Alinteri J. of Agr. Sci. (2023) 38(3): 38-46
e-ISSN: 2587-2249
info@alinteridergisi.com
http://dergipark.gov.tr/alinterizbd
http://www.alinteridergisi.com/

DEEP LEARNING-BASED MODELS FOR ACCURATE HUMAN ACTIVITY RECOGNITION

[1]Dr.N. SATHYAVATHI, [2]ANIL KUMAR YERRAKUNTLA,[3]POLEPAKA PRASHAMSA,
[4]SWAPNA CHUNCHU, [5]NALLELLA RAKESH,[6]BANOTH SANDHYA
[1]Associate Professor,[234] Assistant Professor,[56]Student
Department Of CSE
Vaagdevi College of Engineering, Warangal, Telangana

ABSTRACT

Human Activity Recognition (HAR) is a rapidly growing field with applications ranging from healthcare and fitness tracking to smart home systems and security. Traditional methods often rely on handcrafted features and conventional machine learning algorithms, which may not effectively capture the complex temporal and spatial patterns inherent in human activities. Deep learning-based models have emerged as powerful tools for addressing these challenges, leveraging their ability to automatically learn hierarchical features from raw sensor data.

This study explores the deployment of deep learning architectures, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer models, for accurate HAR. These models capitalize on their unique strengths: CNNs excel at extracting spatial features, RNNs and LSTMs are adept at handling sequential data, and Transformer models offer state-of-the-art performance in capturing long-term dependencies. Techniques such as data augmentation, multimodal sensor fusion, and transfer learning are also discussed to enhance model robustness and generalization.

Benchmarking against publicly available datasets highlights the superior performance of deep learning models over traditional approaches, achieving higher accuracy and robustness across diverse activity categories. Challenges such as computational cost, the need for large labeled datasets, and real-time inference are addressed, with potential solutions proposed. The findings demonstrate that deep learning-based HAR systems can significantly advance the automation and reliability of activity recognition tasks, paving the way for innovative applications in various domains.

1. INTRODUCTION

Human Activity Recognition (HAR) is a critical area of research that involves identifying and classifying physical activities performed by individuals based on data collected from various sensors, such as accelerometers, gyroscopes, and cameras. HAR has a wide range of applications, including health monitoring, fitness tracking, elderly care, rehabilitation, smart homes, and security systems. Accurate recognition of human activities is essential to improving the quality of life, enabling personalized services, and enhancing user experiences in these domains.

Traditional HAR systems often rely on handcrafted features and classical machine learning algorithms such as Support Vector Machines (SVMs) and Decision Trees. While these methods have demonstrated effectiveness in certain scenarios, their

performance is often constrained by the complexity of feature engineering and the inability to generalize across diverse activities or sensor configurations. These limitations underscore the need for more robust and automated approaches capable of learning complex patterns directly from raw data.

In recent years, deep learning has revolutionized many fields, including computer vision, natural language processing, and speech recognition, due to its ability to learn hierarchical features from data. Deep learning-based models have demonstrated remarkable success in HAR by addressing the limitations of traditional methods. Models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformers have shown significant promise in capturing spatial and temporal dependencies in activity data, thereby enabling more accurate recognition of activities.

This paper explores the potential of deep learning-based models in advancing HAR systems. We examine the strengths and limitations of various deep learning architectures, evaluate their performance on standard datasets, and discuss techniques to enhance their efficiency and robustness. The introduction of techniques such as multimodal data fusion, transfer learning, and real-time inference capabilities further enhances the applicability of these models in practical scenarios.

The remainder of this paper is organized as follows: Section 2 reviews related work and the evolution of HAR methods. Section 3 details the deep learning architectures used for HAR. Section 4 presents experimental results and performance benchmarks. Finally, Section 5 discusses challenges and future directions in the field, followed by a conclusion in Section 6.

Through this exploration, we aim to highlight the transformative potential of deep learning in achieving accurate and reliable human activity recognition, paving the way for innovative applications across diverse domains.

2. BACKGROUND

Determining critical factors [6] and preserving resources related to the implemented models require investigating the effects of various data collecting, data preparation, and deep learning methods on a model's ability to identify activities. Vision-based recognition [7] and sensor-based recognition [8] are the two main categories into which human activity recognition is frequently separated. As the name implies, one or more cameras are used to record video samples of human activity for vision-based recognition models. Working with vision-based models involves either treating the video sample as a human silhouette for prediction [10] or combining many views to determine a single movement [9]. Since sensors from mobile devices or body attachments are much more accessible and effective than cameras for both training and real-world applications, sensor-based recognition is a much wider field of study [11] and application [12]. The OPPORTUNITY [13], Skoda Checkpoint [14], UCI-HAR [15], WISDM [16], MHEALTH [17], and PAMAP2 [18] datasets are a few of the often used datasets for models. Some of the most well-liked machine learning algorithms [19] in the field of human activity recognition have been recognised for their excellent accuracy and consistency across many datasets. These techniques include the Random Forest classifier (RF), k-Nearest Neighbour (kNN), and Support Vector Machine (SVM) [20]. However, as deep learning techniques were used to construct adequate data preparation feature extraction, training, and classification procedures, machine learning [21] methods lost favour since deep learning algorithms [23] are considerably more powerful and resource efficient [22]. Using deep learning techniques based on CNN, LSTM, and hybrid layers within the model's architecture, this survey

examines a few cutting-edge human activity detection algorithms..

3. HUMAN ACTIVITY RECOGNITION USING DEEP LEARNING METHODOLOGIES

A few prominent research that suggest models based on CNN, LSTM, and hybrid deep learning architectures are presented in this section.

A. CNN

A lightweight CNN model for wearable device-based human activity recognition was used in a recent study [24]. It made use of many datasets that collected information from portable sensors and smartphones. These datasets include the WISDM, PAMAP2, UNIMIB-SHAR, OPPORTUNITY, and UCI-HAR datasets. Additional research using CNN-based human activity identification models and smartphone data can be found in [25] and [26], where the models demonstrated good performance.

.

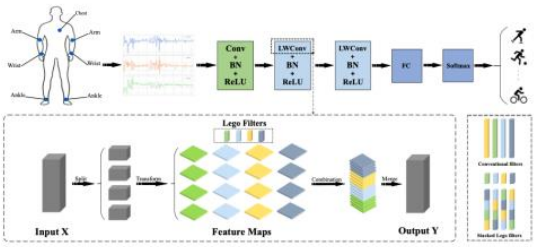

Figure 1. The architecture of a lightweight CNN based model featured in [24]. The model is unique for having a set of lower-dimensional filters which is used as Lego bricks and are stacked for conventional filters resulting in the model being independent of any form of special network structures.

The model's distinctive feature is its collection of lower-dimensional filters, which function similarly to Lego blocks and are stacked for traditional filters. This allows the model to function independently of any kind of unique network topologies. To ensure that operations continue, data signals are divided into windows of predetermined sizes, with overlap between windows allowed. The model is incredibly efficient when ordinary convolution filters are swapped out for lower dimensional Lego filters, which are also optimised during the training phase. The binary mask for optimisations is a straight through estimator (STE). A traditional split transform merge three step approach is used to take use of the intermediate feature maps and speed up convolutions. The various convolution layers employ ReLU activation functions, and the activity label prediction is then made using a Softmax function based on the network's output. A cross entropy between the local liner classifier's prediction and its target is used to implement the local loss functions. A three-by-three kernel with a stride and padding of one form the basis of the alternative loss function. With significantly fewer training parameters and good accuracy thanks to training with local loss functions, the Lego CNN uses less memory and computation than a standard CNN. All of the listed datasets were used to train the model, which was then assessed using the weighted F1 score and total classification accuracy. Thus, the UCI-HAR and WISDM datasets yielded the best results for the model, with accuracy of 96.9% and 98.82% and F1 scores of 96.27% and 97.51%, respectively.

A CNN-based methodology was also proposed in another study [27] that used anonymous binary sensors to identify older people's activities in an unobtrusive manner. In order to identify freezing of gait in individuals with Parkinson's disease, a CNN-based model trained with innovative spectral data techniques was utilised in a related work [28]. Using an Aruba-annotated open dataset, the model from [27] was constructed using a deep convolution network (DCNN) that was able to predict ten actions that were recorded by a single elderly woman over an eight-month period.

.

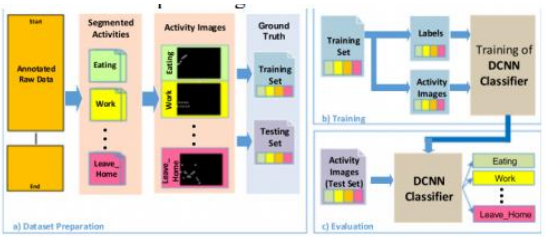

Figure 2. This figure shows the experimental setup [27] used for its DCNN classifier. It focused on using data collected from elderly people using anonymous binary sensors.

As previously stated, since models demonstrated efficacy with such methods in [30], it is typical practice to gather data through the use of specialised body-worn sensors [29]. Segmenting the activities according to their occurrences in the dataset, putting them through a sliding window to obtain fragmented samples, and finally turning the activity into an activity image are the steps taken to preprocess the raw data for the model. The action was represented by a straightforward binary picture in black and white with black and white on and off signals. Three convolution layers made up the DCNN classifier, which was then followed by pooling layers for feature extraction. The last layer is ultimately connected to ten outputs after the output of the final max-pooling layer was flattened and sent to the neurones of the connected layers. The accuracy, precision, recall, F1 score, and error rate were used to assess the activity's predictive capacity. With accuracy scores of 99.99% and 99.85%, respectively, walking from the bed to the toilet and working were the most accurate activities. When predicting 10 activities, the DCNN model's average accuracy was 98.54%, and when predicting 8 activities, it was 99.23%..

B. LSTM

The effectiveness of bi-directional LSTM (BiLSTM) approaches in feature extraction and prediction has led to its growing popularity in human activity recognition. The prediction capabilities of models from studies [31] and [32] are based on a BiLSTM, whereas [33] uses it for special feature extraction methods. In order to extract spatial information from the multidimensional data of MEMS inertial sensors, the model makes use of the residual block. The residual block has a CNN-based design because it can automatically extract local spatial data. A 2D CNN residual network with 23 2 x 2 kernels is utilised to extract the spatial information from various sensor data. To speed up training and prevent covariate shift problems, a batch normalisation layer is added, and the stride length is set to two. After that, it passes through a second convolution layer with the same configuration but a stride of 1 after using a ReLU activation function. Prior to being entered into the model, the data was standardised. Due to its highest recognition accuracy, the convolution kernels' size were 2 x 2. In order to balance the model size and training cost, 32 convolution kernels were chosen. With an ideal learning rate of 0.0003, 0.0006, and 0.00003, the ADAM optimiser was used to minimise the model's cross entropy. The models were trained 80 times with a batch side of 64. The BiLSTM layer then uses the features' forward and backward dependencies before feeding the features onto a Softmax layer for HAR. Three distinct datasets were used to evaluate the model: PAMAP2, WISDM, and a created dataset. Compared to previous deep learning-based models, the suggested models also produced even better outcomes with a lot less training parameters. With an accuracy of 96.95% for the handmade dataset, 97.32% for WISDM, and 97.15% for PAMAP2, it demonstrated strong performance across all three datasets..

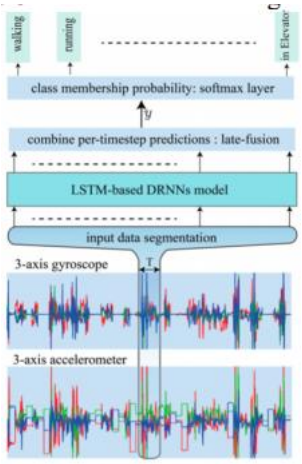

Figure 3. The LSTM based model proposed in [34] that utilizes accelerometer and gyroscope data.

In [34], LSTM-based unidirectional, bidirectional, and cascaded topologies were introduced. The model captures long-range dependencies in variable-length input sequences by using a deep recurrent neural network. It may therefore categorise windows of human activity of varying length. The model was given data from 3-axis accelerometers and gyroscopes that had been divided into time series windows. After receiving input, the model produces a series of scores that correspond to activity labels, each of which has a label prediction for a different time step. The prediction is represented by a vector of scores. Prediction scores are then translated into probabilities using the SoftMax layer. Three distinct models were tested using three different setups of the LSTM-based classifier. There are three types of LSTM-based DRNN models: unidirectional, bidirectional, and cascaded bidirectional and unidirectional. The UCI-HAD dataset, USCHARD dataset, Opportunity dataset, Daphnet FOG dataset, and Skoda dataset were among the datasets used to assess the models' efficacy. Eighty percent of the data was utilised to train each model, with the remaining twenty percent being used for validation. The mean cross entropy between the ground truth labels and the expected output labels was used to continuously update the model's weights, which were initially random. The cost of updating model parameters and backpropagating gradients was minimised using the Adam optimiser. The USC-HAD dataset yielded the best overall performance from the unidirectional LSTM-based DRNN. Its average precision was 97.4% and its overall accuracy was 97.8%. Additionally, the hybrid model outperformed standalone deep learning algorithm-based models like CNN and conventional machine learning algorithms [36] like KNN. Similar methods of utilising LSTM layers in the context of an RNN-based model for human activity recognition [35] are also used in models from [37] and [38].

C. Hybrid Models

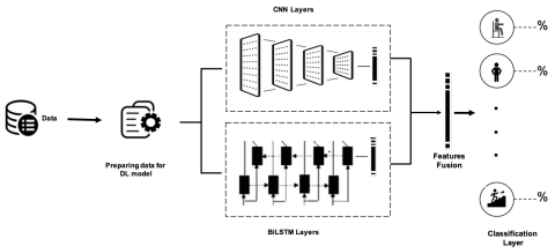

Figure 4. : A hybrid model that utilizes both CNN and (Bi)LSTM layers from [39].

Introducing a model for human activity recognition using a CNN with varied kernel dimensions that collaborate with a bi-directional long short-term memory (BiLSTM) layer to record features at various resolutions was the aim of another study [39]. This is comparable to the suggested models found in research [40], [41], and [42]. For advanced feature extraction and processing methods that strengthen the model's classification performance, all of these research combine

CNN and LSTM layers in the model. When employing CNN and BiLSTM to extract spatial and temporal data from the sensors, the model from [43] performs exceptionally well. An activation function required to be used to convert the model's input into the output value. The CNN then used both maximum and average pooling to reduce the dimensionality of the input data. High-level features relating to long-term strategies are automatically learnt over time steps by the LSTM layer. The BiLSTM component, which can access context both ahead and backward, was utilised to gain temporal representation regarding activity recognition. The model was fitted across 30 epochs with a batch size of 128 samples using the WISDM and UCI datasets. The accuracy of the model was 98.53% on the WISDM dataset and 97.05% on the UCI dataset.

A convolution neural network (CNN) with a Gated Recurrent Unit (GRU), a CNN with a GRU and attention, a CNN with a GRU and a second CNN, and a CNN with Long Short-Term Memory (LSTM) and a second CNN are the four models that were presented in [45]. The paper suggests four distinct deep learning-based models for identifying human activity that make use of WiFi 902.11n's channel state information (CSI). A publicly available dataset was used. Utilising the Linux CSI 802.11n utility, data was gathered. When it comes to explaining WiFi signals, this gadget is perfect. Channels show varying amplitudes for various actions, and actions between a WiFi transmitter and receiver are recorded. Lying down, falling, walking, running, sitting down, and standing up are among the activities that were labelled. In an indoor office, six individuals completed each task 20 times, and the data was divided into 80% and 90% groups for testing and training. The CNN-GRU model was the most effective model. Three components make up the CNN-GRU model: input, feature extraction, and classification. The CSI data is reshaped into 1000 x 30 x 3 matrices because it is found to be a good input for CNNs. Two convolutional layers and a GRU layer are employed for feature extraction. A size 5 x 5 x 128 kernel and a size 1 x 1 stride filter are applied to the input data, and then batch normalisation, ReLU activation, average pooling, and dropout with a value of 0.6 are applied. In order to transform the data into a vector appropriate for the GRU layer, the output goes via a flattening layer with time-distributed input. With a 99.46% accuracy rate, 99.52% precision rate, and 99.90% AUC, it outperformed the other three models as well as other cutting-edge models. In a related study [46], human activity from cameras was classified using a deep-neural network-based model that is trained using shared-weight and transfer learning approaches. The accuracy of the model, which included shared-weight LSTMRes and pre-trained CNN layers, was 97.22%.

.

4. A SUMMARY AND ANALYSIS OF FEATURED WORKS

Title	DL based Methodologies	Experimental Results	Limitations
Layer-wise Training Convolutional Neural Networks with Smaller Filters for Human Activity Recognition Using Wearable Sensors [24]	A lightweight CNN model for human activity recognition based on wearable devices. The model consists of a set of lower-dimensional filters which is used as Lego bricks and are stacked for conventional filters.	The model performed best with the UCI-HAR and WISDM datasets, getting an accuracy of 96.9% and 98.82% and F1 scores of 96.27% and 97.51%.	The use of smaller Lego filters results in a slight decrease in performance compared to a model based on a traditional CNN.
Unobtrusive Activity Recognition of Elderly People Living Alone Using Anonymous Binary Sensors and DCNN [27]	A model based on a deep convolution network (DCNN) and focused on unobtrusive activity recognition of elderly people using anonymous binary sensors.	On average the DCNN model was able to achieve an accuracy of 98.54% when predicting 10 activities, and 99.23% when predicting 8 activities.	Model's dataset was limited to working with a high-cost setup of binary sensors that recorded the movements for training and validation of the model. Untested with more accessible devices such as smartphones or smartwatches.
Human Activity Recognition Based on Residual Network and BiLSTM [33]	Utilized a residual block for extract spatial features from multidimensional signals and bi-directional LSTM (BiLSTM).	It performed well for all three datasets used, with the homemade dataset getting an accuracy of 96.95%, WISDM getting 97.32% and the PAMAP2 getting 97.15%.	Specific activity labels from the datasets that are unbalanced result in lower predicting success. Future work intends to address this to further increase accuracy.
Deep Recurrent Neural Networks for Human Activity Recognition [34]	A unidirectional, bidirectional, and cascaded architectures based on LSTM that uses a deep recurrent neural network for capturing long-range dependencies in variable-length input sequences.	The unidirectional LSTM based DRNN performed best with the USC-HAD dataset. It had an overall accuracy of 97.8% and an average precision of 97.4%.	The model's capabilities were only tested with basic human activities on a small scale and not tested with large scale complex activities.
Sensor-Based Human Activity Recognition with Spatio-Temporal Deep Learning [38]	This model uses a convolution neural network (CNN) with varying kernel dimensions that work with a bi-directional long short-term memory (BiLSTM) layer to capture features at different resolutions.	The model performed well for three datasets, with the homemade dataset getting an accuracy of 96.95%, WISDM getting 97.32% and the PAMAP2 getting 97.15%.	Power and memory usage was not considered for this model, hence the potential for low-powered devices to struggle with such a model.
Utilizing deep learning models in CSI-based human activity recognition [44]	The study proposes four different human activity recognition models based on deep learning that utilizes channel state information (CSI) in WiFi 902.11n.	The best performing model was the CNN-GRU model. It obtained an accuracy of 99.46%, precision of 99.52% and AUC of 99.90%, outperforming the other three models.	No use of denoising in the signals from the data before training.

Table 1 summarizes the features studies that were surveyed as state-of-the-art human activity recognition models that utilize deep learning-based architecture to achieve its predictive capabilities. The papers include a division of models that utilize CNN layers, LSTM layers and hybrid models that utilize more than one algorithm, such as employing both CNN and LSTM layers in the model.

5. CONCLUSION

Human Activity Recognition (HAR) has emerged as a crucial domain with diverse applications in healthcare, fitness, smart environments, and security systems. Traditional approaches relying on handcrafted features and classical machine learning techniques, while effective in controlled scenarios, often fall short in handling the complexities and variability of real-world activity data. Deep learning-based models, with their ability to learn hierarchical and high-level features directly from raw sensor data, have demonstrated significant potential in overcoming these limitations.

This study has highlighted the efficacy of various deep learning architectures, including

Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer models, in achieving state-of-the-art performance in HAR tasks. These models excel at capturing both spatial and temporal dependencies in data, providing robust and accurate recognition of activities. Furthermore, advanced techniques such as multimodal sensor fusion, transfer learning, and data augmentation have been shown to enhance model performance and adaptability to diverse application scenarios.

Despite these advancements, challenges remain. Computational costs, the need for large labeled datasets, and difficulties in deploying models for real-time inference in resource-constrained environments are critical areas that require attention. Future research should focus on developing lightweight models, semi-supervised and unsupervised learning methods, and optimization techniques for real-time processing. Additionally, addressing privacy concerns associated with sensor data collection and sharing will be vital for the widespread adoption of HAR systems.

In conclusion, deep learning-based models represent a transformative approach to HAR, offering unprecedented accuracy and adaptability. By addressing existing challenges and leveraging emerging technologies, HAR systems can play a pivotal role in improving quality of life and enabling intelligent, context-aware applications across multiple domains.

6. REFERENCES

1. Janelle Mason, Rushit Dave, Prosenjit Chatterjee, Ieschecia GrahamAllen, Albert Esterline, and Kaushik Roy. 2020. An Investigation of Biometric Authentication in the Healthcare Environment. Array 8, (2020), 100042. DOI: https://doi.org/10.1016/j.array.2020.100042
2. Mousse, Mikaël & Motamed, Cina & Ezin, Eugène. (2017). Percentage of human-occupied areas for fall detection from two views. The Visual Computer. 33. 10.1007/s00371-016-1296-y
3. Joseph M. Ackerson, Rushit Dave, and Naeem Seliya. 2021. Applications of Recurrent Neural Network for Biometric Authentication & Anomaly Detection. Information 12, 7 (2021), 272. DOI:https://doi.org/10.3390/info12070272
4. Parker SJ, Strath SJ, Swartz AM. Physical activity measurement in older adults: relationships with mental health. J Aging Phys Act. 2008;16(4):369-380. doi:10.1123/japa.16.4.369
5. Nyle Siddiqui, Rushit Dave, and Naeem Seliya. 2021. Continuous User Authentication Using Mouse Dynamics, Machine Learning, and Minecraft. 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET) (2021). DOI: https://doi.org/10.48550/arXiv.2110.11080
6. Sam Strecker, Willem Van Haaften, and Rushit Dave. 2021. An Analysis of IoT Cyber Security Driven by Machine Learning. Algorithms for Intelligent Systems (2021), 725-753. DOI: https://doi.org/10.1007/978-981-16-3246-4_55
7. Wang, H., Zhang, D., Wang, Y., Ma, J., Wang, Y., & Li, S. (2017). RT-Fall: A Real-Time and Contactless Fall Detection System with Commodity WiFi Devices. IEEE Transactions on Mobile Computing, 16, 511-526.
8. L. Chen, J. Hoey, C. D. Nugent, D. J. Cook and Z. Yu, "Sensor-Based Activity Recognition," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no.

6, pp. 790-808, Nov. 2012, doi: 10.1109/TSMCC.2012.2198883.

9. Liu AA, Xu N, Su YT, Lin H, Hao T, Yang ZX. Single/multi-view human action recognition via regularized multi-task learning. Neurocomputing. 2015; 151:544–553. https://doi.org/10.1016/j.neucom.2014.04. 090
10. Chaaraoui AA, Climent-Pe´rez P, Flo´rez-Revuelta F. Silhouette-based human action recognition using sequences of key poses. Pattern Recognition Letters. 2013; 34(15):1799–1807. https://doi.org/10.1016/j.patrec.2013.01.021