# Improved Sensor Based Human Activity Recognition via Hybrid Convolutional and Recurrent Neural Networks

1st Sonia Perez-Gamboa
*Sch. of Computer Science and Eng.*
*California State University*
San Bernardino, USA
Sonia.PerezGamboa@csusb.edu

2nd Qingquan Sun
*Sch. of Computer Science & Eng.*
*California State University*
San Bernarnido, USA
qsun@csusb.edu

3rd Yan Zhang
*Sch. of Computer Science & Eng.*
*California State University*
San Bernarnido, USA
Yan.Zhang@csusb.edu

2021 IEEE International Symposium on Inertial Sensors and Systems (INERTIAL) | 978-1-7281-5099-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/INERTIAL51137.2021.9430460

***Abstract*—Non-intrusive sensor based human activity recognition (HAR) is utilized in a spectrum of applications including fitness tracking devices, gaming, health care monitoring, and smartphone applications. In this paper, we design a multi-layer hybrid architecture with Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM). Based on the exploration of a variety of multi-layer combinations, we present a lightweight, hybrid, and multi-layer model which can improve the recognition performance by integrating local features and scale-invariant with dependencies of activities. The experimental results demonstrate the efficacy of the proposed model which can achieve a 94.7% activity recognition rate on a benchmark dataset. This model outperforms traditional machine learning and other deep learning methods. Additionally, our implementation achieves a balance between accuracy and efficiency.**

***Index Terms*—deep learning, LSTM, CNN, human activity recognition, inertial sensor**

## I. Introduction

HAR is the ability of a system to detect and identify specific human activities with the data collected through a sensor or camera. Digital images and videos from cameras are at the core of computer vision based HAR. Unfortunately, data collected through cameras has disadvantages including the inevitable capture of non-human activities that may be occurring in the background. Additionally, images and videos can be negatively affected by inconsistent lighting of the environment in which they are captured. Sensor data does not have these limitations, and therefore is a more promising way to collect human activity data [1]. Inertial sensors which can provide accelerometer and gyroscope data have been used to collect human activity because of their small size, low energy consumption, and non-intrusiveness on subjects [2], [3].

Traditional machine learning methods have been successful in accomplishing HAR on sensor based data. For instance, [1], [4] and [5] used a Support Vector Machine (SVM) for activity classification, [6] uses a decision tree, and [7] uses discriminant analysis, nearest neighbor classifiers, and ensemble classifiers. As successful as traditional machine learning methods are, in order to complete classification, prior feature extraction by an expert within the domain needs to be completed. This requirement for feature extraction limits the flexibility of these methods.

Deep learning models such as RNNs and CNNs have more recently been used to complete HAR due to their capability of automatically completing feature extraction on raw data without requiring a human expert to intervene, while obtaining high recognition accuracy. Training deep neural networks can be both computationally intensive and time expensive, taking hours or several days to train models.

## II. Related Works

Deep Learning techniques have been intensively investigated and applied to accomplish sensor-based HAR. CNNs have proven to work for HAR due to their capability of capturing local dependencies on signal data, as well as their preservation of feature scale invariance when completing feature extraction [8]. Some CNN models have been proved to achieve a higher HAR accuracy than other state of the art methods, while investigating the effect that different parameter values would have on the CNN model [9], [10]. CNN model parameters such as pooling size, weight decay, and drop out are shown to affect the overall accuracy of the model.

LSTM RNNs, which are excellent at properly handling long term dependencies on input sequences, have also been used to achieve HAR. A multi-layer LSTM RNN model, which had a lower recognition time than CNN based models, is used in to achieve a high HAR accuracy [11]. In [12] a bidirectional LSTM was found to outperform CNN and regular LSTM models when using a large benchmark dataset.

Recently, the combination of CNN and RNN models has been explored to further improve the performance of sensor based HAR. A DeepConvLSTM model was presented in [13], which combines convolutional, recurrent, and softmax layers to achieve a higher HAR rate than a baseline CNN model.

Our work is based on a hybrid multi-layered convolutional and recurrent neural network model that presents the following contribution to the field of sensor-based HAR: 1) We develop a lightweight, 2-layer CNN combined with 1 layer LSTM deep learning model that outperforms traditional and similar deep

978-1-7281-5099-4/21/$31.00 ©2021 IEEE
Authorized licensed use limited to: BIBLIOTECARIO DI INGEGNERIA. Downloaded on August 11,2023 at 14:15:24 UTC from IEEE Xplore. Restrictions apply.

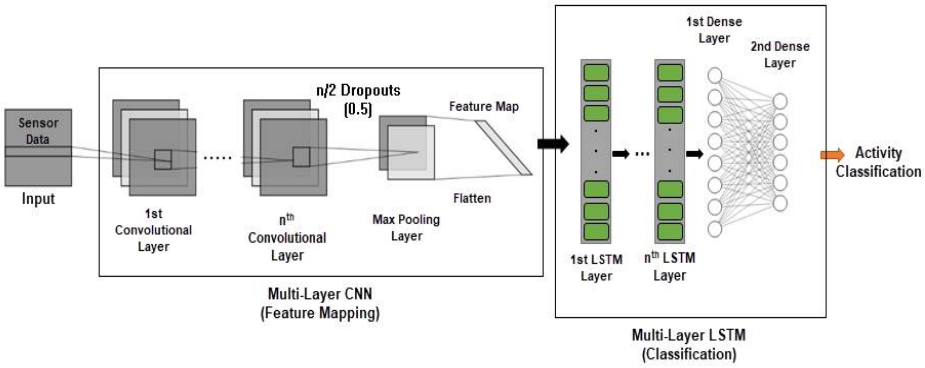


Fig. 1. Architecture of the multi-layer CNN and LSTM hybrid model.

learning models; 2) we develop and investigate a variety of CNN and LSTM hybrid models in the same environment to properly compare performances; 3) we develop models that have a balance between HAR accuracy and efficiency.

## III. Hybrid Deep Convolutional and Recurrent Neural Networks

### A. Convolutional Neural Networks (CNNs)

CNNs are deep learning models that utilize convolution, pooling, fully connected layers, and hidden layers to accomplish classification and recognition. Input data is fed into convolutional layers, which apply filters to the input data and identify local correlations. The pooling layers help with reducing the sensitivity of the output feature map by down sampling the detected features. Pooling helps the network identify the same feature, even if the exact location of the feature changes from one input sequence to the next. CNNs have proven to effectively perform independent, non-handcrafted feature extraction on raw sensor data, which enhances classification accuracy [15]. For our sensor based HAR case, we choose to use 1D CNN to fit and process time series data. In our models, convolutional layers, pooling layers, and dense layers are incorportated and a multi-layer structure is investigated. Chosen number of filters, kernel size, and the activation function were influenced by [8], where they tested various parameter values on a similar dataset, although we made modifications to increase our own performance.

### B. Long Short-Term Memory RNN (LSTM)

LSTM networks are a type of RNN that address the problem of vanishing and exploding gradients in regular RNNs with the addition of a special memory cell within each LSTM unit. The memory cell enables the LSTM to learn and retain dependencies on long input sequences, which has been shown to be ideal for HAR using sensor data [16]. The LSTM layers use the LSTM units to learn temporal relationships along the input sequence. Each LSTM layer is stacked with a dropout layer, which deters the problem of over fitting by randomly "dropping" some hidden units while the network trains. Finally, the fully connected layers complete activity classification. To further explore the dependencies of temporal features, a multi-layer structure of LSTM is investigated and implemented in this work. This multi-layer LSTM model also works as a benchmark to demonstrate the efficacy of the proposed hybrid model.

### C. Multi-layer CNN and LSTM Hybrid

Based on the multi-layer CNN and multi-layer LSTM architectures, we designed and implemented a hybrid multi-layer CNN-LSTM model that aims to 1) extract the local features in depth; 2) fully exploit the temporal dependencies; and 3) combine these two properties to enhance recognition performance. This hybrid model is shown in Fig.1. We investigated multiple combinations by varying the number of CNN and LSTM layers to see how accuracy is affected. The process of HAR for the hybrid model is as follows: 1) sensor data is input in sequences of shape 32x4 through 1D convolutional layer(s) containing 64 filters to generate independent, non-handcrafted feature maps; 2) the convolutional layers are stacked with a dropout layer of rate 0.5; 3) the output is passed through a max pooling layer of size 2, which down samples the feature maps; 4) the remaining feature maps are then flattened to be processed through the LSTM layer(s); 5) the LSTM layer(s) is (are) made up of 100 units which identify long-term temporal dependencies; 6) a fully connected layer of 100 units interprets the LSTM predictions; 7) finally, the predictions are passed through a dense layer made up of 6 units which applies a softmax activation function to complete classification.

## IV. Experiment and Evaluation

### A. Dataset

The dataset we used to test our models is the public domain UCI dataset for HAR [1]. It is a dataset containing inertial data from the embedded accelerometer and gyroscope in a Samsung Galaxy S II smartphone. The accelerometer and gyroscope provided linear acceleration and angular velocity signals, respectively, at a rate of 50Hz. The obtained signals were then pre-processed for noise reduction using a median and Butterworth filter. A total of 30 subjects, ages ranging from 19 to 48, wore the smartphone on their waist and performed six daily living activities: standing, sitting, laying down, walking, walking downstairs, and walking upstairs. Fig.

Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on August 11,2023 at 14:15:24 UTC from IEEE Xplore. Restrictions apply.

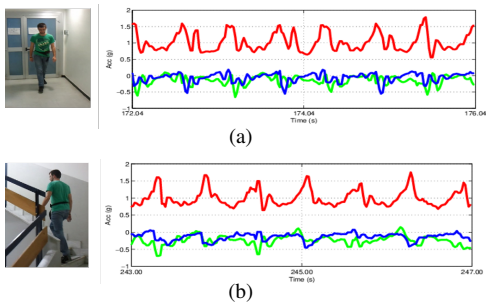


Fig. 2. (a) Sensor data of walking; (b) Sensor data of walking upstairs.

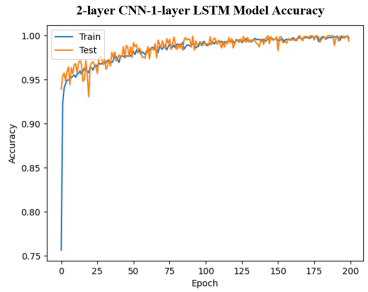


Fig. 3. Model accuracy of 2-layer CNN-1-layer LSTM hybrid.

TABLE I
SUMMARY OF PERFORMANCE METRICS

| Model | Mean Precision(%) | Mean Recall(%) | Mean F1-Score(%) |
|---|---|---|---|
| 1-layer LSTM | 91 | 90 | 90 |
| 2-layer LSTM | 91 | 91 | 91 |
| 1-layer CNN | 91 | 91 | 91 |
| 2-layer CNN | 93 | 93 | 92 |
| 1-layer CNN-1-layer LSTM | 91 | 91 | 91 |
| 1-layer CNN-2-layer LSTM | 91 | 91 | 91 |
| 2-layer CNN-1-layer LSTM | 95 | 95 | 95 |
| 2-layer CNN-2-layer LSTM | 94 | 94 | 94 |
| 2-layer CNN-3-layer LSTM | 93 | 93 | 92 |
| 3-layer CNN-1-layer LSTM | 92 | 92 | 92 |
| 3-layer CNN-2-layer LSTM | 92 | 92 | 92 |
| 3-layer CNN-3-layer LSTM | 93 | 93 | 93 |
| 4-layer CNN-1-layer LSTM | 94 | 94 | 94 |
| 4-layer CNN-2-layer LSTM | 93 | 93 | 93 |
| 4-layer CNN-3-layer LSTM | 91 | 92 | 91 |

TABLE II
SUMMARY OF PERFORMANCE ACCURACY

| Model | UCI HAR Accuracy (%) | Training Time(minutes) |
|---|---|---|
| 1-layer LSTM | 90.2 | 7.3 |
| 2-layer LSTM | 91.0 | 14.9 |
| 1-layer CNN | 91.1 | 3.2 |
| 2-layer CNN | 92.4 | 3.5 |
| 1-layer CNN-1-layer LSTM | 91.9 | 4.4 |
| 1-layer CNN-2-layer LSTM | 91.0 | 4.2 |
| 2-layer CNN-1-layer LSTM | 94.7 | 7.7 |
| 2-layer CNN-2-layer LSTM | 94.3 | 6.7 |
| 2-layer CNN-3-layer LSTM | 92.5 | 9.3 |
| 3-layer CNN-1-layer LSTM | 91.6 | 4.9 |
| 3-layer CNN-2-layer LSTM | 91.7 | 7.4 |
| 3-layer CNN-3-layer LSTM | 92.9 | 9.9 |
| 4-layer CNN-1-layer LSTM | 93.8 | 7.1 |
| 4-layer CNN-2-layer LSTM | 92.5 | 7.2 |
| 4-layer CNN-3-layer LSTM | 91.3 | 10.7 |

2 shows samples of walking and walking upstairs activities, as well as the corresponding sensor signals. The sampling data was randomly separated into training and testing sets, with 70% of the subjects being used for training, and 30% of the subjects being used for testing. We used 33% of the training data as validation data to obtain the accuracy and loss of our models over each epoch.

### *B. Implementation*

Implementation of the models was done using the Python programming language, Keras API, and Tensorflow framework. Results were obtained by running our models using an Amazon Elastic Compute Cloud instance with the following configuration: 1 NVIDIA Tesla V100 GPU, 8 Intel Xeon E5-2686 v4 CPUs, 16 GB GPU Memory, 100 GB SSD.

### *C. Performance Metrics*

In datasets with several activities, an imbalance of training and testing activities can occur, resulting in an overall performance accuracy that can be influenced by one activity that had a high classification, but ignore another activity with a low classification rate. To further evaluate our models, it is beneficial to look at the precision, F1 score, and recall percentages of each.

Table I shows the average of each of these metrics for all our deep learning models. This table shows us that our 2-layer CNN-1-layer LSTM model had the highest overall metrics with a mean precision of 95%, mean recall of 95%, and mean f1 score of 95%, which correctly reflects its high overall accuracy of 94.7% By comparing Table I and Table II, we can see that the overall accuracy of our models is proportional to their performance metrics, which also reveals that our dataset was balanced.

### *D. Performance Results of Multi-layer Hybrid Models*

As shown in Table II, the top 3 performing models were all hybrid models: 2-layer CNN-1-layer LSTM, 2-layer CNN-2 layer LSTM, and 4-layer CNN-1 layer LSTM, which obtained 94.7%, 94.3%, and 93.8% accuracy, respectively. This confirms that hybrid models of CNN and LSTM layers have better performance than standalone models.

The results of all implemented deep learning models and 2 traditional machine learning methods on the UCI HAR Dataset are shown in Table III, with the best performing model highlighted in bold. The table lists the performance accuracy of each model in classifying the activities, as well as the time it took to train each of the models. The same parameters were used across the models, and the same training and testing subsets were used, therefore taking a closer look at the overall training time of the model can provide us with more insight on how lightweight and fast a model is. Since deep learning methods are known for having long training times, a short training time is compelling.

Fig. 5 gives us a closer look at the precision, recall and F1 scores for each of the activities. By comparing Figure 4 and Figure 5, we can see that the results of the confusion matrix are directly proportional to the performance metrics.

### *E. Performance result of 2-layer CNN-1-layer LSTM Hybrid*

Among all these multi-layer hybrid CNN-LSTM models, the one composed of 2-layer CNN and 1-layer LSTM has the

Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on August 11,2023 at 14:15:24 UTC from IEEE Xplore. Restrictions apply.

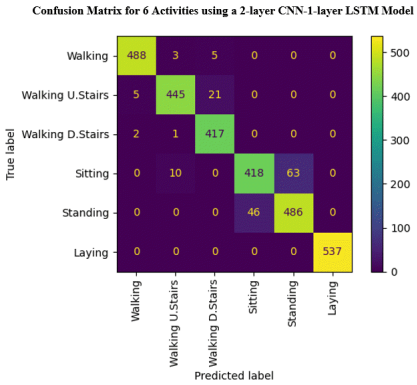


Fig. 4. 2-layer CNN-1-layer LSTM Confusion Matrix.

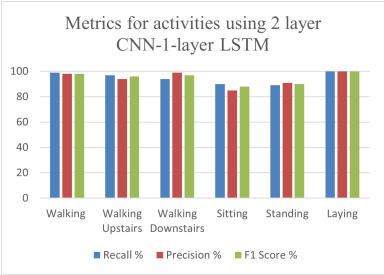


Fig. 5. Performance Metrics for 2-layer CNN-1-layer LSTM.

best performance. Therefore, we offer a closer look and deep analysis to this model here.

Fig. 3 demonstrates the accuracy rate of our model as it was training. We can see here that the accuracy stabilizes around 175 epochs, which would suggest that increasing our epochs over 200 would not benefit the model.

Fig. 4 presents a confusion matrix for our model for each of the 6 activities in the dataset. The confusion matrix provides insight on the number of times our model classified a specific activity correctly, or classified it as another activity. Fig. 4 shows that our model struggled the most with differentiating between the"sitting" and "standing" actions, likely due to the similarity in acceleration and subject orientation while performing these actions. Our model had a 100% classification rate for the "laying" activity, and only misclassified walking downstairs 3 times.

From Table III, we see that the 2-layer CNN combined with 1-layer LSTM outperformed all other models with a strong accuracy of 94.7%. All deep learning models outperformed the traditional methods presented by [4], which upholds that deep learning methods which automatically extract features and complete classification, outperform methods that require hand-crafted features.

## V. Conclusion

In this paper, we present a multi-layer, hybrid and lightweight CNN-LSTM architecture that outperform previously presented traditional machine learning and deep learning methods. Our 2-layer CNN and 1-layer LSTM hybrid model outperforms all other multi-layer hybrid models. Moreover, the lightweight hybrid model not only has a high performance accuracy, but also has a faster model training time. Future work will use larger, more complicated datasets to verify the performance of our lightweight hybrid model.

TABLE III
Summary of Performance Accuracy

| Model | UCI HAR Accuracy (%) | Training Time(minutes) |
|---|---|---|
| MC-SVM[4] | 89.3 | Unknown |
| MC-HP-SVM[4] | 89.0 | Unknown |
| 1-layer LSTM | 90.2 | 7.3 |
| Multi-layer LSTM | 91.0 | 14.9 |
| 1-layer CNN | 91.1 | 3.2 |
| Multi-layer CNN | 92.4 | 3.5 |
| Multi-layer CNN-LSTM hybrid | **94.7** | 7.7 |

## References


[1] D.Anguita, A.Ghio, L.Oneto, X.Parra and J.Reyes-Ortiz,"A Public Domain Dataset for Human Activity Recognition Using Smartphones",*21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013*,2013.

[2] L.Xie, J.Tian, G.Ding and Q.Zhao,"Human activity recognition method based on inertial sensor and barometer",*2018 IEEE International Symposium on Inertial Sensors and Systems (INERTIAL)*, pp.1-4, 2018.

[3] C.Chen, R.Jafari, and N.Kehtarnavaz,"UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and a Wearable Inertial Sensor",*Proceedings of IEEE International Conference on Image Processing*, 2015.

[4] J.-L.Reyes-Ortiz, L.Oneto, A.Samà, X.Parra, D.Anguita,"Transition-aware human activity recognition using smartphones",*Neurocomputing*, vol. 171, pp. 754-767, 2016.

[5] H.Xu, Z.Huang, J.Wang and Z.Kang,"Study on Fast Human Activity Recognition Based on Optimized Feature Selection",*2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES)* , pp. 109-112, 2017.

[6] L.Fan, Z.Wang, H.Wang,"Human activity recognition model based on decision tree",*Proc. CBD*, pp. 64-68, 2013.

[7] N.F.Ghazali, N.Shahar, N.A.Rahmad, N.A.J.Sufri, M.A.As'ari and H.F.M. Latif,"Common sport activity recognition using inertial sensor",*2018 IEEE 14th International Colloquium on Signal Processing and Its Applications (CSPA)*, pp.67-71, 2018.

[8] M.Zeng, L.T.Nguyen, B.Yu, O.J.Mengshoel, J.Zhu, P.Wu, J.Zhang ,*6th International Conference on Mobile Computing, Applications and Services* , pp.197-205, 2014.

[9] K.Nakano and B.Chakraborty,"Effect of dynamic feature for human activity recognition using smartphone sensors",*International Conference on Awareness Science and Technology*, pp.539-543, 2017.

[10] I.Andrey,"Real-time human activity recognition from accelerometer data using convolutional neural networks",*Applied Soft Computing*, pp. 1-8, 2017.

[11] T.Yu, J.Chen, N.Yan and X.Liu,"A Multi-Layer Parallel LSTM Network for Human Activity Recognition with Smartphone Sensors",*2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1-6, 2018.

[12] N Y.Hammerla, S.Halloran and T.Ploetz,"Deep convolutional and recurrent models for human activity recognition using wearables ", 2016.

[13] F.J.Ordonez, D.Roggen,"Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition",*Sensors*, 2016.

[14] N.Tufek, M.Yalcin, M.Altintas, F.Kalaoglu, Y.Li and S.K.Bahadir,"Human Action Recognition Using Deep Learning Methods on Limited Sensory Data",*IEEE Sensors Journal*, vol. 20, no. 6, pp. 3101-3112, 2020.

[15] J.B.Yang, M.N.Nguyen, P.P.San, X.L.Li,S. Krishnaswamy,"Deep Convolutional Neural Networks On Multichannel Time Series For Human Activity Recognition",*24th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3995-4001, 2015.

[16] Y.Guan and T.Plotz,"Ensembles of deep LSTM learners for activity recognition using wearables", *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol*, vol. 1, no. 2, pp. 1-28, 2017.


Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on August 11,2023 at 14:15:24 UTC from IEEE Xplore. Restrictions apply.