https://doi.org/10.48047/AFJBS.6.14.2024.7704-7719

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Long Range Context Network based Deep Learning Model for Recognition of Human Activity.

BANUSHRI S , Dr.JAGADEESHA R

Research Scholar, Computer Science and Engineering
Impact College of Engineering and Applied Sciences, Bangalore.
Visvesveraya Technological University,Belgavi-590018
banushri914@gmail.com

Professor, Computer Science and Engineering
IMPACT College of Engineering and Applied Sciences, Bangalore.
Visvesveraya Technological University,Belgavi-590018
Jagdish.mtech@gmail.com

Volume 6, Issue 14, Aug 2024

Received: 15 June 2024

Accepted: 25 July 2024

Published: 15 Aug 2024

doi: 10.48047/AFJBS.6.14.2024.7704-7719

Abstract

Several studies have been led on video footage using AI deep learning methods. Human activity recognition (HAR), object localization, and event recognition are some of the most common applications. Among these, HAR stands out as a particularly difficult task and a major focus topic for future study in video data processing. Many AI-based models have been created for activity identification. Still, their poor presentation of real-world long-term HAR can be attributed to their inability to extract spatial and temporal data. To improve the system's prediction skills, the authors of this study suggest a hybrid deep learning perfect called the Long-Range Context Network (LRCN), which combines a convolutional neural network with a long-short-term memory network (CNN-LSTM) and is driven by a self-attention algorithm. The transformer's built-in self-attention mechanism can compete with high-end convolutional neural networks with long short-term memory by expressing unique dependencies among signal levels within a time series. The human activity recognition process is then finalized by feeding the collected features into the SoftMax layer. Experiments determine the best values for the model's parameters. To find the best option for HAR, a thorough ablation study is achieved over many classic machine learning and deep learning models. A few of the measures examined in this experimental study of the publicly accessible modified UCF50 dataset (UCF50mini). Using the LRCN method, a 99% accuracy is attained, proving the proposed model's viability for HAR use.

Keywords: Long-Range Context Network; Human activity recognition; Convolutional network; Long-short term network; Event Recognition.

Introduction

Smart homes, fall finding for the elderly, rehabilitation, and misbehavior identification are just some of the areas where HAR has been getting a lot of attention in recent years [1]. The fall behavior of elderly persons living alone, for instance, can be recognized by tracking their movements to alert loved ones promptly. Science-based exercise and fitness management are within the reach of every fitness enthusiast who counts steps and recognizes their workout

status [2]. Gait analysis is useful in the diagnosis of knee problems. Data on patients' mobility throughout rehabilitation can be used to fine-tune treatment strategies for conditions affecting the lower limbs. It is possible to classify HAR technology as either camera-based or sensor-based [3]. The camera-based technique [4] uses a camera set up in a human environment to extract human activity elements from the video stream. This method has the potential to graphically represent the intricacies of human behavior, but it has privacy concerns and is sensitive to lighting conditions. Wearable Sensor-Based Activity Recognition uses sensor network technologies to monitor a person's behavior and environment. In this case, there are sensors attached to humans. However, wearable sensors can only detect the activity of the person carrying them and do not support multiple-person detection. Most wearable sensors are unsuitable for real applications due to their size or battery life. However, the HAR system built with wearable sensors has certain people in performing daily activities such as discomfort in wearing sensors to their body parts for a long time, elderly people often forgetting to wear proper suit types of equipment of wearable sensors, directional control issues cause unreliable data recording during complex subject's movements (i.e., especially in smartphones). There is a relative difficulty in terms of energy consumption and device settings. Therefore, camera-based methods are preferable for recognizing human activities [5]. In this study, we emphasize primarily the issue of HAR based on sensors

Human activity detection using mobile devices holds considerable promise in the medical field, where it may be used to keep tabs on patients with a wide range of conditions, ensure they are following their treatment plans, and even deter them from engaging in harmful behaviors [6]. interaction, robotics, and sports are just a few of the many potential applications [7] of this cutting-edge technology. The basic goalmouth of human activity recognition [8] is to properly represent human behaviors and their relationships based on a data sequence that has never been seen before**.** Recognizing human actions in video data is difficult because of issues including motion blur and poor video quality. For example, when comparing human activity identification methods, there are two primary concerns: "Which action is performed?" falls recognition job, and "Where exactly in the video?" falls under the localization task [9]. Images in a series are called frames. For this reason, the fundamental goal of an action recognition challenge is to analyze the provided video clips to identify the next human activities.

The actions of humans are reflections of their routines, making it difficult to identify them as such. Another difficult problem [10] is creating a deep learning-based perfect that can anticipate human activity within sufficient benchmark datasets for assessment. The massive success of the ImageNet dataset in the arena of image processing has led to the publication of many benchmark action recognition datasets for researchers to explore. Similarly, if we consider video data processing in comparison to image processing, we find that training the deep learning model necessitates massive computational capacity and a huge number of input parameters [11]. Human activity recognition using deep neural networks is an area of intense research and development. Time series classification of sensor signals typically employs several deep neural network architectures, [12], fully convolutional neural networks [13], multiscale CNN [14], time-LeNet [15], stacked autoencoder [16], deep belief neural networks [17], LSTM deep recurrent neural network [18], echo state networks [19], and inception Res Net [13]. Deep learning (DL) has replaced classical machine learning for activity identification [14] because of its ability to analyze pre-processed IMU data without the need to extract hand-

crafted features. There have been several suggested CNN-based HAR approaches that automatically retrieved characteristics [15] in recent years.

Nonetheless, activity recognition is a time series classification challenge. Improving the model's performance is challenging due to CNN's inability to easily extract the long-term dependency inside the time series [16]. To address this issue, HAR has made extensive use of the Long Short-Term Memory (LSTM) network for its ability to efficiently extract long-term dependency within time series. This paper suggests a deep convolutional neural network (CNN-LSTM) with self-attention to solve the activity categorization issue. A CNN layer is used to extract spatial data, and then an LSTM network is used to learn temporal information, in this study's hybrid method of recognizing physical activity. The model compared several machine learning and deep learning representations to regulate which was the most real at recognizing activities.

The outstanding sections of the paper are as shadows: The literature review is obtainable in Section 2, the materials and procedures employed in this investigation are described in Section 3, the experimental assessment is discussed in Section 4, and the results are summarized in Section 5.

2. Related works

For HAR, Li et al. [17] offer a sensor status frequency-based sensor data contribution significance analysis (CSA) approach. This technique is used to quantify the impact of a certain sensor on a given behavior recognition task. We then construct a context-aware, noise-reduction spatial distance matrix by mapping out the locations of ambient sensors. Finally, we present a HAR algorithm for daily behavior recognition based on a wide time domain (HAR_WCNN). Experiments conducted on the CASAS dataset reveal that the proposed HAR_WCNN achieves better approaches associated with it in terms of HAR accuracy and time required to complete the task.

The residual recurrent neural network (RCNN-Bi GRU) was used in the development work of Helmi et al. [18]. We create new feature selection methods using the MPA to help us choose the best features to include in our models. To guarantee the quality of the suggested MPA variations' performance, we compare them in depth to several optimization methods using a variety of evaluation indicators and do statistical tests. We find that MPA performed the best out of all the MPA variations and other approaches we examined.

Using a multi-head Convolutional multi-source sensor information, Islam et al. [19] propose a multi-level feature fusion procedure for recognition. Three different types of CNN and a convolutional bias adjustment model (CBAM) are used to assess and extract feature dimensions. For effective activity recognition, the Conv LSTM network is tailored to extract temporal characteristics from time-series data collected by several sensors. Experiments and assessments are conducted using a HAR dataset known as the UP-Fall detection dataset to gauge the efficacy of the created fusion construction. Lastly, we implemented an IoT infrastructure to put the suggested fusion network through its paces in practical smart healthcare application settings. Experiment results show that the created multimodal HAR outline outdoes the current state-of-the-art approaches in a variety of ways.

Using a combination of a CNN and a Gated Recurrent Unit (GRU), Dua et al. [20] present a new classifier called "ICG Net" for HAR. It applies convolutional filters of varying

sizes concurrently to the input, allowing it to learn from data at a variety of scales. Convolutional networks with many filters of the same size are used to calculate more generalized features for small regions of data. The model additionally employs 11 convolutions to pool data across channel dimensions; this is done on the hunch that it would aid in revealing previously concealed information. By combining CNN and GRU, the proposed ICG Net is better able to detect both short series. It's an end-to-end paradigm for HAR that doesn't require any human intervention in the form of feature engineering to interpret raw data from wearable sensors. Human-Computer Interaction (HCI) disciplines surveillance can all benefit from the proposed HAR system's integration of adaptive user interfaces. Accuracy levels of 99.25% and 97.64% were attained on the MHEALTH and PAMAP2 benchmark datasets, correspondingly.

To make the most of the temporal and spatial info in time-series data, Mim et al. [21] presented a model called Gated Recurrent Unit-Inception (GRU-INC), which is an Inception-Attention-based technique employing GRU. On publicly accessible datasets including UCI-HAR, OPPORTUNITY, PAMAP2, WISDM, and Daphnet, the suggested model was achieved respectively. When it came to the model's temporal components, GRU and Attention Mechanism (AM) were used, while the spatial components benefited from Inception and Convolutional Block Attention Module (CBAM). The suggested design was tested against industry standards and precedents. An improved recognition rate at a reduced computational cost has been demonstrated for the GRU-INC model. Therefore, our paradigm may be used in clinical and therapeutic settings where activities are included.

With the Multi CNN-Filter LSTM model proposed by Park et al. [22], feature vectors are effectively processed hierarchically by combining a CNN with an LSTM via a residual connection. To this purpose, we suggest a unique way of employing LSTM cells—which we refer to as filter-wise LSTM, or Filter LSTM—so that the HAR model may discover the interdependencies between characteristics across multiple tiers of the hierarchy. The suggested HAR model has undergone stringent testing, with positive results, on two open-source datasets. In contrast to state-of-the-art models, the suggested HAR model improves accuracy by 2.3%-4.4% while using 21%-70% less operations. For additional deployment study, the suggested model is also installed on a Raspberry Pi 4. To prove that the suggested HAR model is greater to the state-of-the-art models in terms of linkages, and to provide substantial insights into the efficacy of the system, experimental findings are provided.

A unique learning approach to semi-supervised HAR is proposed by Qu et al. [23]. To combat the overfitting of a single network, we first design a semi-supervised mutual learning approach. The primary and secondary networks in this architecture share supervised training data. Second, to prevent the distribution from deviating, the distribution-preserving loss is developed. This loss minimizes the difference in the distribution of labelled data. Finally, a context-aware aggregation module is used to incorporate the adjacent sequences' contextual information. This component can glean more nuanced data from a wider variety of sequences. Our approach has been tested on four distinct human activity identification datasets that have been publicly released. The experimental consequences validate that the suggested procedure outperforms four benchmark semi-supervised HAR methods.

2.1. Problem statement

Many deep learning-based HAR models have been projected because of the widespread use of HAR in fields like sports and medicine. Extraction of geographical and temporal

information from data is often overlooked in existing models. To solve these problems, the suggested model utilizes the Long-Range Context Network, which is comprised of convolutional layers for LSTM layers for collecting temporal patterns, allowing it to successfully deal with sequential picture input.

3. Proposed scheme

In this section, the activity of humans is detected from the publicly available dataset, and the whole process is explained in Figure 1.

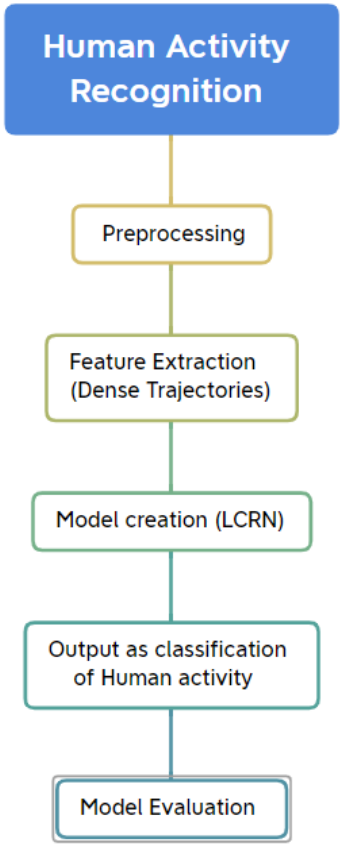

Figure 1. Projected model.

3.1. UCF50 Dataset

There were 50 distinct behavioral categories in the UCF50 dataset. The dataset included authentic YouTube videos included. Camera movement, clutter, lighting, etc. all varied greatly across the sample. Some of the videos in a set may feature the same person, have the same setting, or have the same point of view. In Figure 2 we show the UCF50 data set.

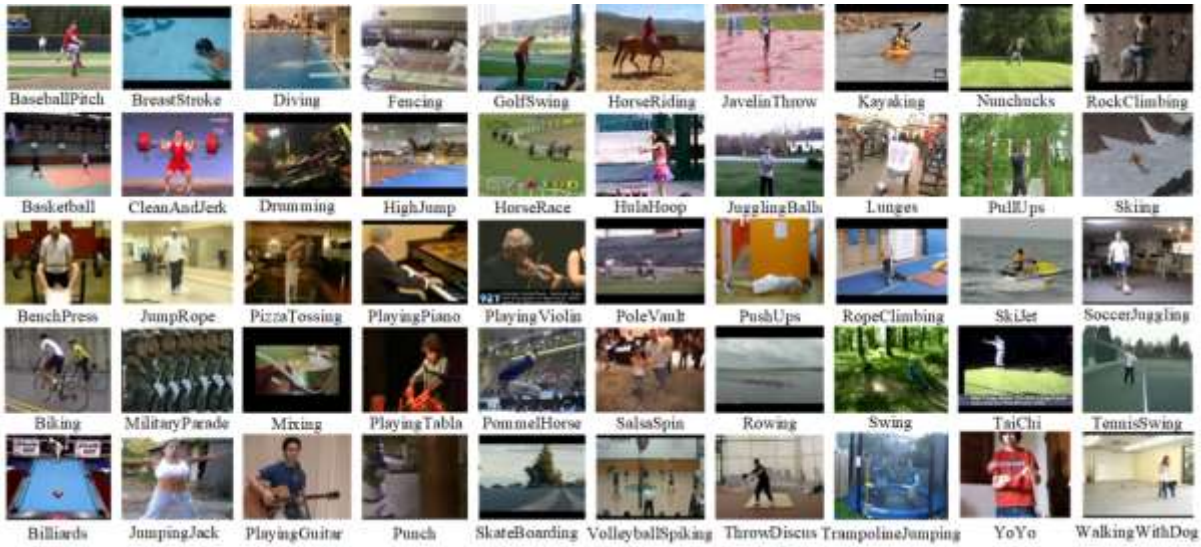

Figure 2. Example of the UCF50 dataset [24].

The fifty classes that made up the UCF50 dataset. The dataset includes events like a baseball pitch, a basketball shot, a billiards break, a clean and jerk, a dive, a drum solo, a fencing stroke, a golf swing, a guitar solo, a horse race, a horseback ride, a high jump, a horseback ride, a hula hoop, a javelin a jump rope, a jumping jack, a kayak Indoor rock climbing, rowing, salsa spins, skateboarding, skiing, ski jet, juggling soccer balls, swing dancing, playing the dog, and yo-yo is all in the list of 24 activities. The dataset was partitioned into 10 layers, and both datasets were utilized to train the network and evaluate its performance. We have designated the smaller data set as Mini-UCF50. Training, test, and validation sets were created from both the reduced and full datasets.

3.2. Data Pre-processing

For every application that relies on video, the frame extraction process is a necessary first step in the processing pipeline. To interpret video, machine learning algorithms need to see each frame individually. First, we read each video file in the dataset's class directories using the OpenCV Video Capture method, then we extract the video frames, and finally, we resize each video frame to a fixed size (64x64 pixels) and normalize its pixel values by dividing them by 255 [25]. Each class's pre-processed frames are subsequently appended to a temporary list. After that, we'll run a script that will pull still images from every action video ever made. Normalizing the pixel values to a range from 0 to 1 and resizing the images to a consistent height and width facilitates training and improves model performance and accuracy.

3.3. Feature Extraction

Dense trajectory refers to a concept in computer vision and video analysis where motion information is captured densely from a video sequence. Unlike sparse optical flow, which tracks motion only at specific interest points, dense trajectories are computed at every pixel or a dense grid of points in a video frame. Dense trajectory extraction involves the following steps:

A) Dense Sampling includes the points that are densely sampled from video frames. These points are not necessarily distinct features but can be regular grid points.

B) Feature Point Tracking: Each point is tracked across subsequent frames, capturing its motion trajectory over time.

C) Descriptor Extraction: For each tracked trajectory, local descriptors are extracted, providing information about the appearance and motion characteristics along the trajectory. These descriptors are often LBP representations.

In our experiment, we use the HOG [26] to extract characteristics of the region of interest. The HOG feature requires an RGB picture. This is why we've layered the RGB image on top of the clean foreground. This is where the RGB picture of the foreground comes from. The image is reduced in size to 128 by 64 pixels at the outset.

$$\text{Gradient magnitude, M} = \sqrt{S_x^2 + S_y^2} \quad (1)$$

$$Gradient\ orientation, \Theta = arctan\left(\frac{S_y}{S_x}\right) \quad (2)$$

The gradient of the layered picture is then calculated in the horizontal and vertical directions using a 1-dimensional kernel. These kernels are represented by the notations [-1 0 1] and [-1 0] of the gradient are determined to be 3 and 4, correspondingly. Here, the value, S_x, and the y-direction gradient value, S_y, are denoted.

The cells that make up the resizable window are 8 pixels by themselves. A histogram is calculated from each cell by dividing each bin 20 degrees apart. In this configuration, blocks of four cells overlap by a factor of the half. After the histograms of four blocks are combined, the vector is normalized such that it is independent of the contrast between the blocks. In the end, 3780 feature vectors are compiled from the block histograms.

The local edge shape of humans may be described with the help of the HOG features descriptor. However, human body texture information is also required for precise and effective human detection. Our suggested approach uses consistent LBP [27] to describe textures. With LBP, a picture becomes a picture of integer labels that characterize the picture's textures. It is thought that a pattern's strength and the strength of the texture as a whole are complimentary. At first, the value of each neighbouring pixel is set to a threshold relative to the block's central pixel. The calculated value is multiplied to convert from binary to the decimal of the center pixel. Then, a central pixel's label is calculated by adding the values around it. Using this method, 256 unique labels may be generated. There are 58 uniform models, and by treating all other models as a single category, we may obtain a 59-D vector. The formula for determining the values is (3).

$$LBP_{P,R} = \sum_{P=0}^{P-1} S(g_p - g_c).\, 2^p \quad (3)$$

Where, g_p is neighborhood block, g_c is center pixel charge, p is sampling radius. If the centre pixel's rate is better than the neighbor's charge, then the threshold function t(x) of (6) will be 0. Then, it will be 1. In this generated.

$$t(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \quad (4)$$

If U in (5) is less than or equal to 2, the pattern is considered uniform; otherwise, it is considered non-uniform. Each block's 59-dimensional properties are derived from the LBP features of its four individual cells. After that, the photos are combined with HOG feature vectors and stored in a matrix format, with 3839 features vectors per image.

$$U(LBP_{P,R}) = \left|s(g_{p-1} - g_c) - s(g_p - g_c)\sum_{p=0}^{p-1} s(g_p - g_c) - s(g_{p-1} - g_c)\right| \quad (5)$$

D) Feature Representation: Trajectories and their descriptors can be used as features for numerous requests, such as action recognition, object tracking, or behavior analysis in videos.

Dense trajectories capture detailed motion information, making them valuable for action recognition tasks where fine-grained motion patterns are essential for accurate classification. They enable the analysis of complex and subtle motions in videos, contributing to the development of sophisticated computer vision algorithms.

3.4. Classification using Long-Range Context Network (LRCN)

A 16-filter convolutional neural network receives the pre-processed input data before being followed by a batch rate. A 64-neuron information is then fed the output. Second, the self-attention layer has received the LSTM layer's output. The primary function of the attention layer is to zero in on a certain layer of the network. Similar numbers of neurons and a similar dropout rate were used in the second LSTM. In order to activate, a ‘sigmoid' function is used. Finally, the output of the perfect is obtained via an output layer (a dense layer equipped with a ‘SoftMax' classifier). When compared to Adagrad, RMSprop, and SGD, the Adam optimizer came out on top. By iteratively testing all of the limits and the binary and cross-entropy, the optimal one was determined.

3.4.1. Model implementation

As can be seen in Figure 3, a dense network makes up the human activity recognition network [28]. An attention model, CNN, long short-term memories, and residual concatenation are used to detect the subject's actions.

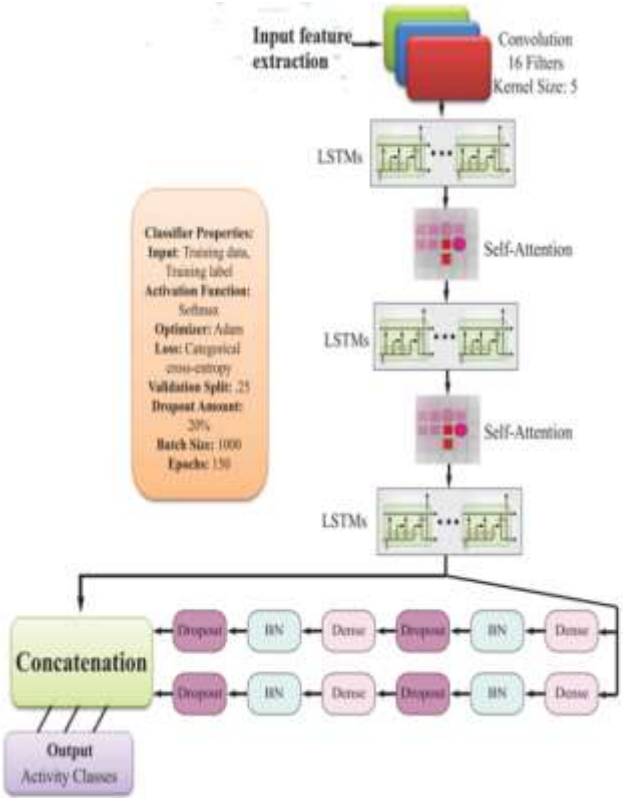

Figure 3: Architecture of the proposed model

When applied to real-world challenges in the field of computer vision, CNNs have shown remarkable success over the past two decades. The normalizing, and fully connected layers that CNNs use to extract from the input data are crucial to their success in solving computer vision issues [29]. Typically, an input picture is sent through a convolutional layer, which then generates a set of 2D feature maps with spatial info. The layer then just applies down-sampling operations to these feature maps to reduce their size. Whereas a convolutional layer is represented mathematically as follows.

$$C_{i,j,k}^{l} = f\left(\left(w_k^l\right)^T x_{i,j}^l\right) + b_k^l \quad (6)$$

where b_k^l is kth, $x_{i,j}^l$ characterizes the input region in the 1st layer. To improve activation accuracy, a normalizing layer is often utilized before the activation function. The recovered feature maps are converted to the 1D feature vectors layer, and then the computed list probabilities are sent on to the layer or output layer. The work suggests a manner for efficient categorization of preset indoor activities under the observation test with varying circumstances.

3.4.2. Temporal Extraction using LSTM

This model is suitable for tasks where the input data consists of sequences of images (videos or image sequences) and the goal is to classify these sequences into diverse classes represented as "WalkingWithDog", "Taichi", "Swing", "Horserace". The model uses both convolutional layers for LSTM layers for capturing temporal patterns, making it capable of handling sequential image data effectively.

Each gate (input, forget, and output) in LSTM's internal architecture forwards it to the next gate, regulating the information movement near the ultimate output [30]. Standard units of the RNN (a) and LSTM (b) are shown in Figure 4.

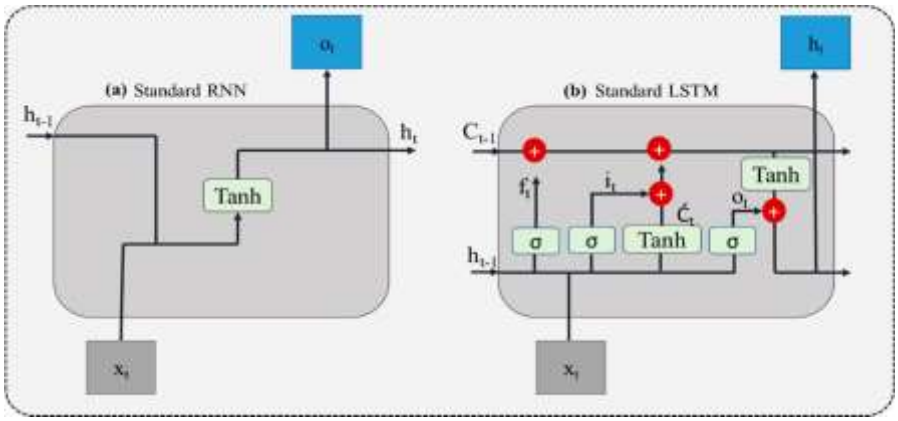

Figure 4. (a) Characterizes the typical RNN unit, (b) characterizes the typical LSTM unit.

The input gate, which is responsible for updating the information, is often regulated by a sigmoid function, however this is not always the case. When necessary, the forget gate will delete data from the current state C_t as well as the state of the preceding cell C_(t-1). When the output is complete, the output gate ot sends it on to the next LSTM unit, where it is stored until it is needed for a subsequent sequence prediction. In contrast, the tanh activation function is used by the recurrent unit C_t to estimate the state of the previous cell C_(t-1) and the current. In contrast, ht may be calculated by multiplying o_t by the tangent of C_t, a scalar. Finally, the

softmax classifier may be used to get the final result. The following mathematical expression describes how the aforementioned gates work.:

$$f_t = \Phi\left(\widehat{W}_f.[h_{t-1}, x_t] + B_f\right) \quad (7)$$

$$i_t = \Phi\left(\widehat{W}_i.[h_{t-1}, x_t] + B_i\right) \quad (8)$$

$$C_t = tanh\left(\widehat{W}_C.[h_{t-1}, x_t] + B_C\right) \quad (9)$$

$$o_t = \Phi\left(\widehat{W}_o.[h_{t-1}, x_t] + B_o\right) \quad (10)$$

$$h_t = o_t x tanh(\Phi(C_t)) \quad (11)$$

$$Output = softmax(h_t) \quad (12)$$

3.4.3. Proposed Model

The study suggests a hybrid strategy in which characteristics are retrieved from the first model's layers and sent on to a second model for further learning and modelling. Researchers began taking a closer look to 1D CNN because of its aptitude to extract useful spatial and data. Nonetheless, several studies have employed LSTM to demonstrate its effectiveness with sequential and time-series data. We combine the 1D CNN and LSTM to first extract features, which are subsequently sent to the LSTM for further modelling and learning. The filter size varies between 64 and 128 in the first and second layers of a 1D CNN. The filter size is the only difference between the two layers; the kernel size is 3 and the ReLU function is utilized for both levels. The Max pooling layer comes afterward the first two and has a pool size of 2. Two LSTM layers, each with 64-bit cells like those used in the CNN layers, process these characteristics. Next comes a flatten layer, then a dense layer using a SoftMax function, and finally the LSTM layer. Adam, an optimizer with a 0.0001 percent learning rate, is employed in this method. Table 1 provides the suggested model's parameter values.

Table 1. Strictures setting of our projected model.

Layer (Kind)	No. of Param.	**Kernel Size**	**Filter Size**
Flatten	-	-	-
Dense (-12)	780	-	-
1D CNN Layer -1	9664	3	64
1D CNN Layer -2	24,704	3	128
Max Pooling -1D	-	-	-
LSTM-(64)	46,408	-	-
LSTM-(64)	33,024	-	-
Total parameters	117,580	-	-

4. Results and Discussion

To assess the efficacy of models on a variety of data sets, we conduct a series of experiments in this section. In the implementation of this research, Python is used utilizing the Kera's framework and the backend libraries TensorFlow and Scikit-learn. Using machine learning classifiers and deep learning representations, five distinct experiments are conducted on data sets consisting of 30 frames, 60 frame, 120 frame, and 150 frame sequences. In the experimental evaluation we used 1000 dataset image for test and train the model to attained below performance measures. Figure 5 demonstrations the suggested representation's confusion matrix.

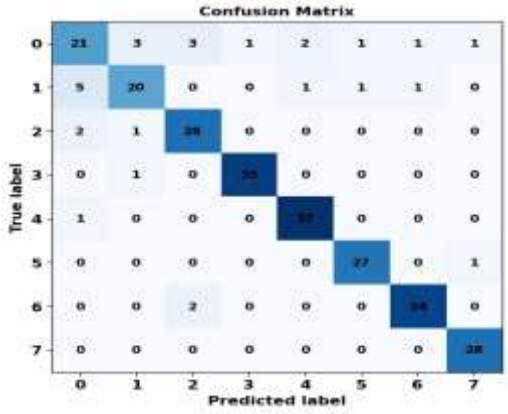

Figure 5: Confusion Matrix of the projected perfect

Figure 6 and 7 presents the loss and accuracy of the projected model data.

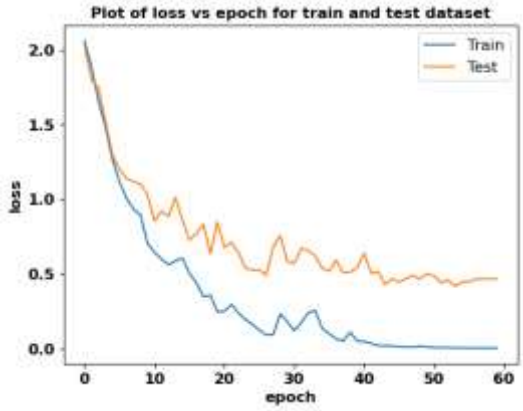

Figure 6: Loss of the projected perfect for data

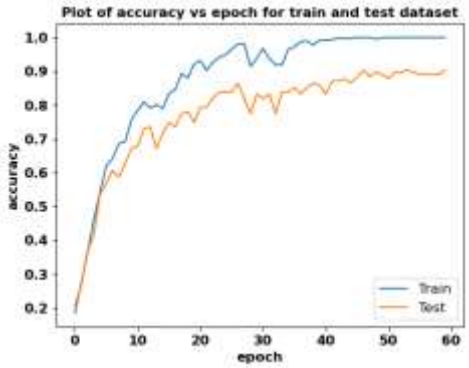

Figure 7: Accuracy of the projected model for training and testing data

The ROC for the whole data is mentioned in Figure 8.

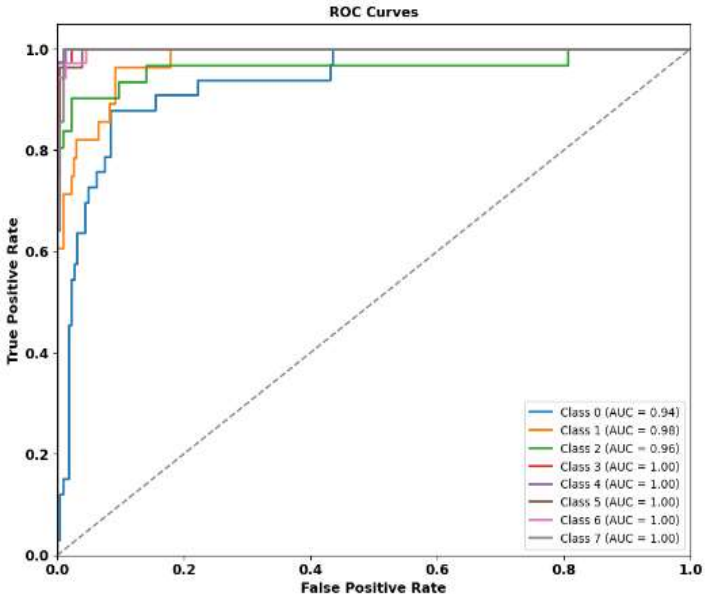

Figure 8: ROC of the projected perfect

4.1. Performance metrics

This paper calculated four main analytical metrics: $true\ positive\ (TP)$, $true\ negative\ (TN)$, $false\ positive\ (FP)$, $and\ false\ negative\ (FN)$ to assess the efficacy of the classification system created using the datasets.

When evaluating the efficacy of a classification model, ACC is described as the ratio of correct assumptions to total assumptions made:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (13)$$

PR, also known as positive prognostic value, measures the proportion of appropriately detected positive examples to all positive examples:

$$Precision = \frac{TP}{TP+FP} \quad (14)$$

RC, also mentioned to as sensitivity or the tp, is the percentage of properly classified positive gears among all positive instances.

$$Recall = \frac{TP}{TP+FP} \quad (15)$$

The F1 is an integrated metric that incorporates PR and RC into a single numerical value:

$$F1 = \frac{Precision*Recall}{Precision+Recall} \quad (16)$$

Table 2: Proportional investigation of predictable perfect with existing representations

Model	ACC (%)	RC (%)	PR (%)	F1 (%)
LR	94.4	92.5	93.8	94.2
MLP	93.8	94.6	94.7	95.8
AE	96.5	96.3	95.4	96.4
DBN	97.7	96.5	96.6	97.7
CNN	97.5	97.6	97.4	98.2
LSTM	98.8	98.3	98.5	98.9
Proposed model	99.8	99.6	99.5	99.8

Table 2 characterizes the Proportional investigation of predictable perfection with existing representations. In the analysis of the LR model the ACC value as 94.4 the RC rate was 92.5 also PR rate was 93.8 and lastly F1 score was 94.2 respectively. After that the MLP model attained the ACC value of 93.8 then the RC rate of 94.6 and also PR rate of 94.8 and lastly F1 score of 95.8 correspondingly. After that the AE model attained the ACC value as 96.8 and then the RC rate of 96.3 and also PR rate of 95 .4 and lastly F1 score as 96.4 correspondingly. After that the DBN perfect attained the ACC value as 97.7 and then RC rate as 96.5 and also PR rate of 96.6 and lastly F1 score as 97.7 correspondingly. After that the CNN model attained the ACC value of 97.5 and then RC rate as 97.6 also PR rate as 97.4 and lastly F1 score of 98.2 correspondingly. After that the LSTM model attained the ACC value as 98.8 and then the RC rate of 98.3 also PR rate of 98.5 and lastly F1 score as 98.9 respectively. After that the Proposed model attained the ACC value as 99.8 then the RC rate of 99.6 also PR rate as 99.5 and lastly F1 score of 99.8 correspondingly.

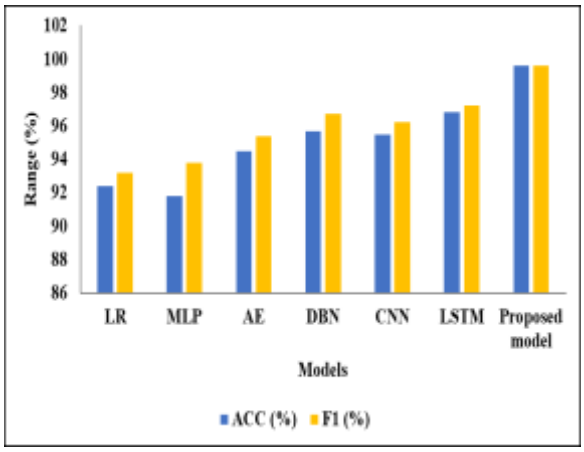

Figure 9: Graphical Description of the proposed model

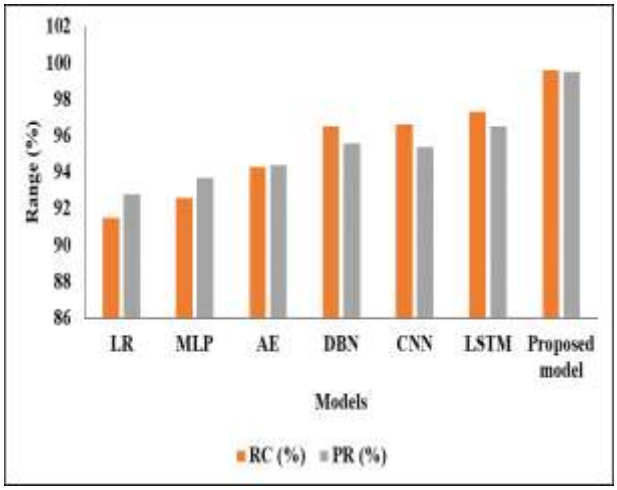

Figure 10: Analysis of different models for HAR

5. Conclusion

Recognizing human behavior from visual sensor data has been a major research focus for decades. This study demonstrates how efficiently these features may be extracted using the automated feature engine included in CNNs and LSTMs. The suggested model was evaluated on UCF50mini, where it showed correctness of 99.6%, recall of 99.6%, precision of 99.5%, and F1 of 99.6% in the activity identification scenario. The suggested approach outperformed statistical machine learning-based methods in terms of both reliability and likelihood of human activity detection. When compared to previous networks developed specifically for video-based human behavior categorization, the suggested architecture performed admirably. It is nevertheless difficult to keep an eye on public spaces like parks and plazas for signs of suspicious activity. First, moving forward, we'll keep adding participants and fine-tuning the network topology to ensure the quality of our dataset. It also emphasizes the creation of tracking systems that may be used with wearable devices and mobile phones. As a result, the established framework has promise for therapeutic usage, and the resulting data may be fruitful for future studies.

REFERENCES

[1] Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., & Liu, Y. (2021). Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. ACM Computing Surveys (CSUR), 54(4), 1-40.

[2] Gu, F., Chung, M. H., Chignell, M., Valaee, S., Zhou, B., & Liu, X. (2021). A survey on deep learning for human activity recognition. ACM Computing Surveys (CSUR), 54(8), 1-34.

[3] Ramanujam, E., Perumal, T., & Padmavathi, S. (2021). Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review. IEEE Sensors Journal, 21(12), 13029-13040.

[4] Zhang, S., Li, Y., Zhang, S., Shahabi, F., Xia, S., Deng, Y., & Alshurafa, N. (2022). Deep learning in human activity recognition with wearable sensors: A review on advances. Sensors, 22(4), 1476.

[5] Nafea, O., Abdul, W., Muhammad, G., & Alsulaiman, M. (2021). Sensor-based human activity recognition with spatio-temporal deep learning. Sensors, 21(6), 2141.

[6] Hayat, A., Morgado-Dias, F., Bhuyan, B. P., & Tomar, R. (2022). Human activity recognition for elderly people using machine and deep learning approaches. Information, 13(6), 275.

[7] Mekruksavanich, S., & Jitpattanakul, A. (2021). Biometric user identification based on human activity recognition using wearable sensors: An experiment using deep learning models. Electronics, 10(3), 308.

[8] Qiu, S., Zhao, H., Jiang, N., Wang, Z., Liu, L., An, Y., ... & Fortino, G. (2022). Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. Information Fusion, 80, 241-265.

[9] Xu, Y., & Qiu, T. T. (2021). Human activity recognition and embedded application based on convolutional neural network. Journal of Artificial Intelligence and Technology, 1(1), 51-60.

[10] Gao, W., Zhang, L., Huang, W., Min, F., He, J., & Song, A. (2021). Deep neural networks for sensor-based human activity recognition using selective kernel convolution. IEEE Transactions on Instrumentation and Measurement, 70, 1-13.

[11] Yadav, S. K., Tiwari, K., Pandey, H. M., & Akbar, S. A. (2021). A review of multimodal human activity recognition with special emphasis on classification, applications, challenges, and future directions. Knowledge-Based Systems, 223, 106970.

[12] Khan, Z. N., & Ahmad, J. (2021). Attention-induced multi-head convolutional neural network for human activity recognition. Applied soft computing, 110, 107671.

[13] Gupta, N., Gupta, S. K., Pathak, R. K., Jain, V., Rashidi, P., & Suri, J. S. (2022). Human activity recognition in artificial intelligence framework: A narrative review. Artificial intelligence review, 55(6), 4755-4808.

[14] Challa, S. K., Kumar, A., & Semwal, V. B. (2022). A multi-branch CNN-Bi LSTM model for human activity recognition using wearable sensor data. The Visual Computer, 38(12), 4095-4109.

[15] Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., & Liu, J. (2022). Human action recognition from various data modalities: A review. IEEE transactions on pattern analysis and machine intelligence.

[16] Huang, W., Zhang, L., Gao, W., Min, F., & He, J. (2021). Shallow convolutional neural networks for human activity recognition using wearable sensors. IEEE Transactions on Instrumentation and Measurement, 70, 1-11.

[17] Li, Y., Yang, G., Su, Z., Li, S., & Wang, Y. (2023). Human activity recognition based on multi-environment sensor data. Information Fusion, 91, 47-63.

[18] Helmi, A. M., Al-cases, M. A., Dahou, A., & Abd Elaziz, M. (2023). Human activity recognition using marine predators' algorithm with deep learning. Future Generation Computer Systems, 142, 340-350.

[19] Islam, M. M., Nooruddin, S., Karray, F., & Muhammad, G. (2023). Multi-level feature fusion for multimodal human activity recognition in the Internet of Healthcare Things. Information Fusion, 94, 17-31.

[20] Dua, N., Singh, S. N., Semwal, V. B., & Challa, S. K. (2023). Inception-inspired CNN-GRU hybrid network for human activity recognition. Multimedia Tools and Applications, 82(4), 5369-5403.

[21] Mim, T. R., Amatullah, M., Afreen, S., Yousuf, M. A., Uddin, S., Alyami, S. A., ... & Moni, M. A. (2023). GRU-INC: An inception-attention-based approach using GRU for human activity recognition. Expert Systems with Applications, 216, 119419.

[22] Park, H., Kim, N., Lee, G. H., & Choi, J. K. (2023). Multi CNN-Filter LSTM: Resource-efficient sensor-based human activity recognition in IoT applications. Future Generation Computer Systems, 139, 196-209.

[23] Qu, Y., Tang, Y., Yang, X., Wen, Y., & Zhang, W. (2023). Context-aware mutual learning for semi-supervised human activity recognition using wearable sensors. Expert Systems with Applications, 219, 119679.

[24] Reddy, K.K.; Shah, M. Recognizing 50 Human Action Categories of Web Videos. Mach. Vis. Appl. J. (MVAP) 2013, 24, 971–981.

[25] Mathew, S., Subramanian, A., MS, B., & Rajagopal, M. K. (2023). Human activity recognition using deep learning approaches and single frame CNN and convolutional LSTM. arXiv preprint arXiv:2304.14499.

[26] Jafari, F., & Basu, A. (2023). Saliency-Driven Hand Gesture Recognition Incorporating Histogram of Oriented Gradients (HOG) and Deep Learning. Sensors, 23(18), 7790.

[27] Lan, S., Fan, H., Hu, S., Ren, X., Liao, X., & Pan, Z. (2023). An edge-located uniform pattern recovery mechanism using statistical feature-based optimal centre pixel selection strategy for local binary pattern. Expert Systems with Applications, 221, 119763.

[28] Khatun, M. A., Yousuf, M. A., Ahmed, S., Uddin, M. Z., Alyami, S. A., Al-Ashhab, S., ... & Moni, M. A. (2022). Deep CNN-LSTM with self-attention model for human activity recognition using wearable sensor. IEEE Journal of Translational Engineering in Health and Medicine, 10, 1-16.

[29] Khan, N.; Ullah, A.; Haq, I.U.; Menon, V.G.; Baik, S.W. SD-Net: Understanding overcrowded scenes in real-time via an efficient dilated convolutional neural network. J. Real-Time Image Process. 2021, 18, 1729–1743.

[30] Ullah, F.U.M.; Khan, N.; Hussain, T.; Lee, M.Y.; Baik, S.W. Diving Deep into Short-Term Electricity Load Forecasting: Comparative Analysis and a Novel Framework. Mathematics 2021, 9, 611.