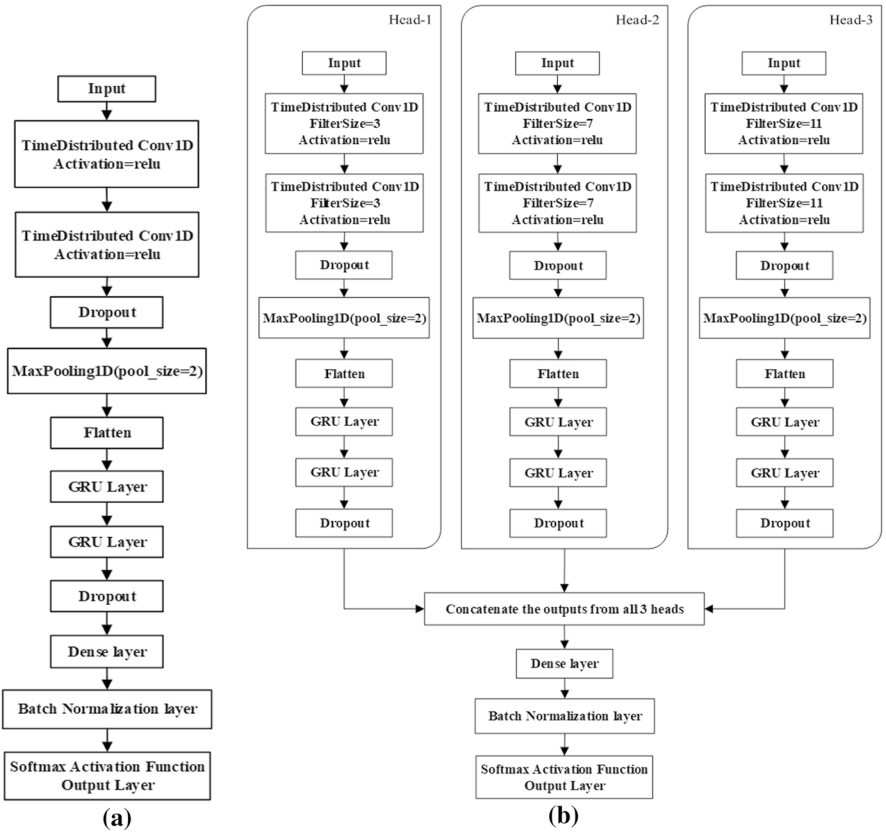


**Fig. 2** **a** Single-input CNN-GRU model for HAR. **b** Multi-input CNN-GRU model for HAR

capture different temporal local dependencies. Hence, the model can extract feature information on different scales. Filter sizes of 3, 7 and, 11 correspond to different timespans, and this is calculated as shown in Eq. 1.

$$T(F_s) = F_s/f \tag{1}$$

**Table 1** Timespans for different filter sizes used

| Dataset | Timespan (in seconds) for filter size '$F_s$' | | |
|---|---|---|---|
| | $F_s = 3$ | $F_s = 7$ | $F_s = 11$ |
| UCI-HAR | 0.06 | 0.14 | 0.22 |
| WISDM | 0.15 | 0.35 | 0.55 |
| PAMAP2 | 0.03 | 0.07 | 0.11 |

where, $T(F_s)$ is the timespan covered by filter size '$F_s$', and f = sampling frequency. Table 1 shows the values of timespans covered by different filter sizes, for the three datasets used in this work for experiments.

The 1D convolution operation is performed on the input vector. The activation function used for Conv1D layers is ReLU. Following the first Conv1D layer, there is a second Conv1D layer with a filter size of 3, 7, and 11 in head-1, head-2, and head-3 respectively. A dropout layer with 50% dropouts follows the second Conv1D layer. Next is a Max-Pooling 1D layer with pool size = 2. The Flatten layer flattens the feature vector from the Convolutional block to make it suitable to be fed to the following GRU layer. The GRU 1 layer contains 32 units, which return sequences to be fed to the GRU 2 layer. The outputs from all the three heads are concatenated and fed to the dense layer. A batch normalization layer follows the dense layer, which is followed by an FC output layer that uses the softmax activation function and classifies the input.

## 3 Experiments and results

The performance of the proposed model is gauged using UCI-HAR, WISDM, and PAMAP2 datasets. Keras framework with TensorFlow backend is used for the implementation of the classifier. The model is trained to minimize the cross-entropy loss. The values of the hyperparameters used are summarized in Table 2. Other hyperparameters used had default values. This section comprises the details of datasets, evaluation metrics employed, and the results obtained.

**Table 2** Values of hyperparameters used in this work

| Hyperparameters | Values used |
|---|---|
| Optimizer | Adam |
| Maximum epochs | 100 |
| Batch size | 400 |
| Learning rate (LR) | Initial LR = 0.001<br>Minimum LR = 0.0001 |
| Dropout rate | 50% |
| Pool size | 2 |
| Filter size | 3 (for Model-A)<br>7 (for Model-B)<br>11 ((for Model-C)<br>3, 7, 11 (for proposed model) |
| Input vector size | 128 |
| No. of input channels | 9 (for UCI-HAR)<br>3 (for WISDM)<br>52 (for PAMAP2) |

### 3.1 Datasets used

*UCI-HAR Dataset* [3]: It is a standard publicly available activity recognition dataset made available by the UCI repository. The UCI-HAR dataset was collected by performing experiments on 30 participants. Each participant wore a smartphone across the waist and performed six activities (laying, standing, sitting, walking downstairs, walking upstairs, and walking). The activities of the volunteers were recorded using the accelerometer and gyroscope sensors data signals, at a sampling rate of 50 Hz. The raw time series data contain nine features (i.e. body acceleration, total acceleration, and gyroscope signals in all three directions). The data signals were divided into 2.56 s windows with an overlap of 50 percent resulting in a total 10,299 number of samples. Activities of 21 volunteers are used for training and the activity data of 9 volunteers are used as a test dataset. The training dataset has 7352 samples and the test set has 2947 samples.

*WISDM Dataset* [23]: It was released by Fordham University's Wireless Sensor Data Mining lab. It is an activity recognition dataset collected for 36 users performing daily activities viz., walking, sitting, jogging, downstairs, upstairs, and standing. The data were captured using an accelerometer sensor embedded in smartphones, at a sampling rate of 20 Hz. For the proposed work, data of 29 users are used as a training set, and data of 7 users are used as test dataset; and values in the dataset are normalized to range between 0 to 1.

*PAMAP2 Dataset* [39]: This dataset contains 18 daily physical activities recorded which include 12 protocol activities (walking, running, vacuum cleaning, rope jumping, etc.) and 6 optional activities (watching TV, computer work, folding laundry, etc.). The activity data were captured for 9 subjects from all the sensors (a Heart rate monitor and 3 inertial measurement units) The data recorded comprised the measurements of gyroscopes, accelerometers, magnetometers, heart rate monitor, and temperature. The dataset has a total of 52 features and was captured at a sampling rate of 100 Hz. For the proposed work the data of users 5 and 6 are used for testing and the data from the remaining 7 users are used for training.

### 3.2 Performance measures

The performance measures used in this paper are F1-score, recall, precision, accuracy, and Confusion Matrix (CM). Accuracy is the ratio of samples that are classified correctly to the total number of samples. Activities can be classified as True Positives (TP) and True Negatives (TN) when they are classified correctly and as False Negatives (FN) and False Positives (FP) when they are wrongly classified. The performance measures can be defined in terms of TP, TN, FP, and FN.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{2}$$

Precision is the ratio of positives predicted correctly to the total number of samples classified as positives.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{3}$$

Recall is the ratio of positives predicted correctly to the actual number of positive samples.

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{4}$$

F1-score is the harmonic mean of the recall and precision, and it is generally applied in case when datasets are unbalanced.

$$\text{F1} - \text{score} = 2 * \text{Precision} * \text{Recall}/(\text{Precision} + \text{Recall}) \tag{5}$$

Confusion Matrices provide the overall misclassification rate. The true labels are represented by rows and the labels predicted by the classifier are represented in

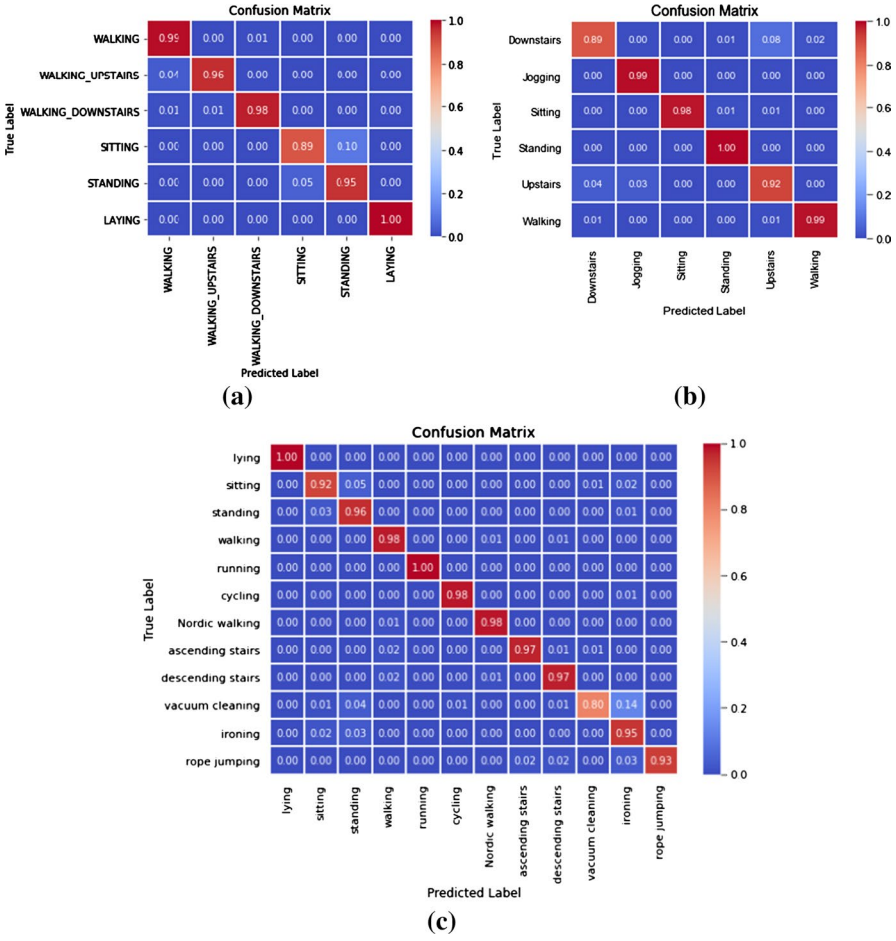


**Fig. 3** **a** CM obtained for UCI-HAR Dataset using multi-input CNN-GRU model. **b** CM obtained for WISDM Dataset using multi-input CNN-GRU model. **c** CM obtained for PAMAP2 Dataset using multi-input CNN-GRU model

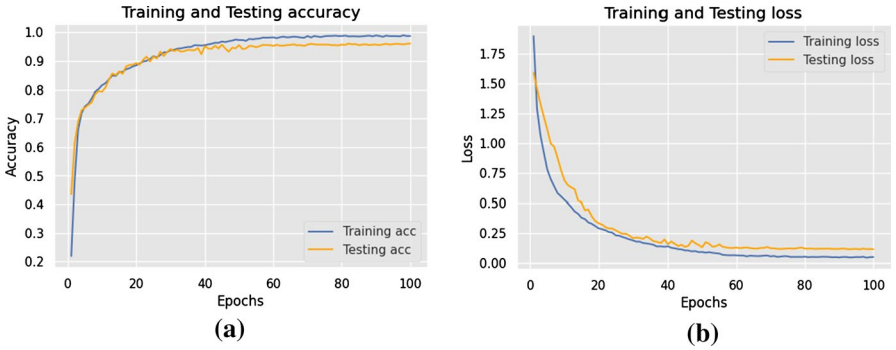


**Fig. 4** **a** Train and Test Accuracy plot obtained for UCI-HAR dataset using the proposed model. **b** Train and Test Loss plot obtained for UCI-HAR dataset using the proposed model

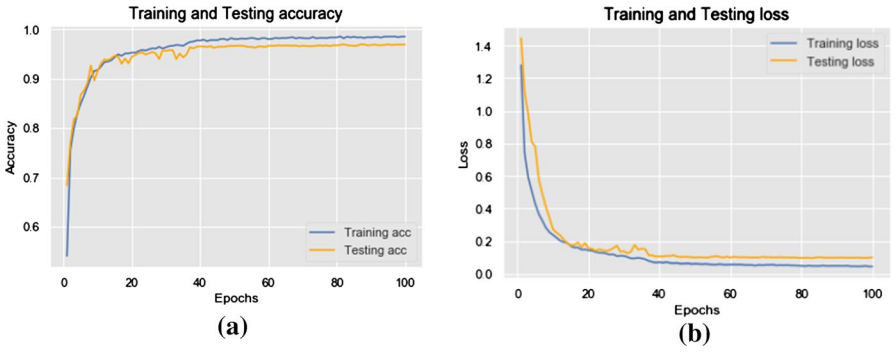


**Fig. 5** **a** Train and Test Accuracy plot obtained for WISDM dataset using the proposed model. **b** Train and Test Loss plot obtained for WISDM dataset using the proposed model

columns. It is an important measure to visualize the classification performance of the model.

### 3.3 Results

Experiments are performed using UCI-HAR, WISDM, and PAMAP2 datasets. Figure 3a depicts the CM obtained for the test data of the UCI-HAR dataset, and it shows that the model achieved good accuracies for 5 of the classes. Figure 3b shows the CM obtained for the test set of the WISDM dataset, and the model has achieved good classification accuracies for 4 out of 6 classes. The CM for test data of the PAMAP2 dataset is depicted in Fig. 3c and the model has obtained good classification accuracies for 11 classes out of 12 protocol activities classes used for this work. Figure 4 presents the accuracy and loss plots obtained using the proposed approach for the UCI-HAR dataset, whereas the accuracy and loss plots for the WISDM and PAMAP2 datasets are depicted in Figs. 5 and 6 respectively.

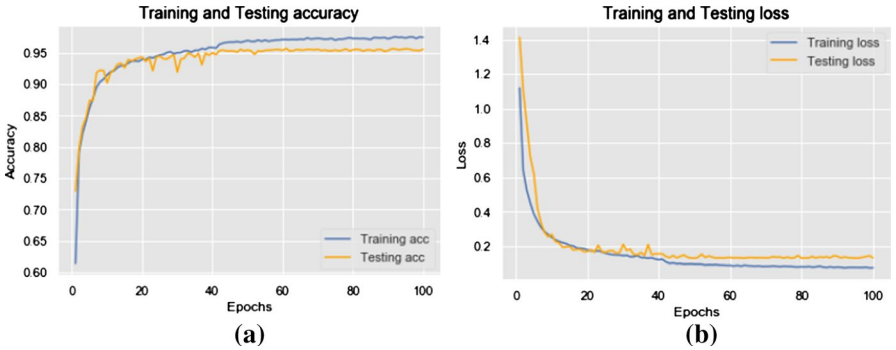


**Fig. 6** **a** Train and Test Accuracy plot obtained for PAMAP2 dataset using the proposed model. **b** Train and Test Loss plot obtained for PAMAP2 dataset using the proposed model

**Table 3** Performance comparison of the models for the UCI-HAR Dataset

| Models | Accuracy (%) | F1-score (%) |
|---|---|---|
| CNN [17] | 92.71 | 92.93 |
| ConvNet + MLP [21] | 94.79 | – |
| Stacked LSTM [24] | 93.13 | – |
| Bidir-LSTM [25] | 93.79 | – |
| Res-LSTM [26] | 91.6 | 91.5 |
| Res-Bidir-LSTM [26] | 93.6 | 93.5 |
| CNN-LSTM [28] | 92.13 | – |
| InnoHAR [27] | – | 94.5 |
| LSTM-CNN [29] | 95.78 | – |
| Multi-Input CNN-GRU (Proposed) | 96.20 | 96.19 |

**Table 4** Performance comparison of the models for the WISDM Dataset

| Models | Accuracy (%) | F1-score (%) |
|---|---|---|
| CNN [20] | 93.32 | – |
| LSTM[22] | 95.78 | 95.73 |
| LSTM–CNN [29] | 95.85 | – |
| Statistical features and reweighted genetic algorithm [36] | 94.02 | – |
| U-Net [37] | 96.4 | 96.5 |
| Multi-Input CNN-GRU (Proposed) | 97.21 | 97.22 |

The proposed model is compared with the models presented in [17, 20–22, 36, 37, 49–52]. Models are compared based on the accuracy and/or F1-score values. The results of the comparisons are presented in Tables 3, 4, 5 and 6.

The accuracy and F1-score values of different approaches obtained on the UCI-HAR dataset are presented in Table 3 and that for the WISDM and the PAMAP2

**Table 5** Performance comparison of the models for the PAMAP2 Dataset

| Models | Accuracy (%) | F1-score (%) |
|---|---|---|
| CNN [17] | 91.00 | 91.16 |
| LSTM + continuous temporal attention [52] | – | 89.96 |
| LSTM-S [49] | – | 88.2 |
| CNN [49] | – | 93.7 |
| CNN(Dfternet) [50] | – | 91.4 |
| InnoHAR [27] | | 93.5 |
| Cond Conv [51] | – | 94.01 |
| DNN [49] | – | 90.4 |
| Multi-input CNN-GRU (proposed) | 95.27 | 95.24 |

dataset are shown in Tables 4, and 5, respectively. Tables 3, 4, and 5 show that the proposed multi-input CNN-GRU model is performing better for HAR than other techniques compared.

To check the effectiveness of the multiple filter sizes used in the proposed model, the model is compared with three different single-input CNN-GRU models (containing single sized filters) namely Model-A, Model-B, and Model-C. The single-input CNN-GRU model is shown in Fig. 2a. Model-A, Model-B, and Model-C differ only in the size of the filter used in convolutional layers. The filter sizes used for Model-A, Model-B, and Model-C are 3, 7, and 11 respectively. All the models are evaluated on all the three datasets used in this work. From the performance results shown in Table 6, the proposed model is observed to outperform the single-input models using single-sized filters.

The proposed multi-input CNN-GRU classifier is also compared with a multi-input CNN-LSTM model and the results are depicted in Table 6. The multi-input CNN-LSTM model structure is similar to the multi-input CNN-GRU model proposed, except for the fact that it contains LSTM layers instead of GRU layers.

The same set of experiments are performed for the multi-input CNN-LSTM and multi-input CNN-GRU model and the results are mentioned in Table 6. Results imply that the use of GRU layers in the proposed model makes it superior when compared to its LSTM counterpart, hence making GRU an obvious choice for the proposed model to perform HAR on data obtained from wearable sensors.

Thus, the results presented in Tables 3, 4, 5 and 6 indicate that the multi-input CNN-GRU model has outperformed the existing approaches compared, and also that the multi-input CNN-GRU model works better than its multi-input CNN-LSTM counterpart.

From the results obtained for all three datasets (UCI-HAR, WISDM, and PAMAP2), it is observed that the proposed model can perform well on datasets containing various features and can recognize a varied range of activities. The model proposed can easily recognize simple activities (walking, jogging, laying, sitting, etc.) as well as complex activities like ironing, rope jumping, vacuum cleaning, etc. The model exhibits good generalization performance on all the three publicly available datasets.

**Table 6** Performance comparison of other CNN-RNN models with proposed multi-input CNN-GRU model

| Models | UCI-HAR dataset | | WISDM dataset | | PAMAP2 dataset | |
|---|---|---|---|---|---|---|
| | Accuracy (%) | F1-SCORE (%) | Accuracy (%) | F1-Score (%) | Accuracy (%) | F1-score (%) |
| Single-input CNN-GRU Model-A | 93.03 | 93.01 | 92.03 | 92.42 | 91.27 | 91.24 |
| Single-input CNN-GRU Model-B | 92.41 | 92.42 | 94.71 | 94.50 | 92.81 | 92.80 |
| Single-input CNN-GRU Model-C | 94.21 | 94.34 | 92.37 | 92.55 | 92.12 | 92.12 |
| Multi-input CNN-LSTM | 95.13 | 95.20 | 95.54 | 95.55 | 94.04 | 94.03 |
| Multi-input CNN-GRU (Proposed) | 96.20 | 96.19 | 97.21 | 97.22 | 95.27 | 95.24 |

## 4 Conclusion

A multi-input CNN-GRU model for HAR proposed in this work, leverage upon the robustness of CNNs in feature extraction, along with the advantages offered by GRUs for time series data classification. The proposed multi-input CNN-GRU model is both temporally and spatially deep, and it outperformed some of the existing DL models applied for the human activity recognition task. The ability of multi-input architecture to capture deep and shallow features helps to predict an activity with less error. Convolutional layers can efficiently capture the local features and the GRU layers are capable of handling long-term dependencies in the sequence data, thus making the model capture the diversity of data. The model can do feature extraction automatically on the raw dataset and hence doesn't rely on manually engineered features. The use of multiple sized convolutional kernels enables the model to capture local correlations at different lengths. The results of experiments carried out on UCI-HAR, WISDM, and PAMAP2 datasets indicate that the overall performance of the proposed model is superior when compared to some of the existing approaches for HAR. The results also show that the proposed multi-input CNN-GRU model performed better when compared to its multi-input CNN-LSTM counterpart. From the results, it is also observed that the model can perform well on datasets comprising various features and can recognize a varied range of activities.

**Funding** The author(s) also like to express thanks to SERB, DST, Govt. of India for funding the project under the schema of Early Career Award (ECR), with file no DST No: ECR/2018/000203 dated 04/06/2019.

**Declarations**

**Conflict of interest** We, the authors declare that we have no conflict of interest with any person or organization. This manuscript is based on original research findings done by the authors themselves.

**Ethical statement** All the ethical issues have been taken care of while writing the manuscript and we have complied with all the standards to the best of our knowledge.

## References

1. Khan ZA, Sohn W (2013) A hierarchical abnormal human activity recognition system based on R-transform and kernel discriminant analysis for elderly health care. Computing 95:109–127. https://doi.org/10.1007/s00607-012-0216-x
2. Cornacchia M, Ozcan K, Zheng Y, Velipasalar S (2016) A survey on activity detection and classification using wearable sensors. IEEE Sens J 17:386–403
3. Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz JL (2013) A public domain dataset for human activity recognition using smartphones. In: ESANN
4. Feng Z, Mo L, Li M (2015) A Random Forest-based ensemble method for activity recognition. In: 37th Annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 5074–5077
5. Jain A, Kanhangad V (2017) Human activity classification in smartphones using accelerometer and gyroscope sensors. IEEE Sens J 18(3):1169–1177

6. Wang Z, Wu D, Chen J, Ghoneim A, Hossain MA (2016) A triaxial accelerometer-based human activity recognition via EEMD-based features and game-theory-based feature selection. IEEE Sens J. 16(9):3198–3207
7. Guo M, Wang Z, Yang N, Li Z, An T (2018) A multisensor multiclassifier hierarchical fusion model based on entropy weight for human activity recognition using wearable inertial sensors. IEEE Trans Hum–Mach Syst. 49(1):105–111
8. Nweke HF, Teh YW, Al-Garadi MA, Alo UR (2018) Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. Expert Syst Appl 105:233–261
9. Zhang X, Zhang Y, Hu Q (2019) Deep learning based vein segmentation from susceptibility-weighted images. Computing 101(6):637–652. https://doi.org/10.1007/s00607-018-0677-7
10. Yu X, Dong H (2018) PTL-CFS based deep convolutional neural network model for remote sensing classification. Computing 100(8):773–785. https://doi.org/10.1007/s00607-018-0609-6
11. Semwal VB, Mondal K, Nandi GC (2017) Robust and accurate feature selection for humanoid push recovery and classification: deep learning approach. Neural Comput Appl 28:565–574
12. Ren S, He K, Girshick R, Sun J(2015) Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint http://arxiv.org/abs/1506.01497
13. Al-Makhadmeh Z, Tolba A (2020) Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. Computing 102(2):501–522. https://doi.org/10.1007/s00607-019-00745-0
14. Le Cun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Handwritten digit recognition with a back-propagation network. In: Proceedings of the 2nd international conference on neural information processing systems, pp 396–404
15. Lee H, Grosse R, Ranganath R, Ng AY (2009) Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th annual international conference on machine learning, pp 609–616
16. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint http://arxiv.org/abs/1412.3555
17. Wan S, Qi L, Xu X, Tong C, Gu Z (2020) Deep learning models for real-time human activity recognition with smartphones. Mobile Networks Appl 25(2):743–755
18. Liu CL, Hsaio WH, Tu YC (2018) Time series classification with multivariate convolutional neural network. IEEE Trans Ind Electron 66(6):4788–4797
19. Cui Z, Chen W, Chen Y (2016) Multi-scale convolutional neural networks for time series classification. arXiv preprint http://arxiv.org/abs/1603.06995
20. Ignatov A (2018) Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. Appl Soft Comput 62:915–922
21. Ronao CA, Cho SB (2016) Human activity recognition with smartphone sensors using deep learning neural networks. Expert Syst Appl 59:235–244
22. Chen Y, Zhong K, Zhang J, Sun Q, Zhao X (2016) LSTM networks for mobile human activity recognition. In: 2016 International conference on artificial intelligence: technologies and applications pp 50–53
23. Kwapisz JR, Weiss GM, Moore SA (2011) Activity recognition using cell phone accelerometers. ACM SIGKDD Explor Newsl 12(2):74–82
24. Ullah M, Ullah H, Khan SD, Cheikh FA (2019) Stacked lstm network for human activity recognition using smartphone data. In: 8th European workshop on visual information processing (EUVIP). IEEE, pp 175–180
25. Yu S, Qin L (2018) Human activity recognition with smartphone inertial sensors using bidir-lstm networks. In: 3rd international conference on mechanical, control and computer engineering (ICMCCE). IEEE, pp 219–224
26. Zhao Y, Yang R, Chevalier G, Xu X, Zhang Z (2018) Deep residual bidir-LSTM for human activity recognition using wearable sensors. Math Problems Eng. https://doi.org/10.1155/2018/7316954
27. Xu C, Chai D, He J, Zhang X, Duan S (2019) InnoHAR: A deep neural network for complex human activity recognition. IEEE Access 7:9893–9902
28. Mutegeki R, Han DS (2020) A CNN-LSTM approach to human activity recognition. In: International conference on artificial intelligence in information and communication (ICAIIC). IEEE, pp 362–366
29. Xia K, Huang J, Wang H (2020) LSTM-CNN architecture for human activity recognition. IEEE Access 8:56855–56866

30. Canizo M, Triguero I, Conde A, Onieva E (2019) Multi-head CNN–RNN for multi-time series anomaly detection: An industrial case study. Neurocomputing 363:246–260
31. Karim F, Majumdar S, Darabi H, Chen S (2017) LSTM fully convolutional networks for time series classification. IEEE access 6:1662–1669
32. Kaushik S, Choudhury A, Dasgupta N, Natarajan S, Pickett LA, Dutt V (2020) Ensemble of multi-headed machine learning architectures for time-series forecasting of healthcare expenditures. Applications of Machine Learning. Springer, Singapore, pp 199–216
33. Patil P, Kumar KS, Gaud N, Semwal VB (2019) Clinical human gait classification: extreme learning machine approach. In: 1st International conference on advances in science, engineering and robotics technology (ICASERT). IEEE, pp 1–6
34. Semwal VB, Gaud N, Nandi GC (2019) Human gait state prediction using cellular automata and classification using ELM. Machine intelligence and signal analysis. Springer, Singapore, pp 135–145
35. Wang Z, Wu D, Gravina R, Fortino G, Jiang Y, Tang K (2017) Kernel fusion based extreme learning machine for cross-location activity recognition. Information Fusion 37:1–9
36. Quaid MA, Jalal A (2020) Wearable sensors based human behavioral pattern recognition using statistical features and reweighted genetic algorithm. Multimed Tools Appl 79(9):6061–6083
37. Zhang Y, Zhang Z, Zhang Y, Bao J, Zhang Y, Deng H (2019) Human activity recognition based on motion sensor using u-net. IEEE Access 7:75213–75226
38. Lu J, Zheng X, Sheng M, Jin J, Yu S (2020) Efficient human activity recognition using a single wearable sensor. IEEE Internet Things J 7(11):11137–11146
39. Reiss A, Stricker D (2012) Introducing a new benchmarked dataset for activity monitoring. In: 16th International symposium on wearable computers. IEEE, pp 108–109
40. Ignatov AD, Strijov VV (2016) Human activity recognition using quasiperiodic time series collected from a single tri-axial accelerometer. Multimed Tools Appl 75(12):7257–7270
41. Sadouk L. CNN approaches for time series classification (2019) In: Time series analysis-data, methods, and applications. IntechOpen, pp 1–23
42. Semwal VB, Nandi GC (2016) Generation of joint trajectories using hybrid automate-based model: a rocking block-based approach. IEEE Sens J 16(14):5805–5816
43. Gupta A, Semwal VB (2020) Multiple task human gait analysis and identification: ensemble learning approach. Emotion and Information Processing. Springer, Cham, pp 185–197
44. Raj M, Semwal VB, Nandi GC (2018) Bidirectional association of joint angle trajectories for humanoid locomotion: the restricted Boltzmann machine approach. Neural Comput Appl 30(6):1747–1755
45. Nandi GC, Semwal VB, Raj M, Jindal A (2016) Modeling bipedal locomotion trajectories using hybrid automata. In: IEEE region 10 conference (TENCON). IEEE, pp 1013–1018
46. LeCun Y, Bengio Y (1995) Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks 3361(10)
47. Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Networks 5(2):157–166
48. Cho K, Van Merriënboer B, Bahdanau D, Bengio Y (2014) On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint http://arxiv.org/abs/1409.1259
49. Hammerla NY, Halloran S, Plötz T (2016) Deep, convolutional, and recurrent models for human activity recognition using wearables. arXiv preprint http://arxiv.org/abs/1604.08880
50. Yang Z, Raymond OI, Zhang C, Wan Y, Long J (2018) DFTerNet: Towards 2-bit dynamic fusion networks for accurate human activity recognition. IEEE Access 6:56750–56764
51. Cheng X, Zhang L, Tang Y, Liu Y, Wu H, He J (2020) Real-time human activity recognition using conditionally parametrized convolutions on mobile and wearable devices. arXiv preprint http://arxiv.org/abs/2006.03259
52. Zeng M, Gao H, Yu T, Mengshoel OJ, Langseth H, Lane I, Liu X (2018) Understanding and improving recurrent networks for human activity recognition by continuous attention. In: Proceedings of the 2018 ACM international symposium on wearable computers pp 56–63