OPEN ACCESS JOURNALS
Indiana Journal of Multidisciplinary Research
Abbreviated Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.21567145

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Article

Volume-06|Issue-04|2026

A Deep Dive into Human Activity Recognition using Deep Learning: A Comparative Study on Sensor and Video Datasets

Pallavi Hallappanavar Basavaraja[1], Nandeeswar Sampigehalli Basavaraju[2], Usha H N[3], Anusha Preetham[4], Pooja Nayak S[5]

[1,5]Associate Professor, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India.
[2]Professor, Department of Computer Science & Engineering - AI&ML, AMC Engineering College, Bangalore -560083, India.
[3]Associate Professor, Department of Electronics and Communication Engineering, APS College of Engineering, Bangalore-5600116, Karnataka, India.
[4]Associate Professor, Department of Computer Science & Engineering, Dayananda Sagar College of Engineering, Bangalore-560111, Karnataka, India.

Article History
Received: 05.05.2026
Accepted: 22.06.2026
Published: 25.07.2026

Citation
Basavaraja, P. H., Basavaraju, N. S., Usha, H. N., Preetham, A. Nayak, P. S. (2026). A Deep Dive into Human Activity Recognition using Deep Learning: A Comparative Study on Sensor and Video Datasets. *Indiana Journal of Multidisciplinary Research, 6*(4), 112-117.

Abstract: At the vanguard of intelligent systems, Human Activity Recognition (HAR) facilitates effortless interaction between people and machines. This research investigates a variety of deep learning techniques to identify the most suitable architecture for real-time human activity recognition. The analysis incorporates two widely recognized datasets: the UCI HAR dataset, which provides structured motion data from smartphone-based accelerometers and gyroscopes, and a selected portion of the UCF101 dataset, comprising diverse human action video clips. The models examined include single- and double-layer configurations of Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRU), Convolutional Neural Networks (CNN), and a combined CNN-LSTM approach.
The CNN model achieved high accuracy and as well as inference speed across the video dataset and therefore was trained using a custom dataset and deployed to predict human actions in real-time using MediaPipe. Through this work, we present a robust, scalable, and deployment-friendly HAR framework, paving the way towards smart surveillance, ambient health monitoring, and fitness.

Keywords: Activity Recognition, Deep Neural Architectures, Sensor data, Video data, Real-Time Prediction

INTRODUCTION

Recognizing and classifying human activities is a noteworthy area of study. HAR systems have witnessed considerable impact in applications such as health monitoring, fitness, and smart surveillance. In healthcare, it can effectively be used to monitor elderly patients or patients with predicaments such as dementia or Alzheimer's to ensure safety and timely intervention. But the most widely used HAR systems can be seen in the area of fitness. Most smartphones these days facilitate tracking steps, the distance covered, calories burnt, and even sleep monitoring using their in-built sensors, providing users with analytics about their life.[1][10]

Numerous approaches have been used in recognition systems, and a low-cost and scalable solution would be the use of wearable sensors. The sensors are capable of providing precise, real-time motion data and can be efficiently used in scenarios that require continuous monitoring. On the other hand, video-based HAR enables access to visual information and additional environmental contexts, delivering higher accuracy while identifying human activities.

Traditional machine learning algorithms were initially used in HAR systems due to their simplicity and shorter model training duration, but they came at the expense of accuracy. Their inability to handle large datasets paved the way for the adoption of deep learning models. These models have significantly magnified HAR performance through their adaptability to grasp complex patterns directly from raw input data. In this paper, we explore and compare deep learning models ranging from single and two-layer CNNs, RNNs, LSTMs, and GRUs, to a hybrid CNN-LSTM model, using both sensor and video data.

LITERATURE REVIEW

Human activity recognition using deep learning models: A review of CNN and LSTM-based approaches

This paper predominantly used the following datasets – UCI dataset and iSPL dataset. The UCI

***Corresponding Author:** Pallavi Hallappanavar Basavaraja

dataset, consisting of 6 activities, is a larger dataset when compared to the iSPL dataset, with 3 activities. A total of 4 models were employed: a hybrid CNN-LSTM model and a dense layer variant of it, and an LSTM model with its dense layer variant. The LSTM layer showcased the least accuracy, while the hybrid CNN-LSTM model topped with its accuracy of 99.06% on iSPL dataset and 92.13% on the UCI dataset. Despite its high accuracy, this study lacked an inspection of other deep learning models.[2]

A comprehensive survey on video-based human action recognition

The paper assembled an in-depth analysis of the working of various machine learning and deep learning models across several public datasets. Several research papers were utilised as a form of data collection, and their findings were summarised in this paper. For each model, its pros and cons are highlighted for various approaches. Regardless of the vast scope, this work did not explore hybrid deep learning models. [3]

Investigation on Human Activity Recognition using Deep Learning

This is another paper that addresses the merits and demerits of deep learning models across numerous public datasets. It also identifies various research problems, where deploying a HAR system might be beneficial, such as in ATMs, elderly care, employee surveillance at the workplace, etc. However, the paper only offers a theoretical view.[4][9]

Human Activity Analysis using Machine Learning Classification Techniques

The authors primarily used 6 traditional machine learning models, Naïve Bayes, kNN, Random Forest, Logistic Regression, Neural Network, and Stochastic Gradient Descent to predict activities. Among these models, the study stated that the Neural Network and logistic regression models displayed the best accuracy compared to the other models. It also exerted emphasis on the extraction of features, which is paramount in the efficiency of the model. However, the models were restricted to a single dataset, the UCI dataset, and did not explore deep learning approaches. [5]

A Close Look into Human Activity Recognition Models using Deep Learning

This study added that LSTM, CNN, and their hybrid frameworks could achieve high accuracies for activity recognition. A lightweight CNN model achieved high accuracy in multiple datasets, and the use of Bi-directional LSTM accelerated predictions for activity recognition. Although it uses a survey-based approach, other deep learning models were not investigated. [1]

DATASET DESCRIPTION

Sensor dataset:

A popular benchmark for testing time-series classification models is the UCI Human Activity Recognition Using Smartphones dataset. It focuses on sensor-based recognition and consists of recordings from 30 subjects ranging from the ages of 19 to 48, each executing six common daily activities, that is, walking, sitting, standing, walking upstairs, walking downstairs, and lying down. A Samsung Galaxy smartphone, mounted on the waist of each subject, recorded motion data through triaxial accelerometer and gyroscope sensors.

Sensor values were collected under controlled laboratory conditions at 50 Hz sampling rate. The raw signals were cleaned using noise reduction algorithms and divided into windows using a sliding window of 2.56 seconds with an overlap of 50%, thereby collecting 128 signal values per window. For each window, 561 features were calculated using both time-domain and frequency-domain techniques, including measures like mean, standard deviation, energy, etc [6].

Video dataset:

The UCF101 dataset is another standard dataset in the field of video-based action recognition [10]. It comprises 13,320 video clips distributed across 101 different action categories, encompassing a broad spectrum of human behaviours such as sports activities (e.g., Basketball Dunk, Cricket Bowling), daily routines (e.g., Brushing Teeth, Shaving Beard), and human-object interactions (e.g., Typing, Playing Guitar). The videos are collected from YouTube, introducing natural variations in camera motion, lighting, viewpoint, and background clutter. This diversity makes UCF101 a challenging yet realistic dataset for testing the robustness of action recognition algorithms. The videos are provided in .avi format and have a spatial resolution of 320×240 pixels, with frame rates varying around 25 FPS depending on the source. Each action class has a minimum of 100 video clips. [7].

METHODOLOGY

Data Collection: This study employed two types of data sources, namely, the sensor-based dataset, UCI HAR dataset, along with the video-based data, UCF101 dataset. To support the goals of this investigation, five action categories were selected - Apply Eye Makeup, Basketball, Jumping Jack, Push Ups, and Tennis Swing – from the UCF101 dataset. In addition to this, a different video dataset was generated for deployment that was recorded against real-world conditions. These recordings were manually annotated to facilitate supervised learning.

***Corresponding Author:** Pallavi Hallappanavar Basavaraja

Preprocessing of Data: Preprocessing in the sensor database was done by normalizing the values across all features. The six specified activities labels were converted to numbers, and then these numbers were one-hot encoded to assist in classification. In contrast, in the video dataset, the videos were extracted and resized to the required resolution. After this, these frames were then grouped into sequences with the appropriate action labels.

Data Exploration: In order to better understand sensor data, t-SNE was used. It uncovered the clustering patterns within different activities to judge how distinct each class is. Furthermore, the distribution of the activity classes was visualized using bar charts. For the video dataset, bar plots were used to depict the distribution of each action. For verification, sample videos of the selected actions were displayed too.

Model Architecture: For both datasets, CNN, RNN, LSTM, and GRU models (both single-layer and two-layer configurations) models were exercised. Categorical cross entropy function and Adam optimizer was wielded to compile the model. EarlyStopping was used to prevent overfitting issues and Dropout was used wherever necessary.

Performance Measurement: The assessment of the models was done using accuracy, classification reports, and confusion matrices. The matrices effectively visualised the model's accuracy and the classification report provided a detailed report on the F1 score, precision, recall and support for all activities.

Real-Time Deployment: Depending on the above mentioned metrics, the model with the best model was selected for integration into a practical system. The dataset was tailor made for this system and MediaPipe was used to make the predictions

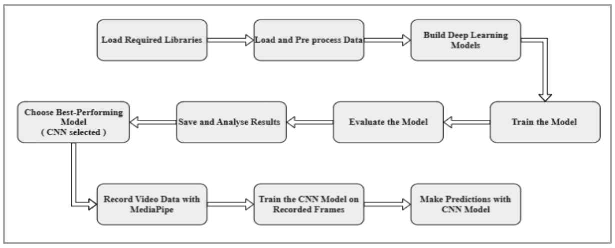

Figure 1: Architecture of HAR system

Algorithm for Human Activity Recognition (HAR)

Step 1: Import the Required Libraries

Import the essential Python libraries for deep learning, computer vision, and data processing.

```
import tensorflow as tf
import cv2
import pandas as pd
```

Step 2: Retrieve and Preprocess the Datasets

Load and preprocess the **UCI HAR** dataset and extract frames from the **UCF101** video dataset for model training.

```
X, y = load_and_segment_uci_data()
video_frames = extract_ucf_frames()
```

Step 3: Construct the Deep Learning Models

Design and initialize the deep learning architectures, including **CNN**, **LSTM**, **RNN**, **GRU**, and **CNN-LSTM**, for activity recognition.

```
cnn_model, lstm_model, rnn_model, gru_model, cnn_lstm_model = build_models()
```

Step 4: Train the Models

Train each model using the training dataset and validate its performance using the validation dataset.

```
model.fit(
   X_train,
   y_train,
   epochs=30,
   validation_data=(X_val, y_val)
)
```

Step 5: Evaluate Model Performance

Assess the trained models on the testing dataset using standard evaluation metrics.

```
model.evaluate(X_test, y_test)
```

Step 6: Save and Visualize the Results

Store the trained model and generate performance visualizations such as accuracy and loss curves.

```
save_results(model, output_dir='results/')
visualize_results(model)
```

Step 7: Select the Best-Performing Model

Compare the performance of all models and select the one with the highest accuracy. In this study, the **CNN model** achieved the best performance.

```
best_model = cnn_model
```

Step 8: Capture Live Video Data

Access the webcam and capture a frame for real-time activity recognition.

```
cap = cv2.VideoCapture(0)
ret, frame = cap.read()
```

Step 9: Fine-Tune the CNN Model (Optional)

If required, further train or fine-tune the CNN model using the newly captured data.

```
train_cnn_model(frame, cnn_model)
```

Step 10: Perform Real-Time Activity Prediction

Use the trained CNN model to predict the activity from the captured frame and display the predicted result.

```
prediction = cnn_model.predict(frame)
print(f"Predicted Activity: {prediction}")
```

***Corresponding Author:** Pallavi Hallappanavar Basavaraja

EXPLORATORY DATA ANALYSIS

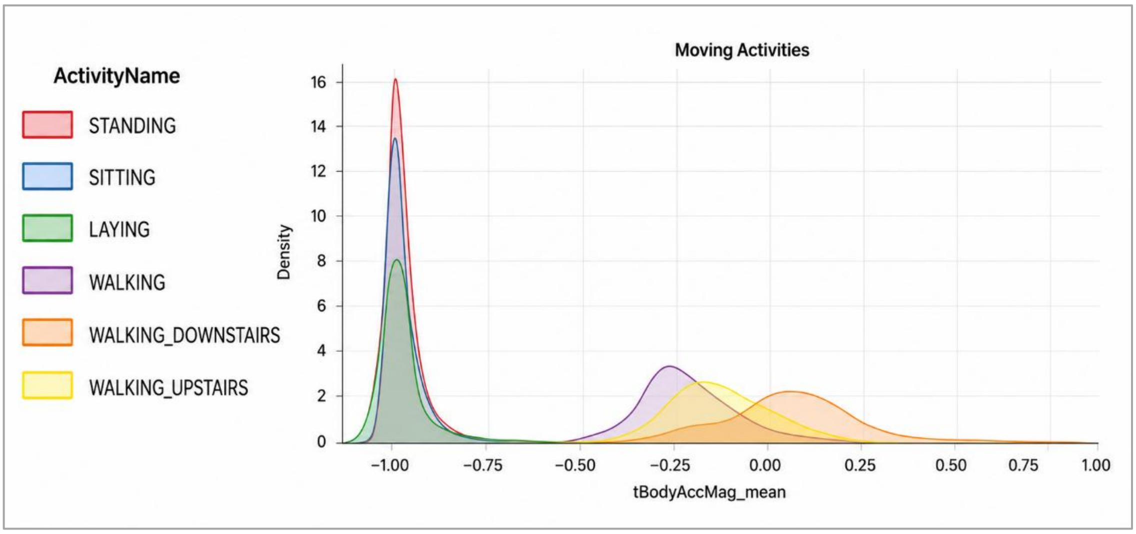

Figure 2: Density Distribution of Stationary and Non - Stationary Activities

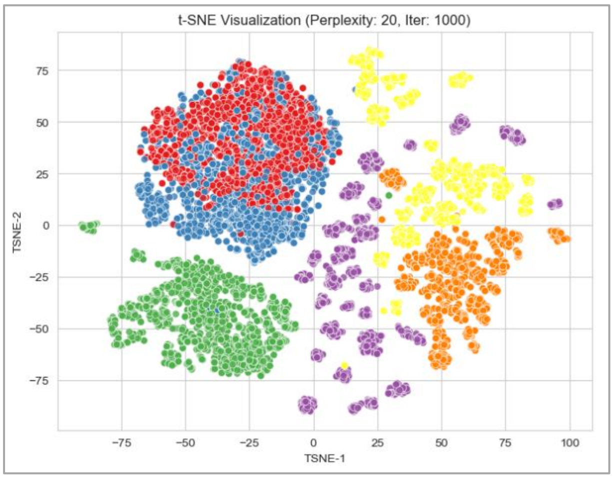

Figure 3: t-SNE Visualization of Human Activities (Perplexity = 20)

Figure 2 depicts the density variations of the feature, tBodyAccMag_mean, across different human activities from the UCI HAR dataset. The stationary activities (Sitting, Standing, and Laying) are seen towards the left and moving activities (Walking, Walking Upstairs, and Walking Downstairs) on the right. We can deduce that the motionless activities have comparable profiles due to the overlapping curves, and it becomes harder to predict Laying, Sitting, and Standing.

In the same dataset, t-SNE image (Figure 3), with perplexity value of 20, highlighted an overlap in the Sitting (blue clusters) and Standing activities (red clusters). In other words, this overlap hinders the model's ability to effectively predict them.

RESULTS AND DISCUSSION

Table 1: Training and test accuracies across both datasets

Model	UCI HAR Dataset		UCF101 Dataset	
	Training Accuracy	Test Accuracy	Training Accuracy	Test Accuracy
Single Layer LSTM	95.76%	90.46%	80.03%	80.39%
Two Layer LSTM	96.18%	91.21%	80.03%	78.43%
Single Layer RNN	71.10%	66.07%	84.30%	90.20%
Two Layer RNN	53.29%	51.88%	85.67%	82.35%
Single Layer GRU	95.97%	92.77%	81.74%	90.20%
Two Layer GRU	95.80%	91.14%	92.49%	83.33%
Single Layer CNN	98.61%	90.91%	98.63%	94.12%
Two Layer CNN	95.69%	91.92%	97.27%	89.22%
Hybrid LSTM-CNN	95.51%	91.41%	80.89%	76.47%

Among the evaluated models, GRU model achieved the highest test accuracy in UCI (92.77%) dataset and the CNN model in UCF101 (94.12%). It is interesting to note that the performance of two-layer variants of most models was diminished when compared to its predecessor, except the CNN and LSTM models in the UCI dataset. Contrary to the assumptions made, the hybrid model did not observe a high accuracy, specifically in the UCF dataset. Overall, it can be stated that the GRU model is suitable to be deployed on a sensor-based approach to a HAR system, owing to its accuracy and lightweight structure. The CNN model, on

***Corresponding Author:** Pallavi Hallappanavar Basavaraja

the other hand, were most suited to be employed in a video-based approach due to their ability to handle visual data.

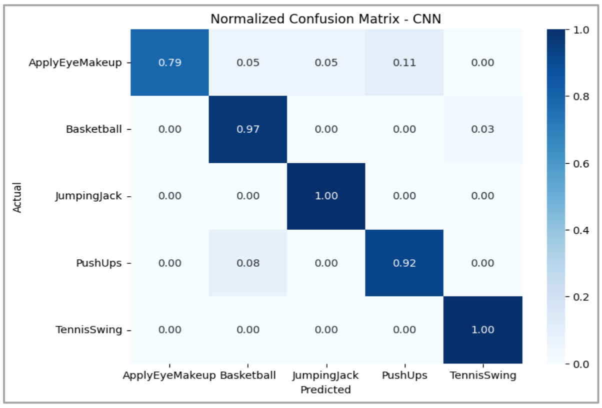

Figure 4: Normalised Confusion Matrix – CNN Model on UCF101 (5-Class Subset)

The CNN model's confusion matrix on the UCF101 dataset demonstrates that actions like JumpingJack and TennisSwing were classified correctly, with 100% accuracy each. The model exhibited minor misclassifications for ApplyEyeMakeup, especially in confusing it with PushUps and Basketball, which resulted in a lower class accuracy of 79%.

Additionally, it was observed that the models trained on the UCF dataset that were quite accurate at low input resolution of 28×28. But when the size of the images was increased to 112×112, the overall accuracy of the model decreased quite a lot to about 35%, reflecting how sensitive it is to the resolution of the input. This reduction of accuracy can be connected to the fact that increased resolution provided more parameters for a model to train on. This process and be computationally expensive for the model and thus the low accuracy.

```
2/2 ━━━━━━━━━━━━━━━━━━━━ 1s 154ms/step
Predicted Activities for batch 1: {'Walking'}
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 113ms/step
Predicted Activities for batch 2: {'Walking_Downstairs', 'Walking'}
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 102ms/step
Predicted Activities for batch 3: {'Walking_Downstairs'}
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 97ms/step
Predicted Activities for batch 4: {'Walking_Downstairs', 'Walking'}
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 90ms/step
Predicted Activities for batch 5: {'Walking_Downstairs', 'Walking'}
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 103ms/step
Predicted Activities for batch 6: {'Walking_Downstairs', 'Walking'}
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 97ms/step
Predicted Activities for batch 7: {'Walking_Downstairs', 'Walking'}
```

Figure 5: Predictions made by hybrid CNN-LSTM model in deployment using sensor data

To determine the best implementation approach, we deployed the UCI dataset initially. The accelerometer and gyroscope readings were recorded using a smartphone placed inside the user's pocket to mimic the original dataset setting. For this purpose, we used the SensorStream app and websockets were used to receive these readings to the laptop. [8] However, it was challenging to record this data due to noise as the smartphone was not completely still while in the pocket. This noise prevented clean input signals from being captured.

It was, in fact, not possible to process all 561 features in real-time prediction with processing constraints. Hence, the choice was to pick just 62 features, some by manual selection and others by feature importance analysis on a machine learning algorithm, Random Forest Regressor. Using the entire UCI HAR dataset, consisting of 561 features, during training, the best test accuracy was achieved by the GRU model with 92%. But, when limited to the 62 features, its performance differed. In such scenarios, the CNN-LSTM model performed better than the GRU model, with an accuracy of about 85%, whereas the accuracy of the GRU model decreased to about 75-80%. Interestingly, the CNN-LSTM model exhibited better temporal learning and more stable predictions.

Though the CNN-LSTM model worked, it misclassified certain activities, like "Walking" was confused with "Walking Downstairs." Such confusion is likely due to the noise of the sensor. The time consumed for making predictions was maintained at less than 120ms per prediction, enabling real-time recognition, deployment with sensor data was not investigated further. Video data has more visual information and is less prone to the sensor noise and position issues that are typically available in gyroscope and accelerometer data.

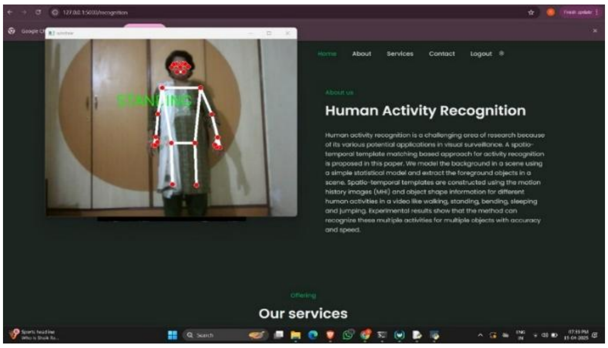

Figure 6: Predictions made by the CNN model

A CNN-trained model was applied to identify human actions using MediaPipe, which records necessary body landmarks from a live webcam. MediaPipe was used with the model due to its low-latency keypoint extraction, suitable for activity recognition using videos. Preprocessed landmarks were passed as input to the model, and it produced predicted labels under 120ms per frame.

The final system performs prediction through a simple interface using Flask and is quite swift in responding. Its precise classification, however, relies on being able to view the entire body, that is, if significant portions such as arms or legs are obscured or not visible, MediaPipe can't obtain all key points, reducing the accuracy. The model was working correctly for movements such as Standing, Crossing Arms, and

***Corresponding Author:** Pallavi Hallappanavar Basavaraja

Praying, but mistakes were observed with Sitting as it was classified as Standing.

CONCLUSION AND FUTURE WORKS

In reflection of the results obtained, the present research offers a universal framework for Human Activity Recognition (HAR) that integrates theory and practice. To provide a comprehensive review, architectures such as RNNs, LSTMs, GRUs, CNNs, and CNN-LSTM hybrid were employed on benchmark datasets. The CNN structure was found to be of significance due to its classifying accuracy and processing speed. It's simple and efficient design makes it versatile for many applications, while its seamless integration with MediaPipe makes it especially well-suited for regular use. The system seamlessly furnishes predictions, and can be utilized in medicine, interactive systems, fitness tracking, and surveillance. The difficulty still remains of full body viewability since it diminishes the accuracy of the predictions.

One of the potential lines of work in the near term is an investigation into hybrid deep learning architectures, such as CNN-BiLSTM, GRU-attention, and CNN-Transformer.

Additionally, occlusion and partial visibility are to be dealt with as well. Methods like multi-camera viewpoints or 3D pose estimation, or simulating partial visibility augmented training data, will perform better in actual situations.

Further enhancement can be done with attention mechanisms and self-supervised learning, focusing attention on significant joints, and boosting activity sets in an unsupervised environment. Multimodal HAR, combining pose with audio, sensor, or depth map, has promise for richer context and improved recognition accuracy, particularly when visual information is scarce.

Scalability and personalization will be key to facilitate broader adoption. Future releases will explore on-device learning and edge-optimized models for wearable devices and mobile. For inclusivity, incremental learning can be introduced to learn to adapt to each person's activity pattern. By exploring these avenues, the HAR framework can be a more flexible and intelligent system, deployable in an enormous variety of applications, ranging from industrial safety and smart homes.

REFERENCES

1. Tee, W. Z., Dave, R., Seliya, N., & Vanamala, M. (2022). *A close look into human activity recognition models using deep learning*. In *Proceedings of the 3rd International Conference on Computing, Networks and Internet of Things (CNIOT)*.
2. Mutegeki, R., & Han, D. S. (2020). A CNN-LSTM approach to human activity recognition. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 362–366). IEEE.
3. Pareek, N., & Thakkar, K. (2021). A survey on video-based human action recognition: Recent updates, datasets, challenges, and applications. *Artificial Intelligence Review, 54*(3), 2259–2286. https://doi.org/10.1007/s10462-020-09904-8
4. Sarveshwaran, Shanmuganathan, S., & Anandan, S. M. (2019). Investigation on human activity recognition using deep learning. *Procedia Computer Science, 165*, 328–336. https://doi.org/10.1016/j.procs.2020.01.047
5. Gulzar, Z., Leema, A. A., & Malaseren, I. (2019). Human activity recognition using machine learning classification techniques. *International Journal of Innovative Technology and Exploring Engineering, 9*(2), 3252–3256.
6. Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013). A public domain dataset for human activity recognition using smartphones. In *Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)* (pp. 437–442). https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones
7. Jansen, M. (2022). *UCF101 – Action recognition data set* [Data set]. Kaggle. https://www.kaggle.com/datasets/matthewjansen/ucf101-action-recognition
8. Kumar, K. R. (2023). *SensorStreamServer* [Computer software]. GitHub. https://github.com/priyankark/SensorStreamServer
9. Basavaraja, P. H., Basavaraju, N. S., et al. (2023). A framework for brain tumor image analysis using convolution with ReLU. *International Journal of Computer Networks and Applications, 2*(4).
10. Basavaraja, P. H., & Ganesarathinam, S. (2022). An ensemble-of-deep learning model with optimally selected features for osteoporosis detection from bone X-ray images. *International Journal of Intelligent Engineering Systems*.
11. Basavaraja, P. H., & Ganesarathinam, S. (2022). Superlative augmented densely connected convolutional neural network for osteoporosis diseases detection. *International Journal of Science and Technology*.
12. Basavaraja, P. H., & Ganesarathinam, S. (2021). Weighted DenseNet-121 for osteoporosis disease detection using X-ray images. *International Journal of Engineering Trends and Technology*.

***Corresponding Author:** Pallavi Hallappanavar Basavaraja