The Visual Computer
https://doi.org/10.1007/s00371-021-02283-3

ORIGINAL ARTICLE

A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data

Sravan Kumar Challa[1] · Akhilesh Kumar[1] · Vijay Bhaskar Semwal[2]

Accepted: 5 August 2021

Abstract

Human activity recognition (HAR) has become a significant area of research in human behavior analysis, human–computer interaction, and pervasive computing. Recently, deep learning (DL)-based methods have been applied successfully to time-series data generated from smartphones and wearable sensors to predict various activities of humans. Even though DL-based approaches performed very well in activity recognition, they are still facing challenges in handling time series data. Several issues persist with time-series data, such as difficulties in feature extraction, heavily biased data, etc. Moreover, most of the HAR approaches rely on manual feature engineering. In this paper, to design a robust classification model for HAR using wearable sensor data, a hybrid of convolutional neural network (CNN) and bidirectional long short-term memory (BiLSTM) is used. The proposed multibranch CNN-BiLSTM network does automatic feature extraction from the raw sensor data with minimal data pre-processing. The use of CNN and BiLSTM makes the model capable of learning local features as well as long-term dependencies in sequential data. The different filter sizes used in the proposed model can capture various temporal local dependencies and thus helps to improve the feature extraction process. To evaluate the model performance, three benchmark datasets, i.e., WISDM, UCI-HAR, and PAMAP2, are utilized. The proposed model has achieved 96.05%, 96.37%, and 94.29% accuracies on WISDM, UCI-HAR, and PAMAP2 datasets, respectively. The obtained experimental results demonstrate that the proposed model outperforms the other compared approaches.

Keywords Bidirectional LSTM · HAR · Wearable sensor data · CNN · Deep neural networks

1 Introduction

In recent decades, HAR has gained substantial attention due to its broad range of applications in the fields of healthcare [1], intelligent surveillance systems [2], smart homes [3], rehabilitation [4], etc. The main aim of HAR is to infer the behavior of humans and inherently anticipate human intentions with the help of sensors. Different data modalities, such as acceleration, infrared, RGB, Wi-Fi signal, audio, depth, skeleton, etc., can be used to depict human actions. These modalities encode different sources of valuable yet distinct information and offer varying benefits depending upon the application environment [5]. Activity recognition can be performed using two systems which are sensor-based [6, 7] and video-based [8, 9]. Video-based systems use videos or cameras to continuously capture and identify a person's physical activity. Even though video-based systems have shown good performance in recognizing human activities, it has some limitations such as having privacy issues, being too expensive, and covering only limited areas [10]. Unlike video-based systems, sensor-based systems such as smartphones, smart devices, inertial measurement units (IMU), etc., engage the human body and move along with the body to record various activities.

Currently, the proliferation of smart devices and smartphones implanted with multi-sensor systems has allowed researchers to gather human physiological signals for tracking daily living activities. Mainly smartphones have access to a wide variety of sensors, namely accelerometers,

✉ Sravan Kumar Challa
2016rsec002@nitjsr.ac.in

Akhilesh Kumar
akumar.ece@nitjsr.ac.in

Vijay Bhaskar Semwal
vsemwal@gmail.com

[1] Department of Electronics and Communication Engineering, NIT Jamshedpur, Jamshedpur, Jharkhand, India

[2] Department of Computer Science Engineering, MANIT Bhopal, Bhopal, Madhya Pradesh, India

Published online: 20 August 2021

proximity, gyroscopes, magnetometers, etc., that can be used to infer details of various activities. Human activity data captured by smartphones or wearable sensors are in time-series format [11, 12]. A time-series is a sequential collection of data points that are typically calculated over successive time stamps. In the past few decades, researchers have adopted numerous machine learning (ML) methods to recognize different human activities using data obtained from smartphones or wearable sensors. For example, anguita et al. [13] adopted support vector machines to identify various human activities. The authors attached a smartphone to a person's waist and collected activity recognition data from sensors. In [14], the authors introduced an algorithm to segment the time-series data obtained from a smartphone sensor, i.e., accelerometer, and used the K-nearest neighbor method for activity recognition.

Apart from these, researchers have also been working on random undersampling, random oversampling, ensemble learning methods, etc. Among all, ensemble learning [15] has proven to be robust in handling imbalanced data. It generally combines the performance of several network classifiers to reduce statistical uncertainty (variance) and generalization error. The authors in [16] adopted an ensemble learning approach that uses a combination of multiple independent random classifiers based on various sensor feature sets to build a more fast, accurate, and stable classifier for HAR. In [17], the authors utilized multi-layer perceptron, J48 decision tree, and logistic regression to form an ensemble of classifiers. They applied the voting algorithm by merging the strength of these three models and conducted various experiments to demonstrate the efficiency of their ensemble.

Although the abovementioned ML techniques achieved decent performance in HAR, these techniques still have some limitations in using them for activity recognition. ML-based techniques require a significant amount of manual effort in data pre-processing, feature extraction, etc. Furthermore, the features derived using these methods are application-specific and cannot be used with other models [18].

Recently, DL-based algorithms have shown excellent results in various areas such as object detection [19–21], classification [22–24], natural language processing [25], etc. DL frameworks can automatically learn features from extensive data collection without needing any human effort and intervention. In [26], the authors collected motion data (movements of the arm) from four different human subjects using an accelerometer (tri-axial). They developed a CNN model to predict the various movements of an arm. The authors in [27] developed a smartphone sensor-based model for HAR that uses a CNN framework to classify multiple human activities. In [28], the authors proposed a DL-based framework for the real-time classification of human activities. The authors combined both CNN and statistical features that help to preserve information of time series data globally. The authors in [29] introduced an attention-based HAR method to deal with weekly labeled activity data collected from wearable sensors. The standard CNN pipeline and attention submodules were adopted to calculate the compatibility between local and global features obtained at the convolutional and fully connected layers, respectively. In [30], the authors proposed a novel local loss-based DL architecture for activity recognition by replacing the global loss of baseline CNN with local loss. This approach significantly reduced the memory requirements of the sensor-based activity recognition framework. Similarly, the authors in [31] designed a CNN-based HAR technique to capture local dependencies and scale invariance of the signals obtained using mobile sensors.

Recurrent neural networks (RNN), another prominent member of the DL family, have shown excellent results in handling sequential data. But traditional RNNs have failed to learn long-term dependencies in sequential data [32]. Long short-term memory (LSTM), a variant of the RNN, solved the vanishing gradient problem of traditional RNNs and proved to be good at handling long-term dependencies. In [33], the authors stacked five LSTM cells to design a robust classifier for identifying various human activities collected from smartphone sensor data. In [34], the authors proposed a Bidirectional LSTM-based network for HAR using smartphone sensor data. The authors in [35] proposed a deep residual bidirectional LSTM for HAR, which includes a bidirectional connection that concatenates both forward and backward states. This work was able to show improvements in the temporal and spatial dimensions and achieved an enhanced recognition rate. In [36], the authors combined both CNN and LSTM to do spatiotemporal sequence forecasting and achieved improvement in the predictive accuracy of human activities from the raw sensor data. An LSTM-CNN-based network was designed for HAR in [37], which consisted of LSTM layers followed by convolutional layers. The authors used a global average pooling layer to reduce the model parameters. A generalized hybrid CNN-LSTM-based architecture for HAR was proposed in [38], which is suitable for multimodal wearable sensors. With the advantages seen in employing ensemble learning in ML, researchers were motivated to adopt ensemble learning techniques in deep learning architectures. In [39], the authors developed a novel training algorithm for LSTM models and designed an ensemble classifier by combining multiple LSTM learners. In [40], the authors designed a hybrid model for HAR by connecting a fully convolutional and an LSTM block in parallel.

Some more DL-based approaches such as temporal convolution network (TCN), attention, and transformer networks have also been applied for HAR. The authors in [41] proposed a temporal convolution network-based architecture for identifying various human activities collected from smartphone sensors. This framework employs dilations and causal convolutions adapted for sequential data with its large receptive fields and temporality. In [42], the authors proposed an RNN-based attention network (RAN) consisting of CNN, LSTM, and attention modules. This model assisted in identifying and locating multiactivity types in sequential weakly labeled activity samples collected by wearable sensors. In [43], the authors have introduced a new benchmark dataset for human activity recognition captured by unmanned aerial vehicle (UAV). They also introduced a guided transformer I3D model for action recognition.

Even though DL-based architectures have demonstrated great success in HAR, feature extraction remains a significant challenge in this field due to erroneous/noisy data, class imbalance, etc. Moreover, most of the HAR approaches rely on manual feature engineering. In this paper, a DL-based model is proposed, which uses raw sensor data with minimal data pre-processing.

In this work, a multibranch CNN-BiLSTM model with different kernel sizes is proposed for HAR using wearable sensor data. The proposed model exploits the benefits of both CNNs and BiLSTMs. Convolutional neural networks are good at extracting local features, whereas BiLSTMs are capable of handling long-term dependencies. BiLSTMs, unlike baseline LSTMs (use only past information), make use of both past and future information when the complete sequence of time series data is available. This enables the network to make more accurate predictions due to the additional context provided. The model uses convolutional kernels of different sizes, which helps the model to capture various temporal local dependencies in sequential data. To evaluate the performance of the proposed model in identifying multiple human activities, datasets such as UCI-HAR, WISDM, and PAMAP2 are adopted.

The remaining data of this work are divided into sections as follows: The proposed methodology, including data pre-processing, the feature extraction process, and network architecture, is described in Sect. 2. Section 3 describes experimental results and datasets adopted for the proposed work. Finally, Sect. 4 concludes the proposed work.

1.1 The key contributions of this work

- A multibranch CNN-BiLSTM model is proposed for HAR based on wearable sensor data. This model operates directly on raw data with nominal pre-processing.
- The model exploits the capabilities of CNNs and BiLSTMs to capture local features as well as long-term dependencies in sequential data.
- Different filter sizes used in the proposed model help to capture various temporal local dependencies, making the feature extraction process more effective.
- The model is validated using three publicly accessible datasets, namely PAMAP2, WISDM, and UCI-HAR, and achieved better results than other DL techniques for HAR in the literature.

2 Methodology

2.1 Pre-Processing

Generally, signals generated by smartphones and wearable sensors are continuous and time series. The activity data captured by these sensors are in time-series format. The first step in the process of human activity identification is building segments from the sensor data. A sliding window technique is mainly employed to divide the sensor data and convert it into fixed-size windows. The proposed model uses a sliding window of size 128 (128 readings/window) with 50% overlapping for all adopted datasets. The datasets used in the proposed work are WISDM, UCI-HAR, and PAMAP2. The data are segmented into frames of size (128, N), where N is the number of features/ channels. The number of features for WISDM, UCI-HAR, and PAMAP2 is 3, 9, and 52, respectively.

2.2 Feature Extraction

Human activity data captured through wearable sensors and smartphones are a time series. The time-series data have a clear 1-D structure, which means that the variables in close proximity are strongly correlated over time [44]. Hence, it is imperative to extract these local features. CNN can extract local features of data by virtue of local receptive fields. Figure 1 shows the feature extraction process using CNN on time-series data. The input to the CNN is the time-series data of dimension (L x N), where L denotes the length of the time-series data, and N represents the number of features/

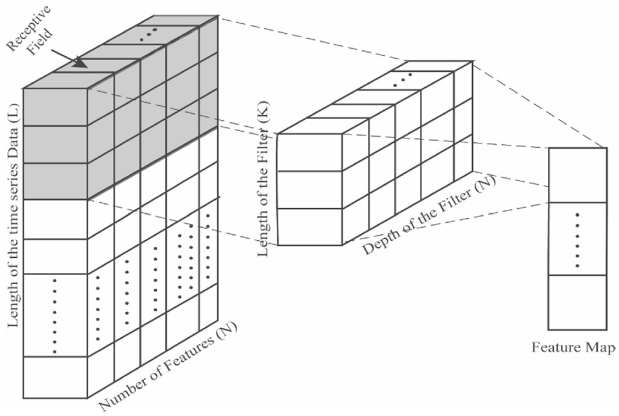


Fig. 1 The feature extraction process using CNN on time series data

channels. Convolution filters are used to extract the features from time-series data. The lengths of the filters used in the proposed model are 3, 7, 11, and the depth of each filter is equal to the number of features/channels of the input time-series data (i.e., N, as is shown in Fig. 1). Feature maps are generated using convolution operation, and the number of feature maps generated is equal to the number of filters used.

HAR data are segmented into multiple frames by using the sliding window technique. CNN considers each frame as an individual piece of data and does feature extraction for these frames in isolation without considering any temporal context outside the frame boundaries. While in HAR data, to identify the activities with precision, it is imperative to consider the temporal context among the data frames. Therefore, various techniques for HAR have applied RNNs to capture temporal features for the activity recognition task.

However, traditional RNN units could not capture long-term dependencies in the data as they suffer from the vanishing gradients problem [32]. LSTM [45] is one of the variations of the RNN that has become very powerful for classifying sequential data. They have solved the drawbacks of RNNs, such as vanishing gradient problems, short-term memory, among others. They are also good at dealing with long-term dependencies in sequential data. LSTM is a chain-like structure where the repeating module has various gates, unlike standard RNN (the repeating module is a simple tanh layer). The basic components of the single block LSTM, such as input gate, forget gate, output gate, and cell state, help the LSTM unit to adaptively capture various time scale dependencies. However, in scenarios when a complete sequence of information is available, LSTMs only use past information. Bidirectional LSTMs [46] are an enhancement of typical LSTMs that can increase the model's efficiency on sequence classification problems by using future information along with past information. The BiLSTMs simultaneously train two LSTMs, one in the forward direction of the input sequence and the other in the backward. This enables the network to make predictions more accurately due to the additional context provided. In the proposed model, the advantages of CNN and BiLSTM are exploited to capture local as well as long-term dependencies in sequential data.

2.3 Proposed Model

In this paper, a multibranch CNN-BiLSTM model is proposed for HAR and is implemented on the Keras API with TensorFlow backend. Figure 2 depicts the flow diagram of the proposed model that contains three branches (branch-I, branch-II, and branch-III). All the branches are similar in structure except for the convolutional filter sizes used. Filter sizes 3, 7, and 11 are used in branch-I, branch-II, and branch-III, respectively. Different filter sizes used in the proposed model can capture various temporal local dependencies. Initially, the same input is fed to all three branches, as shown in Fig. 2. Each of the branches consists of two convolutional 1D (Conv1D) layers to extract local features from the sequential data.

The TimeDistributed Conv1D layer in each branch accepts input data in four dimensions containing samples, n_seq, n_steps, and channels. Here, samples represent the number of windows in the adopted datasets, n_seq represents the number of sub-sequences per window, n_steps represent the length of each sub-sequence, and channels represent the number of input features. For all the datasets used in the proposed work, the length of each sample (i.e., 128) is further divided into four sub-sequences of length 32 each. ReLU activation function is used for introducing nonlinearities in Conv1D layers. A dropout layer (0.5) is introduced after the second Conv1D layer to prevent the network from overfitting. After that, a 1D MaxPooling layer of pool size two is adopted for dimensionality reduction. The output from all three branches is concatenated and fed into a BiLSTM layer. The BiLSTM layer helps the model to captures the long-term dependencies in the input sequence. The output from the last BiLSTM layer is given to a dense layer, as depicted in Fig. 2. The Batch normalization [47] process is employed after the dense layer to normalize its output to zero mean and unit standard deviation. Finally, the classification layer containing the softmax activation function takes the normalized output from the last dense layer and predicts the input class. The proposed architecture is trained to reduce categorical cross-entropy loss using Adam optimizer.

3 Experiments and results

3.1 Dataset description

3.1.1 UCI-HAR dataset

This dataset was developed by Anguita et al. [11]. A total of 30 volunteers were participated to perform different daily activities, i.e., sitting, lying, walking, standing, walking upstairs, and downstairs. The authors captured linear acceleration (tri-axial) and angular velocities (tri-axial) with the help of an accelerometer and a gyroscope implanted on a smartphone. A total of 9 features were recorded at a sampling rate of 50 Hz, and these data are sampled into fixed-width sliding windows (128 readings/window) with a 50 percent overlap. This dataset has already been segmented by user id and contains a total of 10,299 samples.

3.1.2 WISDM dataset

This dataset was developed by Kwapisz et al. [12] by recording different basic activities of human's namely

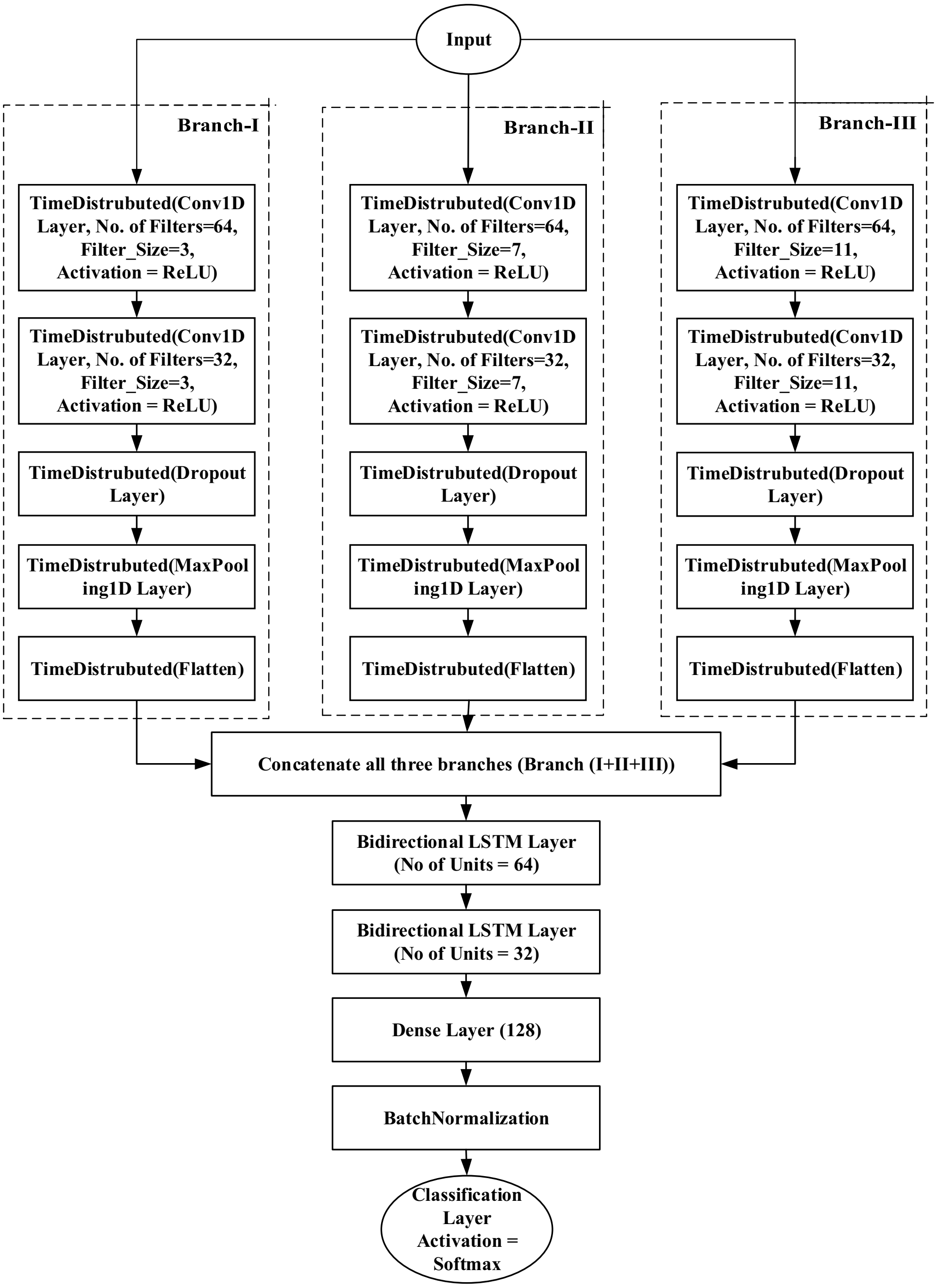

Input
Branch-I
Branch-II
Branch-III
TimeDistrubuted(Conv1D Layer, No. of Filters=64, Filter_Size=3, Activation = ReLU)
TimeDistrubuted(Conv1D Layer, No. of Filters=32, Filter_Size=3, Activation = ReLU)
TimeDistrubuted(Conv1D Layer, No. of Filters=64, Filter_Size=7, Activation = ReLU)
TimeDistrubuted(Conv1D Layer, No. of Filters=32, Filter_Size=7, Activation = ReLU)
TimeDistrubuted(Conv1D Layer, No. of Filters=64, Filter_Size=11, Activation = ReLU)
TimeDistrubuted(Conv1D Layer, No. of Filters=32, Filter_Size=11, Activation = ReLU)
TimeDistrubuted(Dropout Layer)
TimeDistrubuted(MaxPooling1D Layer)
TimeDistrubuted(Flatten)
Concatenate all three branches (Branch (I+II+III))
Bidirectional LSTM Layer (No of Units = 64)
Bidirectional LSTM Layer (No of Units = 32)
Dense Layer (128)
BatchNormalization
Classification Layer Activation = Softmax

Fig. 2 Flow diagram of the proposed multibranch CNN-BiLSTM architecture

sitting, walking, jogging, standing, walking downstairs, and upstairs using an accelerometer embedded on their smartphones at a sampling rate of 20 Hz. A total of 36 volunteers performed the abovementioned six activities, and acceleration was measured in the x, y, and z (3 features) directions for each activity. To perform experiments in this study, all of the values in this dataset are converted into a range of − 1 to 1 with a mean zero and standard deviation of one. Then, these raw sensor data are sampled into fixed-width sliding windows (128 readings/ window) with a 50 percent overlap.

3.1.3 PAMAP2 dataset

A. Reiss and D. Stricker introduced a physical activity monitoring database [48] that consists of various activities performed by nine subjects. All the subjects were instructed to perform 18 multiple activities (including six optional activities) such as rope jumping, running, playing soccer, etc. The activity data are recorded with the help of three IMU sensors mounted at various locations of the human body. A total of 52 features were captured at a sampling rate of 100 Hz. In this work, out of 18 daily activities recorded, 12 activities

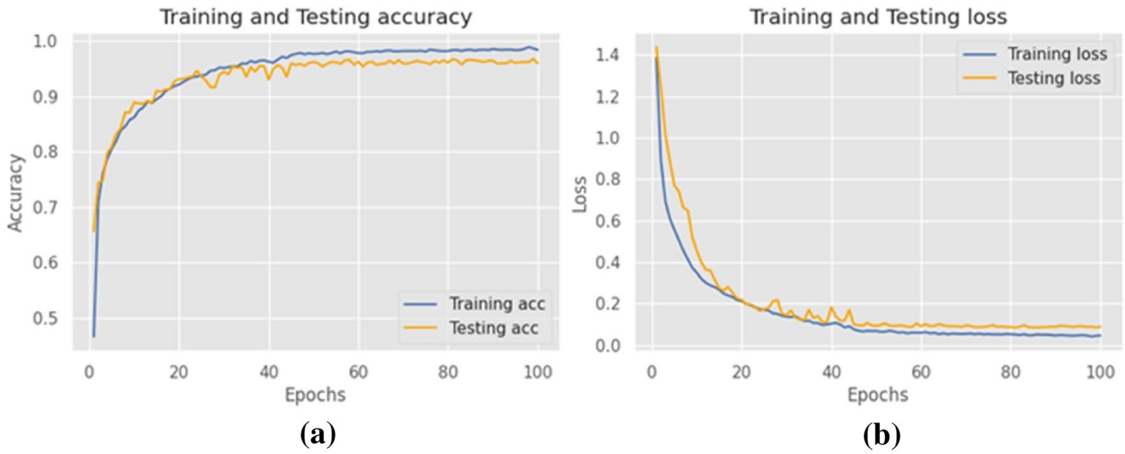


Fig. 3 **a** Accuracy vs. Epochs plot obtained from the proposed model on UCI-HAR dataset. **b** Loss vs. Epochs plot obtained from the proposed model on UCI-HAR dataset

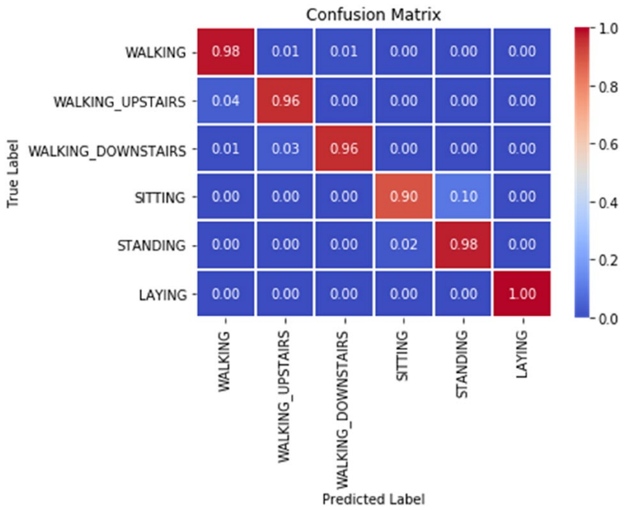


Fig. 4 Confusion matrix evaluated from testing the proposed model on UCI-HAR dataset

were considered for experimental purposes. All the values in this dataset are normalized to zero mean and standard deviation of one. These sensor data are sampled into fixed-width windows (128 readings/window) with a 50 percent overlap to match the temporal resolution with the above-adopted datasets.

3.2 Performance metrics

In this paper, the effectiveness of the proposed multibranch CNN-BiLSTM model is calculated by using different performance metrics [49] as described as follows:

Table 1 Performance comparison of various models with the proposed model on the UCI-HAR dataset

Model	F1-Score (%)	Accuracy (%)
Res-LSTM [35]	91.50	91.60
CNN-LSTM [36]	–	92.13
Bidir-LSTM [34]	–	92.67
CNN [27]	92.93	92.71
Stacked-LSTM [33]	–	93.13
Residual-BiLSTM [35]	93.50	93.60
LSTM-CNN [37]	–	95.78
Dilated-TCN [41]	93.80	93.80
ED-TCN [41]	94.60	94.60
Proposed Method	**96.31**	**96.37**

Accuracy: It is defined as the fraction of samples predicted correctly to the total number of samples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP = True Positives, FN = False Negatives, TN = True Negatives, and FP = False Positives.

Precision: The fraction of positive samples recognized correctly out of the total number of samples recognized as positive.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall: The fraction of positive samples that are recognized correctly out of the total positive samples.

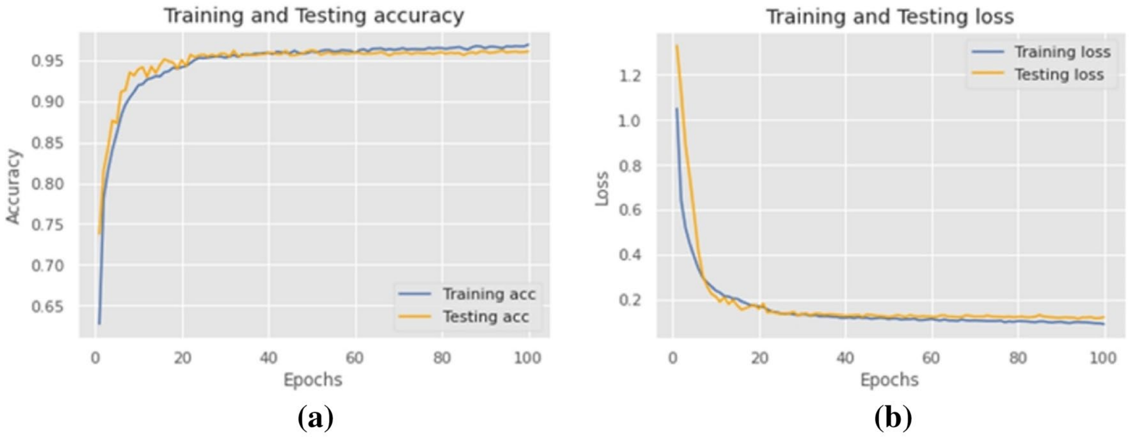


Fig. 5 **a** Accuracy vs. Epochs plot obtained from the proposed model on the WISDM dataset. **b** Loss vs. Epochs plot obtained from the proposed model on the WISDM dataset

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

F1-score: It is a comprehensive estimate of the model's accuracy and can be calculated as the harmonic mean of the precision and recall.

$$F1 - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Confusion matrix (CM): It is a square matrix that gives the complete performance of a classification model. The rows of the CM signify instances of the true class labels, and columns signify predicted class labels. The diagonal elements of this matrix define the number of points for which the predicted label is equal to the true label.

3.3 Results

To assess the performance of the proposed multibranch CNN-BiLSTM model, it is essential to segregate the whole dataset for training and testing intelligently. Since this work is based on HAR data, which is sequential data, it is not ideal to randomly split train and test data from the raw sensor data. This is due to the fact that if the data split has been done randomly, the classification model might see the activities of the same person in both train and test sets, thus leading to better accuracy, which does not correctly define the model's true performance. Hence, the ideal solution for splitting the HAR datasets would be to split the data according to the user id, i.e., train the model using data collected from a few users and ask the model to predict the movements of unseen users. In this work, three datasets WISDM, UCI-HAR, and PAMAP2 are adopted, and the data split has been done according to the user id for training and testing purposes. The proposed model is implemented on the Keras API with the TensorFlow backend. Adam optimizer is adopted for the training of the proposed model with a learning rate of 0.001. To measure the loss of the proposed classification model, categorical cross-entropy is used. For all the experiments, the model is trained with a batch size of 400 for 100 epochs on NVIDIA GeForce GTX 1660 Ti GPU.

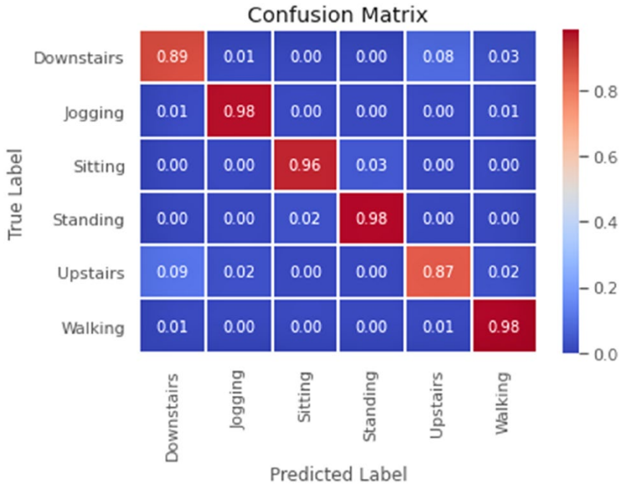


Fig. 6 Confusion matrix evaluated from testing the proposed model on the WISDM dataset

Table 2 Performance comparison of various DL-based models with the proposed model on the WISDM dataset

Model	F1-Score (%)	Accuracy (%)
CNN [28]	–	93.32
LSTM-CNN [37]	–	95.85
RNN-LSTM [50]	95.73	95.78
Proposed Method	**96.04**	**96.05**

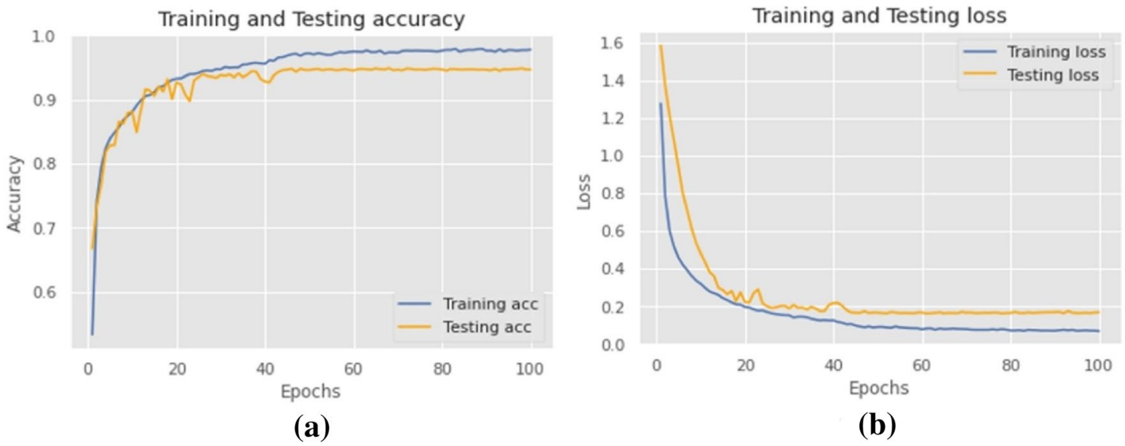


Fig. 7 **a** Accuracy vs. Epochs plot obtained from the proposed model on the PAMAP2 dataset. **b** Loss vs. Epochs plot obtained from the proposed model on the PAMAP2 dataset

3.3.1 Results on UCI-HAR dataset

From this dataset, 7352 samples are used to train the model (training data), while 2947 samples (testing data) are used to assess the model's performance. Figure 3 depicts the performance of the proposed model during both training and testing on the UCI-HAR dataset. The CM obtained by evaluating the trained proposed model on the testing data is depicted in Fig. 4. According to the CM, the proposed model has achieved above 95% classification accuracy in five out of six dataset classes.

The performance of the existing models [27, 33–37, 41] is compared with the proposed model in terms of F1-score and/or accuracy is presented in Table 1. The results show that the proposed model has outperformed the other compared techniques for HAR.

3.3.2 Results on WISDM dataset

The samples in the WISDM dataset are divided according to the user-id for training and testing the proposed model. Out of 36 user's data, the first 28 users are selected for training (13,042 samples), and the following eight users are selected for testing (4114 samples). The performance of the proposed model during both training and testing on the WISDM dataset is displayed in Fig. 5. The CM obtained by evaluating the trained proposed model on the test data is depicted in Fig. 6. According to the CM, the proposed model has achieved above 95% classification accuracy in four out of six dataset classes.

The performance of existing models [28, 37, 50] is compared with the proposed model in terms of F1-score and/

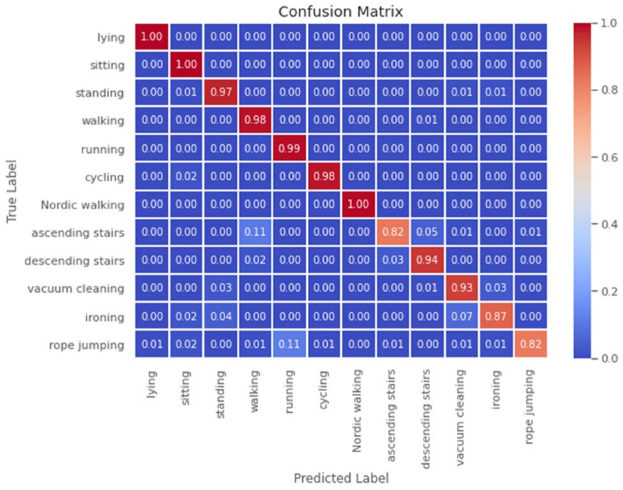


Fig. 8 Confusion matrix obtained using the proposed model for the PAMAP2 dataset

Table 3 Performance comparison of various DL-based models with the proposed model on the PAMAP2 dataset

Model	F1-Score (%)	Accuracy (%)
BiLSTM [27]	89.40	89.52
CNN [27]	91.16	91.00
ARC-NET [51]	90.76	90.91
LSTM-F [52]	92.90	–
COND-CNN [53]	–	94.01
Proposed Model	**94.27**	**94.29**

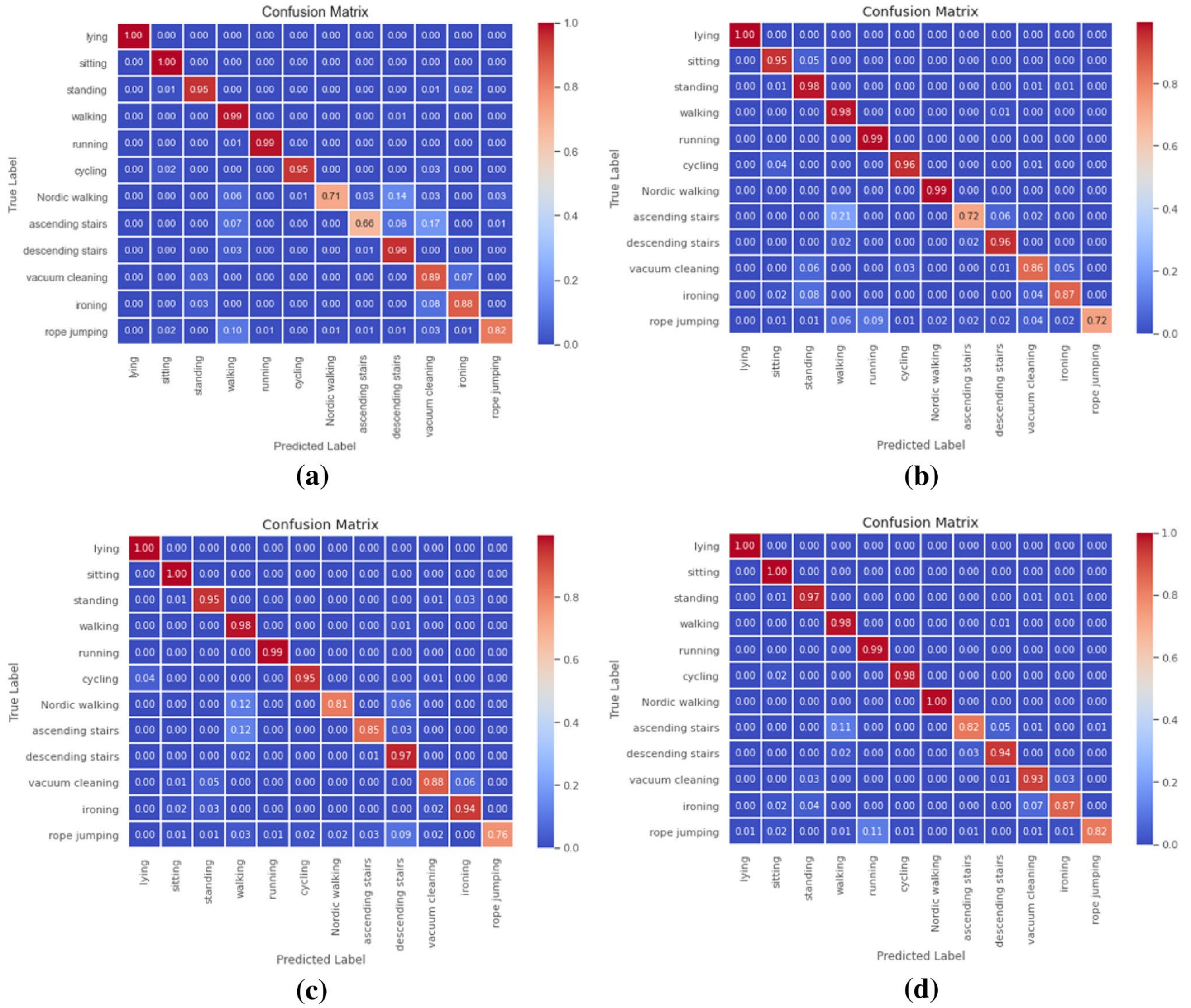


Fig. 9 Confusion matrices for the PAMAP2 dataset. **a** Confusion matrix for model-I (filter size-3), **b** Confusion matrix for model-II (filter size-7), **c** Confusion matrix for model-III (filter size-11), **d** Confusion matrix for the proposed model

or accuracy values and is presented in Table 2. The results show that the proposed model has achieved 96.05% accuracy on testing data, which is higher than the compared models.

3.3.3 Results on PAMAP2 dataset

The samples in this dataset are also divided according to the user-id for training and testing purposes. Among the nine subjects, the sensor data collected from the fifth and sixth subjects are used for testing, while the data from the other seven subjects are used for training the proposed model. The training and testing performance of the proposed model on the PAMAP2 dataset is shown in Fig. 7. The CM obtained from the testing data is depicted in Fig. 8. Except for rope jumping and ascending stairs, the proposed model has shown decent performance in the remaining classes of the PAMAP2 dataset.

The performance of various models [27, 51–53] is compared with the proposed model in terms of F1-score and/or accuracy values and is presented in Table 3. The results show that the proposed model has achieved 94.29% accuracy on testing data, which is higher than the compared models.

3.3.4 Performance comparison of different models using different filter sizes

The idea of the proposed approach is to extract a greater number of significant features during the training process. To achieve this, all the branches in the proposed model are

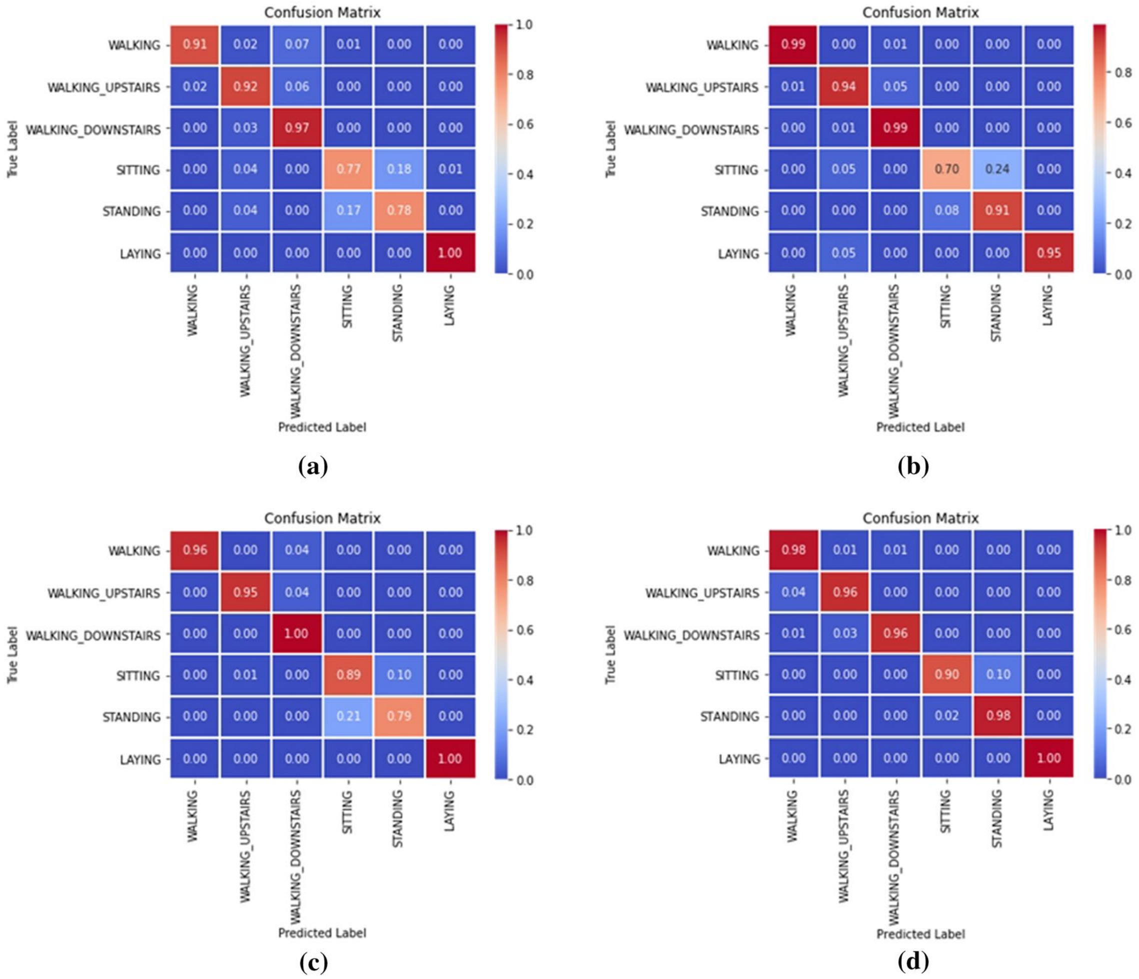
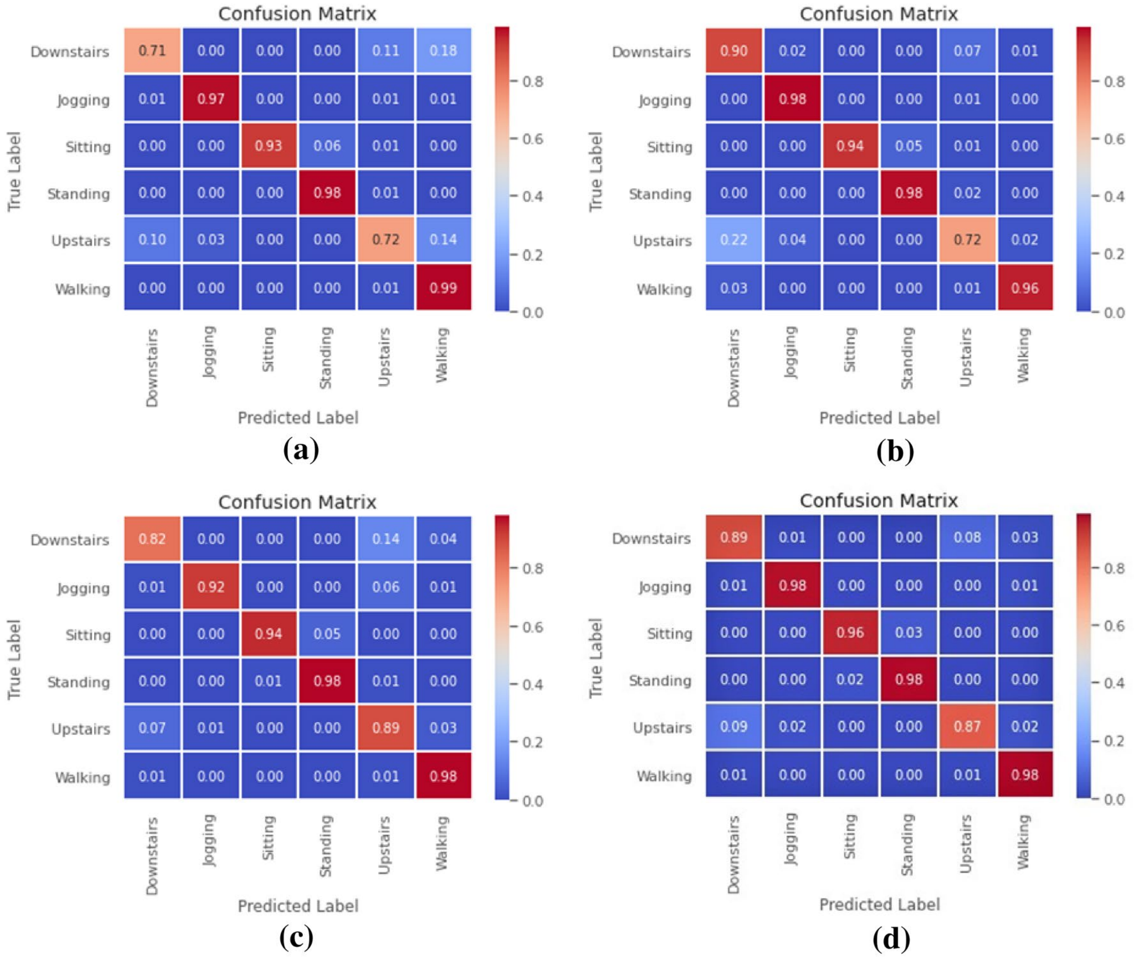


Fig. 10 Confusion matrices for the UCI-HAR dataset. **a** Confusion matrix for model-I (filter size-3), **b** Confusion matrix for model-II (filter size-7), **c** Confusion matrix for model-III (filter size-11), **d** Confusion matrix for the proposed model

kept in a uniform structure and the importance of filter size has been emphasized and thoroughly investigated using different filter sizes. The proposed multibranch CNN-BiLSTM model is compared to three single-branch CNN-BiLSTM models viz. model-I, model-II, and model-III. The single-branch CNN-BiLSTM models differ in the filter sizes used. Filter sizes of 3, 7, and 11 are used for model-I, model-II, and model-III, respectively. All of the individual models are trained using the same set of hyperparameters as the proposed model. Figure 9 shows the confusion matrices of the PAMAP2 dataset obtained using the model- I, II, III, and proposed model. From the confusion matrices, it is observed that model-I with filter size-3 performed well for recognizing the activity of 'rope jumping' while it performed poorly for 'Nordic walking.' Model-II with filter size-7 is excellent at recognizing 'Nordic walking' but performs poorly when recognizing 'rope jumping.' Compared to model-II, the Model-III with filter size-11 performed better on 'ironing' and 'ascending stairs,' but it performed poorly in detecting 'Nordic walking,' Hence, in the proposed model, when the combination of all the three filter sizes (i.e., 3, 7, and 11) is used together, the classification results obtained were optimum for most of the activities. Similar observations were found for UCI-HAR and WISDM datasets. Figures 10 and 11 depict the confusion matrices obtained for UCI-HAR and WISDM datasets, respectively.

The performance comparison of single-branch CNN-BiLSTM models with the proposed model is displayed in Table 4. From the results, it is observed that the concatenation of all three branches with different filter sizes makes the model more efficient compared to the single filter sizes used in the single-branch models.



(a)

	Downstairs	Jogging	Sitting	Standing	Upstairs	Walking
Downstairs	0.71	0.00	0.00	0.00	0.11	0.18
Jogging	0.01	0.97	0.00	0.00	0.01	0.01
Sitting	0.00	0.00	0.93	0.06	0.01	0.00
Standing	0.00	0.00	0.00	0.98	0.01	0.00
Upstairs	0.10	0.03	0.00	0.00	0.72	0.14
Walking	0.00	0.00	0.00	0.00	0.01	0.99

(b)

	Downstairs	Jogging	Sitting	Standing	Upstairs	Walking
Downstairs	0.90	0.02	0.00	0.00	0.07	0.01
Jogging	0.00	0.98	0.00	0.00	0.01	0.00
Sitting	0.00	0.00	0.94	0.05	0.01	0.00
Standing	0.00	0.00	0.00	0.98	0.02	0.00
Upstairs	0.22	0.04	0.00	0.00	0.72	0.02
Walking	0.03	0.00	0.00	0.00	0.01	0.96

(c)

	Downstairs	Jogging	Sitting	Standing	Upstairs	Walking
Downstairs	0.82	0.00	0.00	0.00	0.14	0.04
Jogging	0.01	0.92	0.00	0.00	0.06	0.01
Sitting	0.00	0.00	0.94	0.05	0.00	0.00
Standing	0.00	0.00	0.01	0.98	0.01	0.00
Upstairs	0.07	0.01	0.00	0.00	0.89	0.03
Walking	0.01	0.00	0.00	0.00	0.01	0.98

(d)

	Downstairs	Jogging	Sitting	Standing	Upstairs	Walking
Downstairs	0.89	0.01	0.00	0.00	0.08	0.03
Jogging	0.01	0.98	0.00	0.00	0.00	0.01
Sitting	0.00	0.00	0.96	0.03	0.00	0.00
Standing	0.00	0.00	0.02	0.98	0.00	0.00
Upstairs	0.09	0.02	0.00	0.00	0.87	0.02
Walking	0.01	0.00	0.00	0.00	0.01	0.98

Fig. 11 Confusion matrices for the WISDM dataset. **a** Confusion matrix for model-I (filter size-3), **b** Confusion matrix for model-II (filter size-7), **c** Confusion matrix for model-III (filter size-11), **d** Confusion matrix for the proposed model

3.3.5 Performance comparison of models with the different number of branches

To investigate the impact of the number of branches, the proposed model is also compared with the concatenation of different models such as single-branch model with filter size 3 or 7 or 11, dual-branch model with filter sizes 3 and 7 for branches I and II, tri-branch model with filter sizes 3, 7, and 11 for branches I, II and III, and quad branch model with filter sizes 3, 7, 11, and 15 for branches I, II, III, IV. All four models have similar architecture except for the number of branches. The single, dual, and quad-branch models are trained with the same hyperparameters as the proposed CNN-BiLSTM model. Table 5 displays the performance comparison of the single, dual, and quad-branch models with the proposed model.

Similarly, Table 6 shows the performance comparison of the mentioned models using the total number of parameters and the total time consumed during the training process. From Table 6 it is observed that, with the increased number of operational branches, the complexity of the network gradually increases and thus the computation cost also increases. Keeping aside the accuracy parameter of the quad-branch model (Table 5), the tri-branch model (proposed) maintains a comparatively good trade-off between the accuracy and computational cost, hence it is chosen in this work.

3.3.6 Performance comparison of the proposed model with a hybrid of CNN and other RNN variants

The proposed model, which is a hybrid of CNN and BiLSTM, is also compared against a hybrid of CNN and other RNN variants viz CNN-LSTM and CNN-GRU models. The multibranch architectures of the CNN-LSTM and CNN-GRU hybrid models are the same as that of the proposed multibranch CNN-BiLSTM model except that the LSTM

Table 4 Performance comparisons of single-branch CNN-BiLSTM models with the proposed model in terms of F1-Score (%) and accuracy (%)

Model	UCI-HAR		WISDM		PAMAP2	
	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy
Model-I	89.20	89.22	88.28	88.36	90.12	90.17
Model-II	91.20	91.17	91.73	91.71	91.79	91.81
Model-III	93.23	93.32	92.63	92.67	92.43	92.44
Proposed Model	**96.31**	**96.37**	**96.04**	**96.05**	**94.27**	**94.29**

Table 5 Performance comparison of the proposed model with the single, dual, and quad-branches in terms of F1-score (%) and Accuracy (%)

Model	UCI-HAR		WISDM		PAMAP2	
	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy
Single-branch	89.20	89.22	88.28	88.36	90.12	90.17
Dual-branch	94.12	94.08	93.84	93.88	92.78	92.83
Tri-branch (proposed)	**96.31**	**96.37**	**96.04**	**96.05**	**94.27**	**94.29**
Quad-branch	96.73	96.79	96.63	96.61	93.49	93.55

Table 6 Total training time and parameter configuration of various models with the different number of branches

Model	UCI-HAR		WISDM		PAMAP2	
	No. of parameters	Training Time (sec)	No. of parameters	Training Time (sec)	No. of parameters	Training Time (sec)
Single-branch	321,446	30.17	320,294	41.78	324,524	87.68
Dual-branch	503,750	42.77	499,910	60.93	512,204	122.58
Tri-branch (proposed)	**631,014**	**53.68**	**622,950**	**77.90**	**647,916**	**162.60**
Quad-branch	703,238	66.76	689,414	95.15	731,660	207.12

Table 7 Performance comparison of the proposed model with a hybrid of CNN and other RNN variants in terms of F1- score (%) and Accuracy (%)

Model	UCI-HAR		WISDM		PAMAP2	
	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy
CNN-LSTM	94.76	94.80	94.32	94.28	92.77	92.81
CNN-GRU	94.54	94.58	95.01	95.07	93.16	93.20
CNN-BiLSTM (proposed)	**96.31**	**96.37**	**96.04**	**96.05**	**94.27**	**94.29**

and GRU layers are, respectively, used in the place of BiLSTM layers. These models are trained with the same set of hyperparameters as the proposed model. In scenarios when the complete sequence of time series data is available, the BiLSTM uses both past and future information. The LSTM and GRU units, on the other hand, solely uses past information. Due to the additional context offered, the BiLSTM network can generate more accurate predictions as compared to LSTM and GRU. The same is validated by the results presented in Table 7. The performance measures obtained for CNN-LSTM and CNN-GRU models are comparable. However, the proposed model has a better recognition performance when compared to the other CNN-RNN variants.

The datasets used in the proposed model were recorded using various types of sensors placed at distinct locations of the human body, and the data were recorded at different sampling rates. The proposed model can identify simple and complex activities like walking, sitting, jogging, Nordic walking, ironing, vacuum cleaning, etc., with decent accuracy. According to the results obtained, the proposed multi-branch CNN-BiLSTM model tends to generalize well across all three datasets adopted.

4 Conclusion

In this work, a multibranch CNN-BiLSTM model has been proposed for HAR that directly operates on raw data captured from wearable sensors with minimal pre-processing. This model exploits the advantages of CNNs and BiLSTMs, which allows the model to capture both the local as well as long-term dependencies in sequential data. The proposed architecture makes use of multiple convolutional filter sizes to enhance feature extraction by capturing various local dependencies, and able to identify simple activities like walking, sitting, jogging, etc., along with complex activities like Nordic walking, vacuum cleaning, ironing, etc., with decent accuracy. Based on the obtained results, it can be observed that the proposed multibranch CNN-BiLSTM model could generalize well across all three datasets. The efficiency of the proposed multibranch CNN-BiLSTM model is evaluated on PAMAP2, WISDM, and UCI-HAR datasets and achieved 94.29%, 96.05%, 96.37% accuracies, respectively. The experimental results indicate that the proposed multibranch CNN-BiLSTM model outperformed the other HAR models.

Funding This research work is supported by SERB, DST, Government of India through the Early Career Research (ECR) award (ECR/2018/000203).

Declarations

Conflict of interest When writing the manuscript, we considered all ethical concerns and followed all guidelines to the best of our ability. The authors declare that there is no conflict of interest.

References

1. Mousse, M.A., Motamed, C., Ezin, E.C.: Percentage of human-occupied areas for fall detection from two views. Vis. Comput. **33**(12), 1529–1540 (2017)
2. Vishwakarma, D.K., Dhiman, C.: A unified model for human activity recognition using spatial distribution of gradients and difference of Gaussian kernel. Vis. Comput. **35**(11), 1595–1613 (2019)
3. Yao, L., Sheng, Q.Z., Benatallah, B., Dustdar, S., Wang, X., Shemshadi, A., Kanhere, S.S.: WITS: an IoT-endowed computational framework for activity recognition in personalized smart homes. Computing **100**(4), 369–385 (2018)
4. Mishra, A., Sharma, S., Kumar, S., Ranjan, P., Ujlayan, A.: Effect of hand grip actions on object recognition process: a machine learning-based approach for improved motor rehabilitation. Neural Computing and Applications, 1–12 (2020).
5. Sun, Z., Liu, J., Ke, Q., Rahmani, H., Bennamoun, M. and Wang, G.: Human Action Recognition from Various Data Modalities: A Review. arXiv preprint. (2020).
6. Pedersoli, F., Benini, S., Adami, N., Leonardi, R.: XKin: an open source framework for hand pose and gesture recognition using kinect. Vis. Comput. **30**(10), 1107–1122 (2014)
7. Chen, Z., Jiang, C., Xiang, S., Ding, J., Wu, M. and Li, X.: Smartphone Sensor Based Human Activity Recognition Using Feature Fusion and Maximum Full A Posteriori. IEEE Trans. Instrum. Meas. (2019).
8. Madhuranga, D., Madushan, R., Siriwardane, C. and Gunasekera, K.: Real-time multimodal ADL recognition using convolution neural networks. Vis. Comput., 1–14 (2020).
9. Abdelbaky, A. and Aly, S.: Two-stream spatiotemporal feature fusion for human action recognition. Vis. Comput., 1–15 (2020).
10. Yang, J., Nguyen, M. N., San, P. P., Li, X. and Krishnaswamy, S.: Deep convolutional neural networks on multichannel time series for human activity recognition. In: Ijcai, Vol. 15, pp. 3995–4001 (2015).
11. Anguita, D., Ghio, A., Oneto, L., Parra, X. and Reyes-Ortiz, J.L.: A public domain dataset for human activity recognition using smartphones. In: Esann*,* Vol. 3, p. 3 (2013).
12. Kwapisz, J.R., Weiss, G.M., Moore, S.A.: Activity recognition using cell phone accelerometers. ACM SIGKDD Explor. Newsl. **12**(2), 74–82 (2011)
13. Anguita, D., Ghio, A., Oneto, L., Parra, X. Reyes-Ortiz, J.L.: Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In: International Workshop on Ambient Assisted Living, pp. 216–223, Springer, Berlin, Heidelberg (2012).
14. Ignatov, A.D., Strijov, V.V.: Human activity recognition using quasiperiodic time series collected from a single tri-axial accelerometer. Multimed. Tools Appl. **75**(12), 7257–7270 (2016)
15. Gupta, A., Semwal, V.B.: Multiple Task Human Gait Analysis and Identification: Ensemble Learning Approach. In: Emotion and Information Processing, Springer, Cham, pp. 185–197 (2020).
16. Feng, Z., Mo, L. and Li, M.: A Random Forest-based ensemble method for activity recognition. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5074–5077. IEEE (2015)
17. Catal, C., Tufekci, S., Pirmit, E., Kocabag, G.: On the use of ensemble of classifiers for accelerometer-based activity recognition. Appl. Soft Comput. **37**, 1018–1022 (2015)
18. Nweke, H.F., Teh, Y.W., Al-Garadi, M.A., Alo, U.R.: 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 (2018)
19. Dewangan, D.K., Sahu, S.P.: PotNet: Pothole detection for autonomous vehicle system using convolutional neural network. Electron. Lett. **57**(2), 53–56 (2021)
20. Xi, P., Guan, H., Shu, C., Borgeat, L., Goubran, R.: An integrated approach for medical abnormality detection using deep patch convolutional neural networks. Vis. Comput., 1–14 (2019).
21. Dewangan, D.K., Sahu, S.P.: Deep learning-based speed bump detection model for intelligent vehicle system using Raspberry Pi. IEEE Sens. J. **21**(3), 3570–3578 (2020)
22. Chen, L., Wang, R., Yang, J., Xue, L., Hu, M.: Multi-label image classification with recurrently learning semantic dependencies. Vis. Comput. **35**(10), 1361–1371 (2019)
23. Semwal, V.B., Mondal, K., Nandi, G.C.: Robust and accurate feature selection for humanoid push recovery and classification: deep learning approach. Neural Comput. Appl. **28**(3), 565–574 (2017)
24. Dewangan, D.K., Sahu, S.P.: RCNet: road classification convolutional neural networks for intelligent vehicle system. Intel. Serv. Robot. **14**(2), 199–214 (2021)
25. Zhu, R., Tu, X., Huang, J.: Using deep learning based natural language processing techniques for clinical decision-making with EHRs. In: Deep Learning Techniques for Biomedical and Health Informatics, Springer, Cham, pp. 257–295 (2020).
26. Panwar, M., Dyuthi, S.R., Prakash K.C., Biswas, D., Acharyya, A., Maharatna, K., Gautam, A., Naik, G.R.: CNN based approach for activity recognition using a wrist-worn accelerometer. In: 2017

39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2438–2441. IEEE (2017).

27. Wan, S., Qi, L., Xu, X., Tong, C. Gu, Z.: Deep learning models for real-time human activity recognition with smartphones. Mobile Netw. Appl., pp.1–13 (2019).
28. Ignatov, A.: Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. Appl. Soft Comput. **62**, 915–922 (2018)
29. Wang, K., He, J., Zhang, L.: Attention-based convolutional neural network for weakly labeled human activities' recognition with wearable sensors. IEEE Sens. J. **19**(17), 7598–7604 (2019)
30. Teng, Q., Wang, K., Zhang, L., He, J.: The layer-wise training convolutional neural networks using local loss for sensor-based human activity recognition. IEEE Sens. J. **20**(13), 7265–7274 (2020)
31. Zeng, M., Nguyen, L.T., Yu, B., Mengshoel, O.J., Zhu, J., Wu, P. and Zhang, J.: Convolutional neural networks for human activity recognition using mobile sensors. In: 6th International Conference on Mobile Computing, Applications and Services, pp. 197–205. IEEE (2014).
32. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Networks **5**(2), 157–166 (1994)
33. Ullah, M., Ullah, H., Khan, S.D., Cheikh, F.A.: Stacked Lstm Network for Human Activity Recognition Using Smartphone Data. In: 2019 8th European Workshop on Visual Information Processing (EUVIP), pp. 175–180. IEEE (2019).
34. Hernández, F., Suárez, L.F., Villamizar, J., Altuve, M.: Human activity recognition on smartphones using a bidirectional lstm network. In: 2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA), pp. 1–5. IEEE (2019)
35. Zhao, Y., Yang, R., Chevalier, G., Xu, X., Zhang, Z.: Deep residual bidir-LSTM for human activity recognition using wearable sensors. Mathematical Problems in Engineering, 2018 (2018)
36. Mutegeki, R. and Han, D.S.: A CNN-LSTM Approach to human activity recognition. In: 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 362–366. IEEE (2020).
37. Xia, K., Huang, J., Wang, H.: LSTM-CNN architecture for human activity recognition. IEEE Access **8**, 56855–56866 (2020)
38. Ordóñez, F.J., Roggen, D.: Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors **16**(1), 115 (2016)
39. Guan, Y., Plötz, T.: Ensembles of deep lstm learners for activity recognition using wearables. Proc. ACM Interact. Mobile Wear Ubiquit. Technol. **1**(2), 1–28 (2017)
40. Karim, F., Majumdar, S., Darabi, H., Chen, S.: LSTM fully convolutional networks for time series classification. IEEE Access **6**, 1662–1669 (2017)
41. Nair, N., Thomas, C., Jayagopi, D.B.: Human activity recognition using temporal convolutional network. In Proceedings of the 5th International Workshop on Sensor-based Activity Recognition and Interaction, pp. 1–8 (2018)
42. Wang, K., He, J., Zhang, L.: Sequential weakly labeled multi-activity localization and recognition on wearable sensors using recurrent attention networks. arXiv e-prints, pp. arXiv-2004 (2020).
43. Li, T., Liu, J., Zhang, W., Ni, Y., Wang, W., Li, Z.: UAV-Human: a large benchmark for human behavior understanding with unmanned aerial vehicles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16266–16275 (2021).
44. LeCun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. Handbook Brain Theory Neural Netw. **3361**(10), 1995 (1995)
45. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
46. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**(11), 2673–2681 (1997)
47. Ioffe, S. and Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning, pp. 448–456. PMLR (2015).
48. Reiss, A. and Stricker, D.: Introducing a new benchmarked dataset for activity monitoring. In 2012 16th International Symposium on Wearable Computers, pp. 108–109. IEEE (2012).
49. Dewangan, D.K., Sahu, S.P., Sairam, B. and Agrawal, A.: VLD-Net: Vision-based lane region detection network for intelligent vehicle system using semantic segmentation. Computing, pp. 1–26 (2021).
50. Pienaar, S.W., Malekian, R.: Human activity recognition using LSTM-RNN deep neural network architecture. In: 2019 IEEE 2nd Wireless Africa Conference (WAC) pp. 1–5. IEEE (2019).
51. Damirchi, H., Khorrambakht, R., Taghirad, H.: ARC-Net: Activity Recognition Through Capsules. arXiv preprint (2020).
52. Hammerla, N.Y., Halloran, S. and Plötz, T.: Deep, convolutional, and recurrent models for human activity recognition using wearables. arXiv preprint (2016).
53. Cheng, X., Zhang, L., Tang, Y., Liu, Y., Wu, H., He, J.: Real-time Human Activity Recognition Using Conditionally Parametrized Convolutions on Mobile and Wearable Devices. arXiv preprint (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Sravan Kumar Challa received his Bachelors degree from St. Theressa Institute of Engineering and Technology, Garividi, Andhra Pradesh,India. Masters degree from NIT Jmashedpur,Jharkand, India and currently pursuing Ph.D. from National Institute of Technology,Jamshedpur, India. His research interests include, machine learing, computer vision and pattern recognition.

Akhilesh Kumar received his B.Sc. Engg.degree from Bhagalpur University, Bhagalpur,India. M.Sc. Engg. degree from Bihar Institute of Technology, Sindri, India and his Ph.D.Degree from Magadh University, Bodhgaya,India. Presently, he is working as an Associate Professor in Electronics and Communication Engineering Department, National Institute of Technology Jamshedpur, India. He has guided 18 M.tech thesis and more than 20 publications in reputed international journals and conferences. His research interests include Digital VLSI Circuit Design, Antennas, Machine Learning.

Vijay Bhaskar Semwal is working as Assistant professor (CSE) at NIT Bhopal since February 2019. Before joining NIT Bhopal he was working at NIT Rourkela. He has also worked with IIIT Dharwad as Assistant Professor(CSE) for 2 year (2016-2018) and he has also worked as Assistant professor (CSE) at NIT Jamshedpur . He has earned his doctorate degree in robotics from IIIT Allahabad (2017), M.Tech. in Information Technology from IIIT Allahabad (2010)and B.Tech. (IT) from College of Engineering Roorkee (2008). His areas of research are Bipedal Robotics, Gait Analysis and synthesis,Artificial Intelligence, Machine Learning, IoT and Theoretical Computer Science.He has published more then 15 SCI research papers.He has received early career research award by DST-SERB under government of India . His research areas are algorithm prospective of machine learning, artificial Intelligence and bipedal robotics. He has delivered more than 25 talks. Currently he is working on human health monitoring using gait pattern & reconstruction of impaired gait.