2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)
16-17 June 2023, Gazipur-1707, Bangladesh

BCL: A Branched CNN-LSTM Architecture for Human Activity Recognition Using Smartphone Sensors

Saimoon Al Farshi Oman *, Md. Nafis Jamil†, and S. M. Taslim Uddin Raju‡
Department of Computer Science and Engineering
Khulna University of Engineering & Technology (KUET),
Khulna-9203, Bangladesh
saimoon.oman@gmail.com*, nafisjamil72@gmail.com†, raju.taslim@cse.kuet.ac.bd‡

2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM) | 979-8-3503-1600-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/NCIM59001.2023.10212972

Abstract—**Human activity recognition, or HAR for short, is a process of interpreting specific human motion based on sensor data. HAR has many human-centric applications, notably in eldercare and healthcare as an assistive service. However, due to the noisy sensor data, it requires domain analysis and signal processing to extract features from the raw data to fit into the machine learning models. The recent revolution of Deep Learning Models makes it possible to learn the features automatically instead of handcrafting features. This area extensively utilizes deep learning techniques like CNN and RNN. In this paper, we present a branch CNN and LSTM structure for recognizing human activity that yields cutting-edge outcomes. The experiment is conducted on the SHOAIB Al and UCI HAR datasets, which produce better results than the traditional approach.**

Index Terms—**Human Activity Recognition (HAR); Branch CNN-LSTM (BCL); CNN; LSTM; Deep Learning; Smartphone; Sensors.**

I. Introduction

In the context of Human-Robot Interaction, the process of recognizing human activity is referred to as Human Activity Recognition (HAR). HAR is a key component of healthcare, particularly in assisting with elder care, supporting rehabilitation, and detecting cognitive disorders [1]. Typically, data for HAR is gathered using either cameras or sensors [2]. The disadvantages of using camera data are the large size of the data, insufficient lighting conditions, privacy issues and it can only monitor some specific areas where the cameras are located. Another widely used method is wearable sensors which are mainly used for experimentation, and not particularly used by general people due to their cost. Nowadays smartphones are equipped with multiple sensors capable of human activity recognition [3]. As most people use smartphones anyway, it doesn't require any extra cost.

The traditional approach to machine learning involves the extraction of statistical features from raw sensor data in the time and frequency domains. Feature engineering is a demanding task that necessitates specialized knowledge and can be quite time-consuming. Moreover, there exists a potential risk of losing valuable information, such as the temporal relationships between actions, during the feature extraction process. [4]. Newly developed deep learning models have been proven capable of producing excellent results in HAR without handcrafted feature extraction. These models can learn representative features due to their stacking structure.

In this paper, we introduce a CNN-LSTM branch network model designed to identify human activities using time-series data captured from inertial sensors embedded in smartphones. Our contributions to this study include:

- Introducing a CNN-LSTM branch model capable of automatically extracting features while preserving time dependencies for human activity classification.
- Using a CNN model to extract characteristics from raw data frames, which are subsequently interpreted by an LSTM model.
- Conducting experiments on the proposed model, which demonstrate superior performance compared to conventional machine learning methods, along with deep learning models like CNN, LSTM, or a combined CNN-LSTM architecture.

The structure of the paper is as follows: Section II presents a comprehensive overview of prior research conducted on human activity recognition using sensor data. Section III introduces the proposed methodology in detail. Section IV discusses the results obtained from the proposed method. Finally, in Section V, the paper concludes by highlighting potential future research directions that can be conducted in this area.

II. Related Works

Foerster and Smeja first introduced human activity recognition using sensor data in the 1990s [5], with the most accurate results being achieved by placing various sensors on different parts of the body. In [6], Ling Bao and Stephen S. Intille achieved an 84% accuracy rate using decision tree classifiers and five small biaxial accelerometers placed on different body parts.

For using machine learning algorithms, feature extraction has always been an important task. Widely used hand-crafted features are statistical features, fourier transform and wavelet transform. In [7], Kwapisz, Weiss and Moore performed HAR using J48 decision trees which perform better than other data

979-8-3503-1600-1/23/$31.00 ©2023 IEEE

mining techniques. In [8], Ankita Jain et al. used k-nearest neighbor (KNN) and support vector machine (SVM) for activity recognition over two public datasets. Naïve Bayesian and k-nearest neighbor (KNN) were used by P. Gupta and T. Dellas [9]. All these works have derived their own hand-crafted features and having different experimental grounds it makes difficult to compare them with each other. Inappropriate hand-designed features may affect the result. So, domain knowledge is required for designing hand-crafted features for a specific application [4].

Due to recent advancements in processing power, various deep learning algorithms have been introduced for categorization tasks, which extract features automatically without domain knowledge [10]. Y. Chen introduced an LSTM-based approach on the (WISDM) Lab public dataset, achieving 92.1% accuracy [11]. S. Yu and L. Qin improved the accuracy to 93.79% using Bidir-LSTM Networks [12]. T. Yu et al. obtained similar performance to CNN by using a parallel multi-layer LSTM network on the public UCI HAR dataset, but with lower computational complexity [13]. J. Wang surveyed the progress of deep learning techniques in activity recognition using sensors and found that RNN performed better for short-term activities while CNN was better suited for long activities [14]. K. Xia et al. introduced a combined LSTM-CNN network where sensor data was inputted into a two-layer LSTM (Long Short-Term Memory) network, which was then followed by convolutional layers [15].

III. Proposed Methodology

The objective of this study is to identify human daily activities using smartphone sensor data, specifically accelerometer and gyroscope data. In this section, we present the CNN-LSTM network architecture, which is designed to accomplish this task.

A. Preprocessing

Human Activity Recognition is done on raw sequential sensor data. Because the data is sequential in nature, it cannot be split randomly. Otherwise, data from the same participant can be found in both the testing sets and training sets. As a result, the accuracy may increase, but it fails to accurately represent the true performance of the model. So, the dataset needs to be split participant-wise.

The proposed work utilizes two datasets that are publicly available, namely SHOAIB and UCI-HAR, to conduct the experiments and evaluate the performance of the approach.

1) SHOAIB dataset: Each participant has sensor data for 5 positions: left pocket, right pocket, right leg (using a belt clipper), right wrist, and right upper arm. Only left and right pocket data were considered. Magnetometer sensor data was ignored. (100, 9) sized frames were made on left and right pocket data separately with 50% overlap. Then left pocket and right pocket data were concatenated. The framing was done on each participant individually. Then out of 10 participants, 3 participants were randomly chosen for test data. In particular, 5,032 samples were reserved for testing whereas 20,128 samples were allotted for training. This separation allows for a complete evaluation of the model's performance on unseen or unobserved data, enabling a thorough evaluation of its efficacy.

2) UCI dataset: The dataset is split into two halves at random: 30% of the data are put aside for testing, while the remaining 70% are assigned for training. The data from the sensors were divided into a window of size (128, 9). Specifically, there were 7,352 samples allocated for training purposes, while 2,947 samples were set aside for testing. The accelerometer signal was separated into two components using a low-pass filter: the gravitational force and the body motion. To filter out the gravitational force, a filter with a 0.3 Hz cutoff frequency was used. Features were extracted from the time and frequency domains to create feature vectors for each window of size (128, 9).

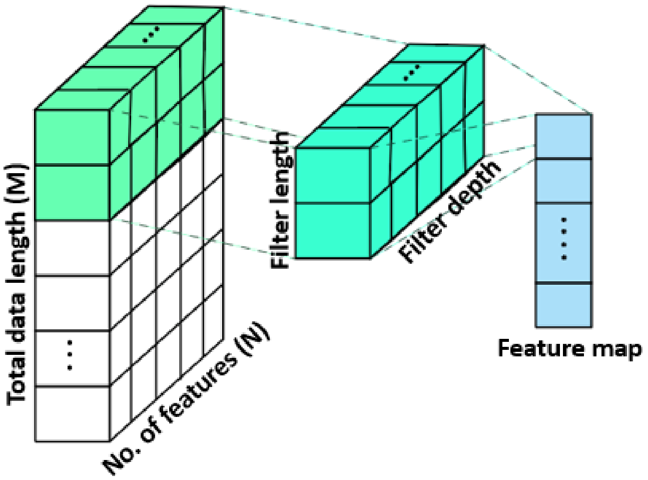


Fig. 1. Feature extraction process of CNN.

B. Feature Extraction

The data collected from smartphone sensors to capture human activity forms a time series, which has a clear one-dimensional structure. This structure means that variables that are close together in time are strongly correlated with each other. So it is crucial to identify and isolate local features within data. CNN is capable of achieving this by using local receptive fields. Time-series data of ($M \times N$) is taken as input to CNN in Fig. 1, where M represents the total data length and N is the no. of features available in the data. For extracting features from time-series data convolution filters are used. The filter length of the proposed model in each branch is 3 and the depth is the same as the no. of features N. The no. of feature maps created by the convolution process depends on the no. of filters employed in the operation. The sliding window technique is utilized to segment the input data into frames. CNN treats each frame as a separate unit of data, ignoring any temporal context outside of frame borders. The temporal context between the data frames is also necessary to identify activities accurately. Therefore, to capture temporal features various techniques for HAR have utilized Recurrent Neural Networks (RNNs).

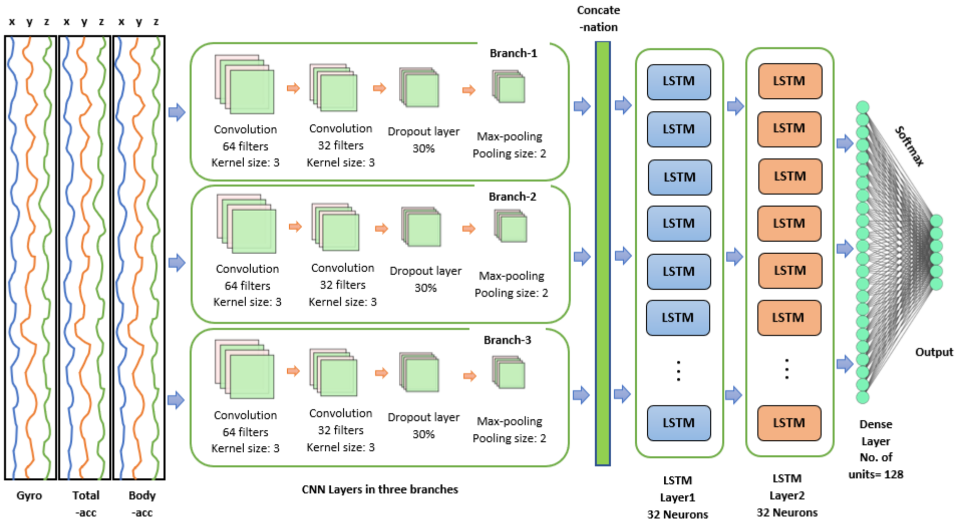

x y z x y z x y z
Gyro
Total -acc
Body -acc
Branch-1
Branch-2
Branch-3
Convolution 64 filters Kernel size: 3
Convolution 32 filters Kernel size: 3
Dropout layer 30%
Max-pooling Pooling size: 2
CNN Layers in three branches
Concate -nation
LSTM
LSTM Layer1 32 Neurons
LSTM Layer2 32 Neurons
Dense Layer No. of units= 128
Softmax
Output

Fig. 2. CNN-LSTM branch architecture.

RNN suffers vanishing gradients problem. For this, it could not capture long-term dependencies. LSTM is good with long-term dependencies. In contrast to normal RNN, LSTM features a chain-like structure with many gates on the repeating module. So, the network can adapt its predictions to the added context more correctly since LSTM can manage long-term dependencies.

C. Model Construction and Validation

In this proposed CNN-LSTM branch network, in each branch we have used convolution layers to extract the features of human activities, then the extracted features are passed through a LSTM layer. Finally, a dense layer is used to classify human activity.

Fig. 2 shows our proposed CNN-LSTM branch approach to classify activities. Keras API can go from idea to result in the least amount of time which makes it a valuable tool for conducting efficient and impactful research. We have implemented a Keras model utilizing TensorFlow as the backend on NVIDIA GTX 1050TI(GPU). Our model takes 9 signals; acceleration (a_x, a_y, a_z), linear acceleration (la_x, la_y, la_z), and angular velocity (g_x, g_y, g_z).

The sensor data is converted into a fixed window size and passed through three parallel convolution layers having 64 filters. Each output of these convolution layers is passed through another convolution layer with 32 filters. These convolution layers extract necessary features which are important for human activity recognition. Rectified linear units (ReLU) are used to construct the feature maps in both convolution layers in three branches with kernel size 3. A 30% dropout layer is added in each branch and passed through a pooling layer having pool size 2 in each branch. In order to lighten the computational load and enhance basic translation invariance in the internal representation, we employ max pooling within the pooling layer. This technique helps in reducing the number of factors that must be taught.

$$f(x) = max(0, x) \tag{1}$$

The output of each branch is concatenated and two LSTM layers are introduced after concatenation having 32 and 64 neurons. The LSTM layers receive extracted features in order to extract the temporal dependencies of the signal which are necessary for short-term human activities like walking, jogging, etc. Then the LSTM layers output is provided to a Dense layer having 128 neurons. Here, the rectified linear units (ReLU) activation function is utilized. Finally, 50% output of the Dense layer is normalized and then passed to a fully linked output layer with a Softmax activation for classifying human activity. To reduce categorical cross-entropy loss, the proposed design was trained using an RMSprop optimizer with a 0.0001 learning rate. The training of the model was conducted using 32 batch sizes over 200 epochs with early stopping(a method that checks the model's performance while it is being trained and stops it if no further improvement is seen).

D. Performance metrics

Various performance metrics are employed to assess the efficacy of the suggested model.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

$$f1 - score = \frac{2 \times precision \times recall}{precision + recall} \tag{5}$$

IV. Results and Discussions

Tab. I displays the effectiveness of several algorithms for recognizing various activities. Tab. II and Tab. III describe the specificity, recall, precision and f1-score of the SHOAIB and UCI dataset respectively.

TABLE I. Performance measurement of HAR using different algorithms.

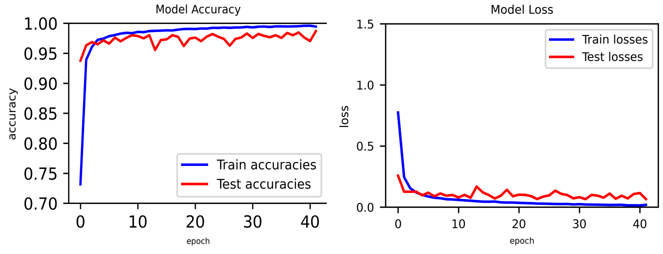
Algorithms	SHOAIB		UCI	
	Accuracy	Loss	Accuracy	Loss
CNN	96.7%	0.12	91.48%	0.27
LSTM	95.3%	0.17	89.03%	0.33
CNN-LSTM(without branch)	96%	0.14	89.21%	0.34
Proposed Method	**98%**	**0.07**	**93.72%**	**0.25**

A. Dataset Description

1) SHOAIB Dataset: Shoaib et al. introduced this particular dataset. In total, ten male participants aged between 25 to 30 performed eight common daily activities like running, walking, standing, sitting, jogging, walking upstairs, biking, and walking downstairs for 3 to 4 minutes. Participants wore five smartphones on different body positions to collect data from sensors such as a gyroscope, accelerometer, linear accelerometer, and magnetometer at a 50 Hz frequency. The data collection took place inside except for biking. Fig. 3 depicts the accuracy and loss values across epochs, while Fig. 4 illustrates the corresponding confusion matrix.

TABLE II. Performance measurement on SHOAIB dataset.

Activity	Specificity	Recall	Precision	F1-score
Biking	0.997217	0.995833	0.983539	0.989648
Walking	0.996058	0.938889	0.975469	0.956829
Jogging	0.998840	0.981944	0.992978	0.987430
Walking upstairs	0.994434	0.965278	0.966620	0.965949
Walking downstairs	0.990509	0.987360	0.944892	0.965659
Sitting	1.000000	0.997222	1.000000	0.998609
Standing	0.999768	0.994444	0.998605	0.996521



(a) (b)

Fig. 3. Accuracy and loss plots of SHOAIB dataset: (a) accuracy (b) loss.

2) UCI Dataset: In this study, A Samsung Galaxy S II smartphone was worn around the waists of 30 volunteers, ages 19 to 48, as they took part in six different activities. The sensors in a smartphone recorded linear acceleration and angular velocity in three directions (x,y,z), at a 50 Hz frequency. Using video recordings the data was labeled manually and the dataset was randomly split into training and test data, with a ratio of 70% for training data and 30% for testing data. Noise was removed from the sensor data before segmenting them into 50% overlapped 2.56-second windows. Fig. 5 depicts the

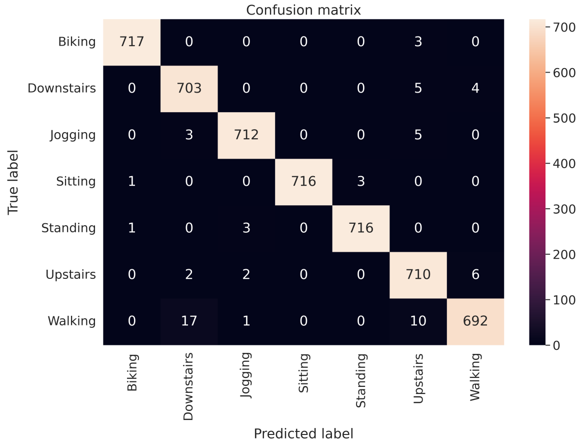
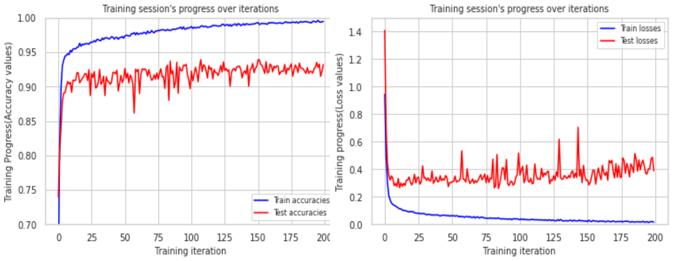


Fig. 4. Confusion matrix on SHOAIB dataset.

TABLE III. Performance measurement on UCI dataset.

Activity	Specificity	Recall	Precision	F1-score
Walking	0.990208	0.993952	0.953578	0.973346
Walking upstairs	0.983037	0.919321	0.911579	0.915433
Walking downstairs	0.989711	0.942857	0.938389	0.940618
Sitting	0.969055	0.818731	0.841004	0.829721
Standing	0.973085	0.851504	0.874517	0.862857
Laying	1.000000	1.000000	1.000000	1.000000



(a) (b)

Fig. 5. Accuracy and loss plots of UCI dataset: (a) accuracy (b) loss.

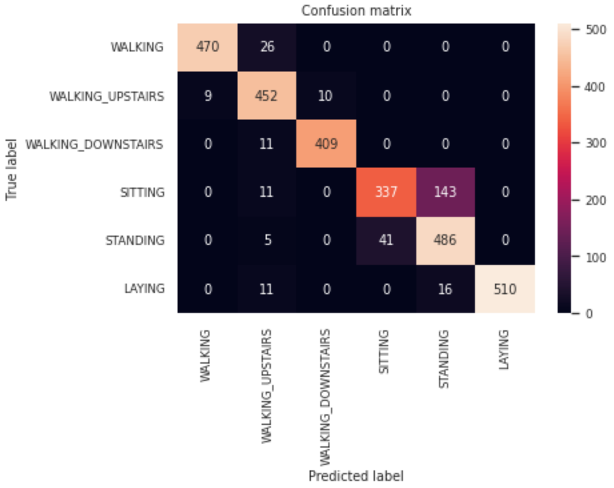


Fig. 6. Confusion matrix on UCI dataset.

accuracy and loss values across epochs, while Fig. 6 illustrates the corresponding confusion matrix.

V. Conclusion

The proposed method successfully distinguishes between a variety of human activities, including walking, running, sitting, standing, jogging, and biking, with high accuracy, proving the power of deep learning techniques to automatically acquire useful features from unstructured data. With this method, the advantages of LSTM networks and convolutional neural networks are combined. It does away with the necessity for manually created features, which is a need for conventional machine learning methods. When compared to traditional machine learning techniques, the proposed method for Human Activity Recognition (HAR) employing smartphone sensors offers the highest accuracy and efficiency. In the future, we want to extend our work for smartphone position-independent activity recognition.

References

[1] C. Chen, R. Jafari, and N. Kehtarnavaz, “A survey of depth and inertial sensor fusion for human action recognition,” *Multimedia Tools and Applications*, vol. 76, pp. 4405–4425, 2017.

[2] Y. Chen and C. Shen, “Performance analysis of smartphone-sensor behavior for human activity recognition,” *Ieee Access*, vol. 5, pp. 3095–3110, 2017.

[3] Z. Chen, Q. Zhu, Y. C. Soh, and L. Zhang, “Robust human activity recognition using smartphone sensors via ct-pca and online svm,” *IEEE transactions on industrial informatics*, vol. 13, no. 6, pp. 3070–3080, 2017.

[4] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. Ieee, 1999, pp. 1150–1157.

[5] F. Foerster, M. Smeja, and J. Fahrenberg, “Detection of posture and motion by accelerometry: a validation study in ambulatory monitoring,” *Computers in human behavior*, vol. 15, no. 5, pp. 571–583, 1999.

[6] L. Bao and S. S. Intille, “Activity recognition from user-annotated acceleration data,” in *Pervasive Computing: Second International Conference, PERVASIVE 2004, Linz/Vienna, Austria, April 21-23, 2004. Proceedings 2*. Springer, 2004, pp. 1–17.

[7] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, “Activity recognition using cell phone accelerometers,” *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.

[8] A. Jain and V. Kanhangad, “Human activity classification in smartphones using accelerometer and gyroscope sensors,” *IEEE Sensors Journal*, vol. 18, no. 3, pp. 1169–1177, 2017.

[9] P. Gupta and T. Dallas, “Feature selection and activity recognition system using a single triaxial accelerometer,” *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1780–1786, 2014.

[10] Y. Tang, R. Salakhutdinov, and G. Hinton, “Robust boltzmann machines for recognition and denoising,” in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2264–2271.

[11] Y. Chen, K. Zhong, J. Zhang, Q. Sun, and X. Zhao, “Lstm networks for mobile human activity recognition,” in *2016 International conference on artificial intelligence: technologies and applications*. Atlantis Press, 2016, pp. 50–53.

[12] S. Yu and L. Qin, “Human activity recognition with smartphone inertial sensors using bidir-lstm networks,” in *2018 3rd international conference on mechanical, control and computer engineering (icmcce)*. IEEE, 2018, pp. 219–224.

[13] T. Yu, J. Chen, N. Yan, and X. Liu, “A multi-layer parallel lstm network for human activity recognition with smartphone sensors,” in *2018 10th International conference on wireless communications and signal processing (WCSP)*. IEEE, 2018, pp. 1–6.

[14] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, “Deep learning for sensor-based activity recognition: A survey,” *Pattern recognition letters*, vol. 119, pp. 3–11, 2019.

[15] K. Xia, J. Huang, and H. Wang, “Lstm-cnn architecture for human activity recognition,” *IEEE Access*, vol. 8, pp. 56 855–56 866, 2020.