Optimizing Human Activity Recognition with Ensemble Deep Learning on Wearable Sensor Data

Nazish Ashfaq[2], Ibtsam Sadiq[1], Muhammad Junaid[1], Muhammad Hassan Khan[2], and Muhammad Shahid Farid[2]

[1]Department of Software Engineering; University of the Punjab, Lahore, Pakistan

[2]Department of Computer Science; University of the Punjab, Lahore, Pakistan

*Correspondence: nazishashfaq30@gmail.com

Citation | Ashfaq. N, Sadiq. I, Junaid. M, Khan. M. H, Farid. M. S, "Optimizing Human Activity Recognition with Ensemble Deep Learning on Wearable Sensor Data", IJIST, Vol. 6 Issue. 4 pp 1751-1767, Oct 2024

Received | Sep 30, 2024 **Revised |** Oct 21, 2024 **Accepted |** Oct 24, 2024 **Published |** Oct 25, 2024.

In recent years, the research community has shown a growing interest in the continuous temporal data gathered from motion sensors integrated into wearable devices. This type of data is highly valuable for analyzing human activities in a variety of domains, including surveillance, healthcare, and sports. Various deep-learning models have been developed to extract meaningful feature representations from temporal sensory data. Nonetheless, many of these models are constrained by their focus on a single aspect of the data, frequently overlooking the complex relationships between patterns. This paper presents an ensemble model aimed at capturing these intricate patterns by combining CNN and LSTM models within an ensemble framework. The ensemble approach involves combining multiple independent models to harness their strengths, resulting in a more robust and effective solution. The proposed model utilizes the complementary capabilities of CNNs and LSTMs to identify both spatial and temporal features in raw sensory data. A comprehensive evaluation of the model is conducted using two well-known benchmark datasets: UCI-HAR and WISDM. The proposed model attained notable recognition accuracies of 97.92% on the UCI-HAR dataset and 98.52% on the WISDM dataset. When compared to existing state-of-the-art methods, the ensemble model exhibited superior performance and effectiveness.

Keywords: Human Activity Recognition; Ensemble Deep Learning Model; Sensory Data Analysis; Wearable Devices; Time-Series Signal Processing.

Introduction:

Human Activity Recognition (HAR) refers to the process of identifying human actions using either visual data [1] or sensor data collected from wearable devices, such as smartwatches, smart glasses, and smartphones [2]. These HAR systems employ machine learning models and signal processing techniques to classify simple actions like standing, to complex activities like running, cycling, and cooking. HAR is widely used in applications belonging to diverse domains like wellness [3], athletics [4], healthcare [5], security [6], etc. It uses sensory time-series data from sensors like Inertial Measurement Units (IMU), pressure sensors, heart rate monitors, and more. Raw data from the wearable devices cannot be used directly for HAR; as it lacks context, is noisy,and contains missing values. It has to be pre-processed for its effective use.

Many approaches have been developed to extract useful features from continuous temporal data as shown in Figure 1. These are handcrafted codebooks and deep-learning approaches. Handcrafted features-based approaches depend on the domain knowledge of researchers to manually extract features from this data, including simple statistical measures like mean, median, variance, maximum, minimum, mode, and standard deviation, as well as more complex and rich components, such as frequency domain-based features,which are related to the Fourier transform of the signals [7]. On the other hand, codebook-based algorithms, such as the Bag-of-Features (BOF) approach, use unsupervised clustering algorithms like K-means to createa histogram-based representation of sensory data which is then used for inference. However, this clustering process can be computationally demanding and is not scalable, must be repeated when new classes are introduced. Deep learning models provide a more effective and efficient method for recognizing human activities as compared to handcrafted and codebook-based approaches because they can learn hierarchical data representations and capture temporal patterns directly from raw sensory data [8]. Given large amounts of data and careful training, these models perform exceptionally better than other mentioned techniques. Moreover, these models are more adaptive and generalize well in production.

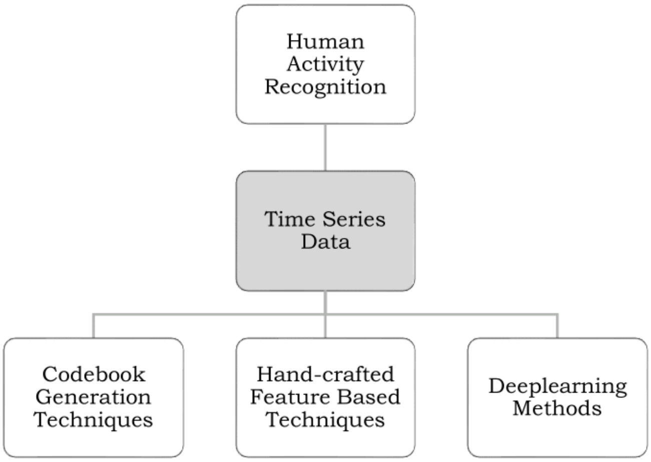
Human Activity Recognition
Time Series Data
Codebook Generation Techniques
Hand-crafted Feature Based Techniques
Deeplearning Methods

Figure 1. The distribution of feature extraction methods for HAR using time-series sensory data.

This research emphasizes HAR using sensory devices, particularly through the implementation of an ensemble model. The proposed CNN-LSTM model serves as an ensemble of models by combining the spatial feature extraction capabilities of CNN with the temporal dependency capturing abilities of LSTM. The proposed architecture comprises convolutional layers followed by LSTM layers, allowing for the effective processing of sequential data and identification of activity patterns. The proposed model is tested on two benchmark datasets: UCI-HAR [9], and WISDM [10]. Experimental results testify to the ensemble model's effectiveness by achieving an accuracy of 97.92% for UCI-HAR and 98.52% for WISDM datasets, showcasing its potential for robust human activity recognition across diverse datasets and applications. The contributions of this paper are summarized as follows:

- A summarized review of existing feature extraction techniques for HAR.
- A novel technique has been proposed to automatically learn features and recognize activities fromwearable smart device data.
- Performance assessment of the proposed CNN-LSTM model over two well-known public datasets.

Literature Review:

The utilization of wearable sensor data obtained from different devices, such as smartphones, smartwatches, and smart glasses, is the most recent trend in the field of HAR research. Such devices contain IMU(accelerometer, gyroscope, and magnetometer) sensors, which record the user's movement as well as pressure and heart rate sensors for environmental and physiological data, respectively. Many approaches have been developed to extract useful features from time-series sensory data. These are (i)handcrafted, (ii) codebook, and (iii) deep learning-based approaches.

Handcrafted Feature-Based Techniques:

Handcrafted features are mathematical and statistical measures calculated from raw data, by using domain knowledge and experience on the subject. Researchers extract mathematical features from cleaned sensory data. The extracted data points are then combined to create a feature vector, which is subsequently used as input for classification algorithms like support vector machines, HMM, logistic regression, etc. These features are computationally inexpensive and take little time to set up. An overview of the entire process, from handcrafted feature extraction to activity classification, is given in Figure 2.

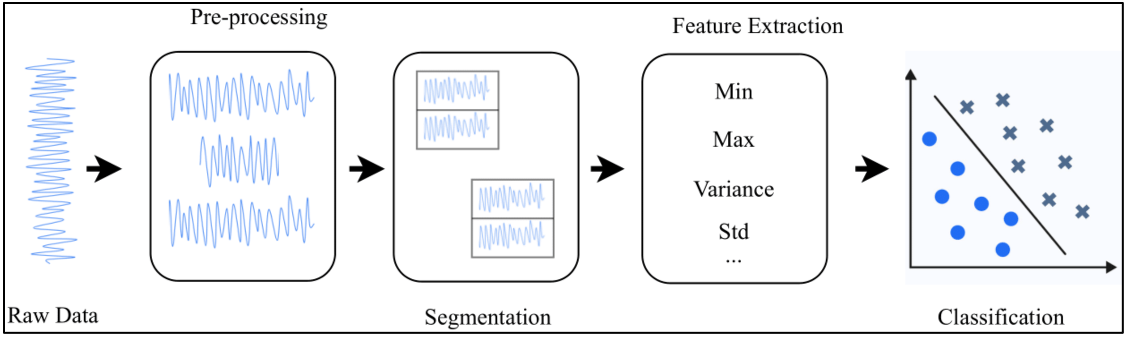

Figure 2. Illustration of HAR process following handcrafted feature extraction, which involves pre-processing, segmentation, feature extraction, and computation.

The researchers in [11][12] used statistical measures like mean, variance, maximum, minimum, average, kurtosis, and standard deviation. While [13] investigated frequency domain features for identifying physical activities. The author [7] Introduced a hierarchical approach for HAR using wearable sensors. They addressed composite daily actions such as cooking, which involve multiple atomic actions, by initially extracting hand-craftedfeatures and employing them for subspace pooling. Moreover, [14] used appearance-based approaches, fuzzy logic, space-time analysis, and local binary patterns to extract these features from raw sensory data. Researchers computed a subset of features from raw sensory data in [15][16] and used them as deep neural network input. They concluded that combining hand-crafted features with deep neural networks improves classification accuracy. Although handcrafted features can be computed quickly, their effectiveness is largely dependent on the researcher's domain knowledge and ability to extract relevant information from unprocessed data [17].

Codebook-based Techniques:

Unlike handcrafted features, these techniques use the BOF approach [18] which has two primary parts: (i) codebook generation and (ii) codeword assignment by feature encoding as shown in Figure 3. Initially, raw time series sensory data is grouped using clustering by their underlying patterns, forming groups, and each group's centroid is assigned as a codeword. Then, the resulting activity sequence data becomes a final representation by assigning these codewords, which are part of the histogram-based representation, which is ultimately used by a classification machine learning algorithm [19].

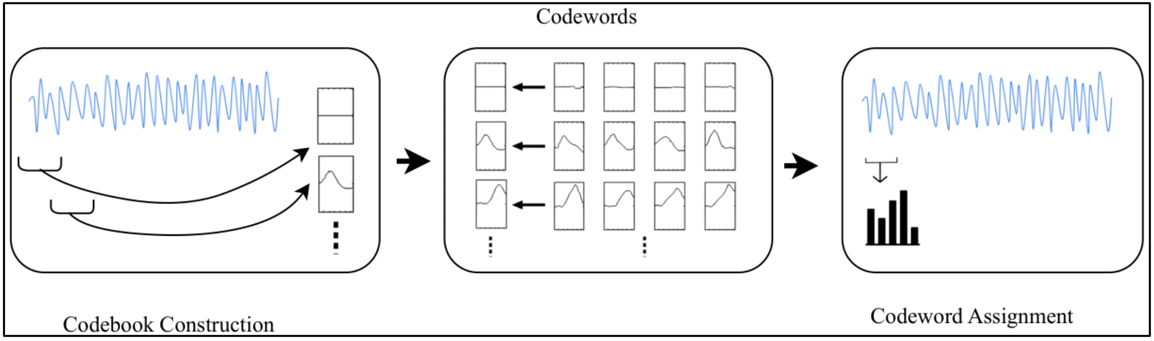

Figure 3. Pictorial representation of the HAR using a codebook-based feature extraction approach, where codewords are derived from raw sensory data. Codebook methods simplify complex data by summarizing raw signals into structured sets of "words," facilitating efficient processing and pattern recognition.

The codebook method was employed for many different recognition tasks that dealt with bodily, mental,and eye-gaze activities [20]. A codebook-based feature learning methodology was presented by Koping et al. [21] to identify human activities from sensory data. They estimated codewords in activity sequences to create a feature vector representation based on a histogram, and they generated a codebook using the k-mean clustering algorithm. A different study [18] employed this technique for person identification using gait. They encoded the static appearance and motion information of the walker. Codebook-based approacheswork incredibly well because they use clustering algorithms to capture hidden patterns in different human activities. Nonetheless, codebook computation with an ideal cluster size is a time-expensive procedure [22].

Deep Learning-Based Techniques:

The use of deep neural network architectures has considerably advanced the use of sensory data for human activity recognition, allowing automatic learning of hidden patterns in the input data. As shown in Figure 4, these networks usually consist of several layers with hidden neurons that input data passes through to learn useful features through weight adjustments using the back-propagation algorithm, which calculates the derivatives to each weight used in the network [23]. These models can automatically extract discriminative features fromunprocessed signal data. For HAR, several deep network architectures have been used, such as recurrent neural networks (RNNs) [24], gated recurrent units (GRUs) [25][26][27][28], convolutional neural networks (CNNs) [11][13][26][29] LSTM [12], and hybrid models [30]. RNNs are tailored for sequential data, withconnections between nodes forming a directed graph over a temporal sequence, enabling the network toretain memory through its hidden states [31].

While deep learning-based methods are excellent at concluding unprocessed sensory data, stand-alone models may find it difficult to fully capture the intricacy of underlying patterns in complex tasks like analyzing continuous time series data of human activities. Ensemble deep networks combine the best features of several models into a single, cohesive framework [32]. The author [15][33] used sensory data to combine CNN andLSTM networks to identify human activities. They found remarkable performance gains as compared to single networks. By utilizing the strengths of various networks within a single framework, these models are excellent at capturing complex spatial features and intricate temporal relationships [34]. This allows them to prevail over the shortcomings of individual models and provide a more all-encompassing solution to the complexities of HAR systems.

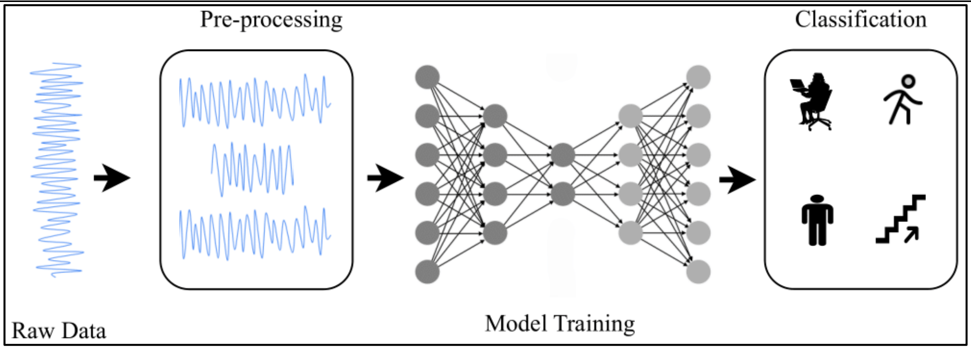

Figure 4. Illustration of HAR using deep learning network, applied to raw sensory data.

Material and Methods:

Ensemble models sequentially integrate multiple deep learning architectures, with the output of one model feeding into the next, forming a layered structure for data processing [35]. In this study, we present a thorough evaluation of a CNN-LSTM ensemble model for human activity recognition. The model's performance is assessed using two well-known public datasets: WISDM and UCI-HAR. After the evaluation of several model configurations, we determined that the proposed ensemble CNN-LSTM model achieved the highest accuracy and efficiency. A detailed breakdown of the model's implementation is also provided to demonstrate our approach.

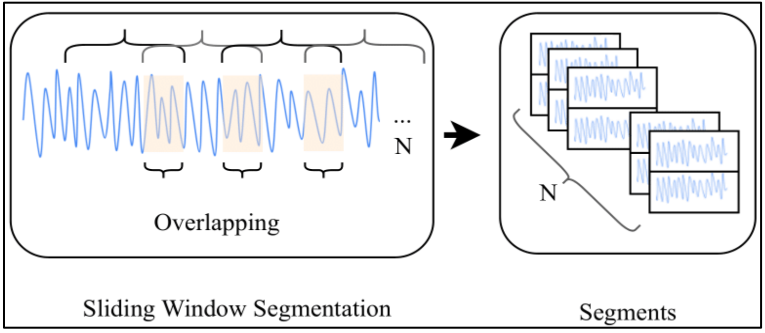

Figure 5. Depiction of the process of segmenting raw time-series data into segments using the sliding window approach.

Data Preprocessing:

We used IMU sensor data, widely available in common wearable devices such as smartphones. These sensorsrecord raw time series data that allow collecting information on the body motion of the subject. These sensors measure different forces acting on the device for instance, accelerometers measure acceleration force, magnetometer measures the magnetic force, and gyroscope sensors measure angular velocities. By integrating data from these sources, we can capture human motion in the three dimensions: x, y, and z. Based on guidelines found in the literature, the raw sensor data was split into samples of 128 frames, with a 50% overlap between frames [21], this segmentation of data into individual segments is illustrated in Figure 5.

Model Architecture Design:

The proposed ensemble model is composed of multiple layers, including CNN, LSTM, dropout, and batch normalization layers. These layers are arranged hierarchically to extract distinguishing features from the pre-processed data. In the final stage, fully connected layers are employed, utilizing the SoftMax activation function for activity classification.

Convolutional Neural Network (CNN):

CNNs with their layered structure of convolution and pooling layers are extremely efficient in spatial feature extraction from raw data. A CNN consists of several levels or layers, with each layer containing several filters that, when applied to the input data, produce feature maps; an abstract view of a CNN is depicted in Figure 6. Convolutional layers are succeeded by batch normalization and dropout layers. The batch normalization layer caters to the covariant shift in the data as it passes from one layer to another. Dropout layers decrease the number of parameters, preventing overfitting in the model [36]. All the filters of a CNN layer are applied on each channel of the input which generates a set of feature maps. These filters perform local pattern recognition. The CNN uses Stochastic gradient descent (SGD) for parameter optimization during training.

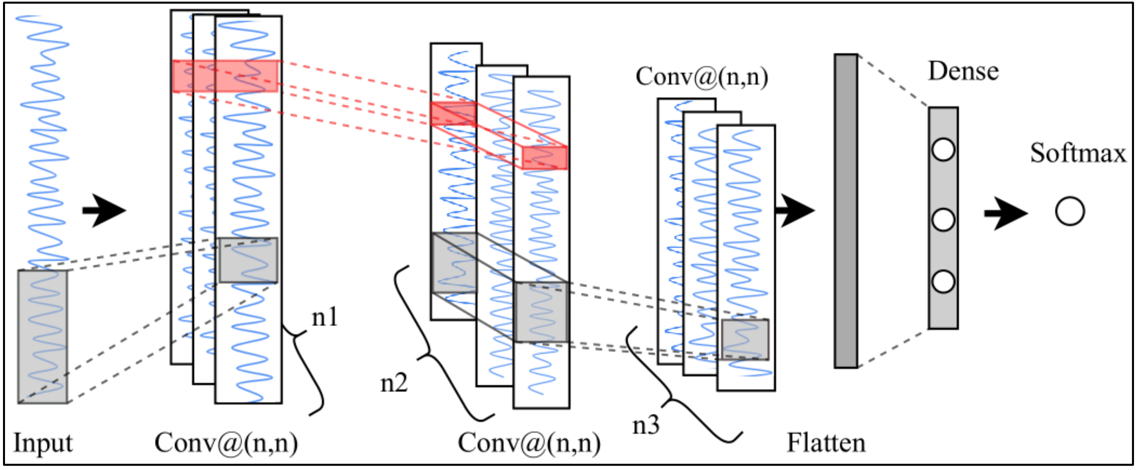

Figure 6. An illustration of a CNN. The network consists of multiple convolutional layers succeeded by dense layers and a SoftMax layer for classification. n represents the dimensions of the layer.

Long Short-Term Memory (LSTM):

LSTMs are a variant of recurrent neural networks (RNNs) that excel at capturing long-term dependencies in temporal data. Unlike traditional RNNs, LSTMs overcome challenges such as vanishing or exploding gradients by utilizing cell states and gating mechanisms specifically the forget, input, and output gates to preserve crucial information over time. The forget gate, using a sigmoid function, determines which portions of the cell state should be discarded. The input gate selects which parts of the input should be retained, while the output gate decides what portion of the cell state should be propagated forward [37]. These mechanisms allow LSTMs to learn from time series data by taking prior data points into account. For instance, considering x_t as the input at time step t, h_{t-1} as the hidden state from the previous time step t $_1$, and W as the weight matrix, the gates' operations can be described mathematically as follows:

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \quad (2)$$

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \quad (3)$$

$$\tilde{C}_t = \tanh(W_c \times [h_{t-1}, x_t] + b_c) \quad (4)$$

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \tilde{C}_t \quad (5)$$

$$h_t = o_t \otimes \tanh(C_t) \quad (6)$$

Here, f_t represents the forget gate activation vector at time t, σ is the sigmoid activation function, W denotesthe weight matrices, i refers to the input gate, o to the output gate, and f to the forget gate. The hidden state is represented by h, while x_t indicates the input vector at the current time step. The concatenation of h_{t-1} (the hidden state from the previous time step) and x_t is denoted by $[h_{t-1}, x_t]$. The bias vectors are represented by b, $\tilde{C}_t$ is the candidate cell state vector, and tanh is the hyperbolic tangent function. The cell state at time t is given by C_t, the element-wise multiplication operation is represented by $\otimes$, and C_{t-1} refers to the previous cell state vector. Figure 7 represents the LSTM architecture.

Proposed Model:

The CNN-LSTM ensemble neural network leverages the strengths of both models. While LSTMs excelat capturing sequential data and time-based dependencies, CNNs are proficient in learning spatial features. The proposed ensemble network integrates two convolutional layers, each followed by a ReLU activation function. After the first CNN layer, a batch normalization layer is applied, and after the second CNN layer, a dropout layer is added. Each convolutional layer consists of filters of various sizes to capture features across different levels of spatio-temporal resolution. The dropout and batch normalization layers are usedin sequence after each convolutional layer to normalize and enhance the training process by adjusting and scaling the data in each training batch. Dropout layers are introduced to prevent overfitting by regularizing the data between network layers [38]. Batch normalization helps the model converge faster by scaling featuresto the same range, addressing covariant shifts, and ultimately improving overall performance. The output of the convolutional layers is flattened and passed to the LSTM network, which comprises two LSTM layers with 128 units each, using a tanh activation function. Similar to the CNN layers, the LSTM layers also utilizedropout and batch normalization layers. The final output from the LSTM is forwarded to two dense layers with ReLU activation functions and a classification layer employing the SoftMax function for classification. The architecture of the proposed model is illustrated in Figure 8.

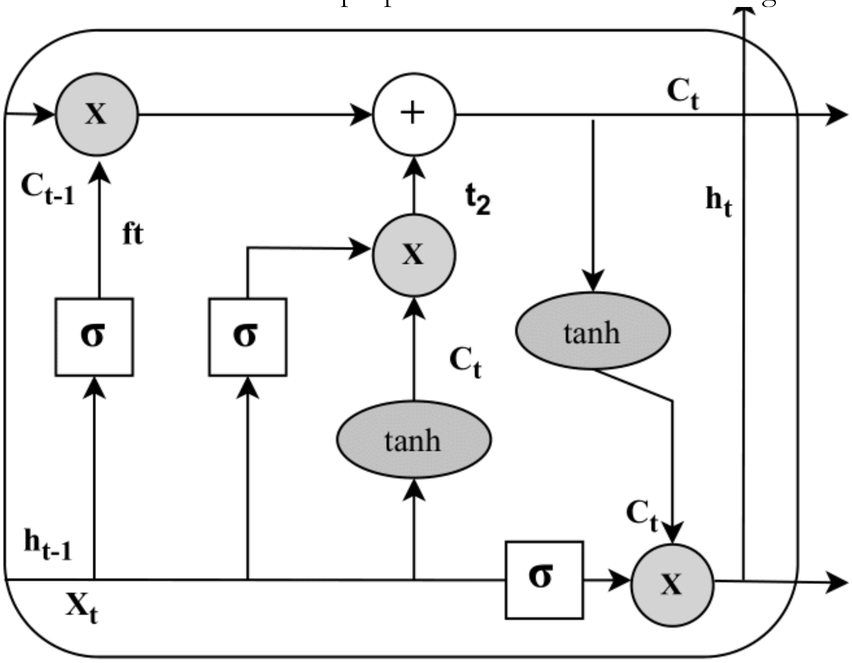

Figure 7. Illustration of the LSTM unit with input, output, and forget gates which control the flow of data.

The Adam optimizer is used for the proposed model, which adjusts the learning rate for each neuronusing the previous and squared gradients. We took k-fold cross-validation into account as a verifying factor in terms of the learned weights. A categorical cross-entropy loss function was adopted to measure the network's accuracy, as a guide for the training stage. Different hyper-parameters used during training are listed in Table 1:

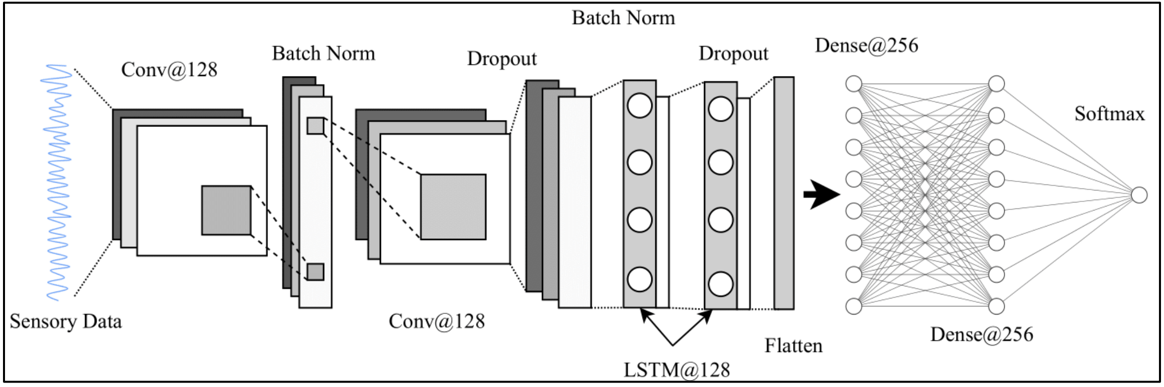

Figure 8. Proposed CNN-LSTM Network Architecture for Analyzing Time Series Sensory Data.

Table 1. Overview of the components and hyperparameters of the proposed CNN-LSTM model, with subscripts indicating the respective layer numbers.

Phase	Parameters	Values
Pre-processing	Window size	128
	Step size	64
	Features	3
Training	Optimizer	Adam
	Learning rate(alpha)	0.001
	Batch size	128
	Loss function	Categorical cross-entropy
Architecture	CNN$_1$	128
	BatchNorm$_1$CNN$_2$	Axis = -1
		128
	Dropout$_1$	0.1
	LSTM$_1$	128
	BatchNorm$_2$LSTM$_2$	Axis = -1
		128
	Dropout$_2$	0.1
	Dense$_1$	256
	Dense$_2$	256

WISDM Dataset:

The WISDM dataset [39] has in total over 1 million samples of smartphone activity data from 36 subjects. The activities include sitting, standing, walking, jogging, and walking upstairs. During these activities, participants held a smartphone in a front leg pocket and took readings from an accelerometer at 20 Hz with 50% overlap of 4-second intervals with a total number of 128 measurements per window. Table 2 below shows the summary of the dataset.

Table 2. Activity distribution and parameters used for collection of the WISDM dataset.

Dataset	WISDM
Total Participants	36
Total Activities	6
Window size	128
Sampling rate	20 Hz
Data dimension	3
Sensor	1 (Accelerometer)
Device & its placement	Smartphone, Front leg pocket
Activities	**Instances**
Downstairs	100,427
Jogging	342,177
Sitting	59,939
Standing	48,395
Upstairs	122,869
Walking	424,400

UCI-HAR Dataset:

The UCI-HAR dataset [9] captures the physical activities of 30 participants aged between 19 and 48. Six distinct activities are recorded: sitting, standing, lying, walking, walking downstairs, and walking upstairs. Participants performed these activities while carrying a Samsung Galaxy S II smartphone in their waist pocket. The smartphone's accelerometer and gyroscope sensors recorded inertial motion at a frequency of 50 Hz. A summary of the dataset is presented in Table 3. When performing the activity recognition, the raw sensory data was divided into frames as follows: a window size of 128 timestamps with 50% overlap, or 64 step intervals, was used. This segmentation provided 128 measurements per window that enabled the generalization capacity for the model. Figure 9 illustrates the segmentation of data into individual segments.

Table 3 Activity distribution and parameters used for collection of the UCI-HAR dataset.

Dataset	**UCI-HAR**
Total Participants	30
Total Activities	6
Window size	128
Sampling rate	50 Hz
Data dimensions	9
Sensor	2 (Gyroscope & Accelerometer)
Device & its placement	Smartphone, Waist
Activities	**Instances**
Walking-downstairs	1406
Walking-upstairs	1544
Sitting	1777
Walking	1722
Standing	1906
Laying	1944

Implementation Details:

The network comprises 12 layers and it is implemented using Kera's framework. The model is trained in a supervised learning setting using the RMSprop optimizer, a variant of gradient descent [37]. The LSTM layers consisted of 128 hidden units, and the dataset was processed over 70 epochs with a batch size of 128. The categorical cross-entropy loss function is employed during training. The validation loss is used after training and the Early Stopping callback process [40] is also used to find appropriate weights. The last layer is formed by the SoftMax function for the task of providing output probabilities. The ratio of train, test, and validation split is 70:20:10 respectively. The process of hyperparameter optimization is done with a multi-resolution search method [41]. This approach is employed to optimize hyperparameters, including the number of LSTM layers, their integration with batch normalization, the number of dense layers, and the choice of activation function. For this multi-class classification problem, the categorical cross-entropy loss function is utilized [42].

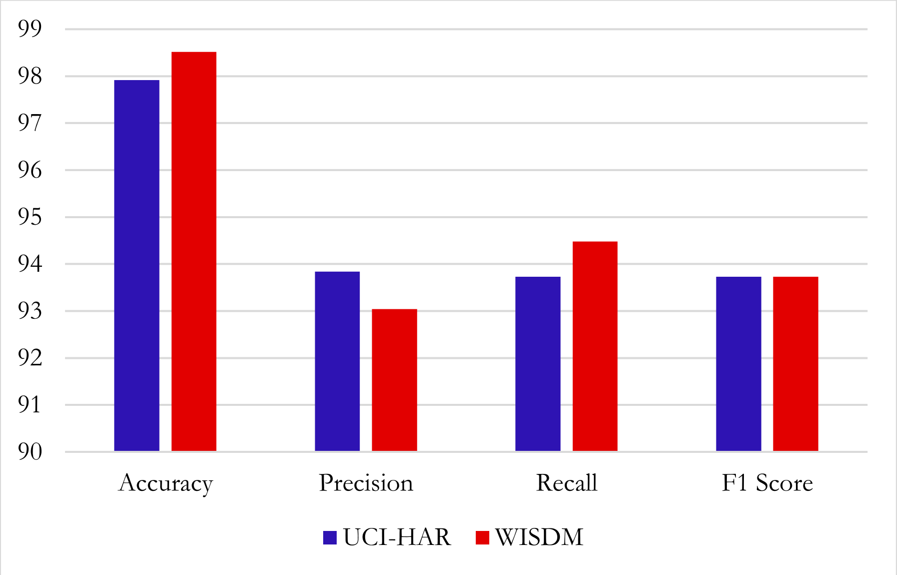

Figure 9. Depicts the performance of the proposed ensemble CNN-LSTM model on WISDM and UCI-HAR datasets. The performance is assessed using various matrices including accuracy, precision, recall, and F_1-score.

Result and Discussion:

In this study, we evaluated the performance of a CNN-LSTM ensemble model for HAR using two popular datasets: WISDM and UCI-HAR datasets. The CNN-LSTM model combines the strengths of CNNs and LSTM networks, making it well-suited for capturing both spatial and temporal features from time-series data. This hybrid approach is crucial for HAR tasks, where sensor data contains both temporal dependencies and patterns in spatial dimensions. Figure 9 presents the recognition scores achieved by the proposed CNN-LSTM model on UCI-HAR and WISDM datasets. Whereas, the training and validation accuracy of the network is visualized in Figure 10. In addition, the training and validation loss of the network is visualized in Figure 11. These visualizations offer a clear representation of how the model's performance progresses over time, providing insight into its learning trajectory and convergence behavior.

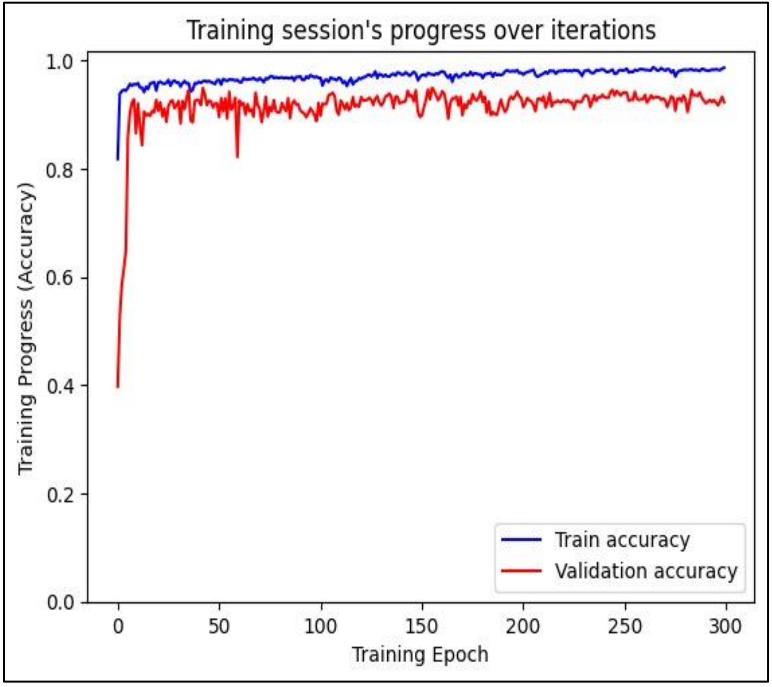

Figure 10. Depicts the training and validation accuracy for the proposed ensemble CNN-LSTM model on the UCI-HAR dataset.

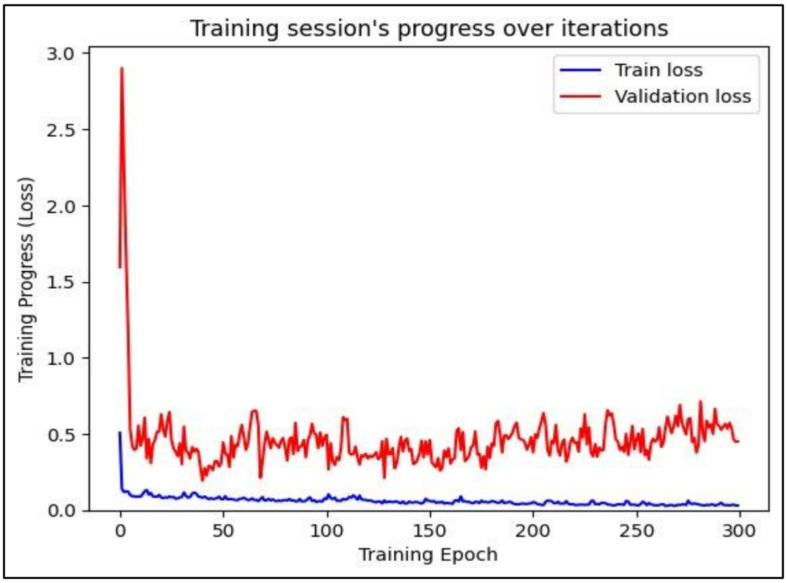

Figure 11. Depicts the training and validation loss for the proposed ensemble CNN-LSTM model on the UCI-HAR dataset.

Table 4 represents the classification performance on the UCI-HAR dataset, evaluated using various metrics. The macro-averaged accuracy across all activities is 97.92%, indicating that the model performs exceptionally well in classifying human activities overall. However, there are notable variations in the performance of different activities. The walking activity achieved an accuracy of 97.79%, a precision of 97.57%, a recall of 89.11%, and an F_1 score of 93.15%. While the high precision indicates that most predictions labeled as" Walking" are correct, the slightly lower recall suggests that the model occasionally confuses walking with other activities, particularly walking upstairs. The F_1 score of 93.15% reflects a good balance between precision and recall for this activity. For walking upstairs, the model performs very well with an accuracy of 97.39%, precision of 91.74%, recall of 91.93%, and an F_1 score of 91.83%. This shows that the model is highly effective at distinguishing walking upstairs, though there is a slight confusion with other activities like walking and walking downstairs. Walking downstairs achieved an outstanding accuracy of 99.29%, with a high precision of 96.50%, recall of 98.57%, and an F_1 score of 97.53%, indicating that the model performs exceptionally well in identifying this activity with minimal misclassifications.

Table 4. Performance metrics (Accuracy, Precision, Recall, and F_1 Score) for activity classification using theUCI-HAR dataset

Activities	Accuracy (%)	Precision (%)	Recall (%)	F_1 Score (%)
Walking	97.79	97.57	89.11	93.15
Walking Upstairs	97.39	91.74	91.93	91.83
Walking Downstairs	99.29	96.50	98.57	97.53
Sitting	95.89	88.70	86.35	87.51
Standing	97.42	90.00	96.43	93.10
Laying	99.73	98.53	100.00	99.26
Average	97.92	93.84	93.73	93.73

The model's performance for more static activities such as sitting and standing is also strong. For sitting, the model achieved an accuracy of 95.89%, a precision of 88.70%, a recall of 86.35%, and an F_1 score of 87.51%. However, sitting exhibits slightly lower performance compared to other activities, which may be attributed to its occasional misclassification as standing. Standing, on the other hand, performs better, with an accuracy of 97.42%, precision of 90.00%, recall of 96.43%, and an F_1 score of 93.10%, showing that the model is highly reliable in detecting this activity. The highest performance is observed for laying, where the model achieved a near-perfect accuracy of 99.73%, precision of 98.53%, recall of 100%, and an F_1 score of 99.26%. This indicates that the model can almost perfectly classify instances of laying without confusion with other activities.

The classification performance on the WISDM dataset is evaluated using several metrics, including accuracy, precision, recall, and F_1 Score as shown in Table 5. The macro-averaged accuracy across all activities is 98.52%, indicating that the model performs very well in classifying human activities overall. However, the performance varies across different activities. The walking activity achieved a precision of 85.37%, a recall of 90.00%, and an F_1 score of 87.62%. This suggests that while the model identifies most walking instances correctly, there are a moderate number of false positives, possibly due to confusion with similar activities like jogging. On the other hand, jogging shows excellent performance, with a precision of 99.11%, a recall of 96.62%, and an F_1 score of 97.85%. The high precision and recall for jogging indicate that the model is highly accurate in predicting this activity with minimal misclassification. For activities such as walking upstairs and walking downstairs, the model demonstrates strong performance, with accuracies of 99.85% and 99.59%, respectively. The F_1 scores for these activities are 98.64% and 94.96%, respectively, indicating that the model effectively distinguishes these activities from others, with slight confusion in certain cases. These results suggest that the model is highly effective in handling dynamic activities. The model's performance for more static activities like sitting and standing also reflects the model's strength, with standing achieving a precision of 98.59% and an F_1 score of 98.11%. However, sitting shows a slightly lower precision of 83.33%, indicating that it is occasionally misclassified as another static activity such as standing. The F_1 score for sitting is 85.20%, reflecting a balance between precision and recall, but with room for improvement in distinguishing this activity from others.

Table 5. Performance metrics (Accuracy, Precision, Recall, and F_1 Score) for activity classification using theWISDM dataset.

Activities	Accuracy (%)	Precision (%)	Recall (%)	F_1 Score (%)
Walking	97.38	85.37	90.00	87.62
Jogging	98.70	99.11	96.62	97.85
Upstairs	99.85	98.91	98.37	98.64
Downstairs	99.59	92.96	97.06	94.96
Sitting	97.08	83.33	87.16	85.20
Standing	98.50	98.59	97.65	98.11
Average	98.52	93.04	94.48	93.73

Discussions:

The results from both datasets underscore the effectiveness of the CNN-LSTM ensemble model for HAR. The CNN component efficiently extracts spatial features from sensor data, identifying important patterns from the provided readings. These patterns are particularly useful for distinguishing activities with unique spatial characteristics, such as walking downstairs or jogging. The LSTM component, on the other hand, captures temporal dependencies, allowing the model to understand sequences of sensor readings over time. This is critical for accurately classifying activities that involve continuous movement patterns, such as walking, jogging, and walking upstairs.

The ensemble CNN-LSTM model proves efficient for both dynamic and static activities, as seen in the consistent performance across both the WISDM and UCI-HAR datasets. In particular, the model excels in identifying dynamic activities, with consistently high F_1 scores for jogging, walking downstairs, and walking upstairs in both datasets. Its ability to handle temporal data effectively is reflected in the high recall and F_1 scores, which indicate that the model can accurately capture the temporal transitions between different activity states. However, the results also reveal some challenges in distinguishing between similar static activities like sitting and standing, where there is occasional misclassification. This can be attributed to the similar sensor readings for these activities, which rely more on subtle differences in posture or minimal movement.

Table 6. Comparison of the proposed model's recognition results (%) against the state-of-the-art method using the UCI-HAR dataset. The top outcomes are highlighted in bold.

Methods	Year	Accuracy
Wang et al. [43]	2023	96.0
Gupta et al. [44]	2021	87.65
Soni et al. [45]	2023	97.15
Khan et al. [46]	2021	95.4
Tong et al. [30]	2022	95.4
Proposed CNN-LSTM	2024	**97.92**

In future work, several strategies could be explored to improve the classification performance of sitting in HAR tasks using the ensemble CNN-LSTM model. Incorporating additional sensor modalities could help better distinguish sitting from other static activities like standing or laying. Furthermore, fusing sensor data from multiple body positions could offer a more comprehensive understanding of posture. Combining these approaches has the potential to reduce the misclassification of sitting and enhance overall classification accuracy.

Table 7. Comparison of the proposed model's recognition results (%) against the state-of-the-art methods using the wisdom dataset. The top outcomes are highlighted in bold.

Methods	Year	Accuracy
Seelwal et al. [47]	2023	87.85
Afsar et al. [48]	2023	88.46
Al-juaifari et al. [49]	2023	89.43
Semwal et al. [50]	2022	90.0
Duan et al. [51]	2022	90.77
Proposed CNN-LSTM	2024	**98.52**

Comparison with State-of-the-Art Techniques:

Tables 6 and 7 present a comprehensive analysis of the results compared to existing techniques. The achieved accuracies greatly surpass numerous existing models in the literature, suggesting that the ensemble approach offers a competitive advantage. For example, conventional machine learning models and independent deep learning architectures often face challenges in achieving such impressive performance levels on these datasets. The suggested model attained macro-averaged accuracy rates of 98.52% on the UCI-HAR dataset and 97.92% on the WISDM dataset. These findings highlight the efficacy of ensemble models in improving the precision and dependability of HAR systems. Subsequent investigations may enhance the CNN-LSTM model by incorporating diverse sensors, employing advanced data augmentation methods, and formulating more intricate ensemble strategies. Furthermore, evaluating the model in practical situations among various demographic cohorts would yield significant insights for enhancement.

Conclusion:

This paper provides a comprehensive assessment of the ensemble model proposed for the HAR using time series sensory data collected from everyday wearable devices like smartphones and smartwatches. The ensemble models work by combining or integrating multiple models to capitalize on their unique strengths, resulting in a more effective solution. The proposed CNN-LSTM-based ensemble network is capable of capturing both spatial and temporal features, enabling it to learn relationships within raw sensory data and accurately identify human activities. The model's performance is evaluated on two widely used datasets: UCI-HAR and WISDM. The recognition outcomes, along with comparisons to recent state-of-the-art approaches, highlight the efficiency of the proposed hybrid CNN-LSTM model.

Acknowledgment: This research was partially supported by the Higher Education Commission Pakistan within the project "HumCare: Human Activity Analysis in Healthcare" (Grant Number: 15041).

Author's Contribution: Conceptualization, I.S., M.J., N.A., M.H.K. and M.S.F.; methodology, I.S., and M.H.K.; software, M.J., and M.S.F.; validation, I.S.; M.J.; investigation, N.A., M.H.K. and M.S.F.; writing—original draft preparation, N.A.; writing—review and editing, N.A., M.H.K. and M.S.F.; supervision, M.H.K. and M.S.F.; project administration, M.H.K. and M.S.F. All authors have read and agreed to the published version of the manuscript.

Conflict of Interest: The authors declare no conflicts of interest.

Project Details: This research was conducted under the project titled "HumCare: Human Activity Analysis in Health Care" under grant number: 15041

List of Abbreviations

HAR:	Human Activity Recognition
CNN:	Convolutional Neural Network
LSTM:	Long Short-Term Memory

- **IMU:** Inertial Measurement Unit
- **RNN:** Recurrent Neural Networks
- **ReLU:** Rectified Linear Activation Function
- **WISDM:** Wireless Sensor Data Mining
- **BOF:** Bag-of-Features
- **SGD:** Stochastic Gradient Descent
- **RMSprop:** Root Mean Squared Propagation

References:

[1] "Human Activity Analysis in Visual Surveillance and Healthcare (Studien Zur Mustererkennung): 9783832548070: Computer Science Books @ Amazon.com." Accessed: Oct. 22, 2024. [Online]. Available: https://www.amazon.com/Activity-Analysis-Surveillance-Healthcare-Mustererkennung/dp/3832548076

[2] T. Haider, M. H. Khan, and M. S. Farid, "An Optimal Feature Selection Method for Human Activity Recognition Using Multimodal Sensory Data," Inf. 2024, Vol. 15, Page 593, vol. 15, no. 10, p. 593, Sep. 2024, doi: 10.3390/INFO15100593.

[3] E. Ferrara, "Large Language Models for Wearable Sensor-Based Human Activity Recognition, Health Monitoring, and Behavioral Modeling: A Survey of Early Trends, Datasets, and Challenges," Sensors 2024, Vol. 24, Page 5045, vol. 24, no. 15, p. 5045, Aug. 2024, doi: 10.3390/S24155045.

[4] Y. Zhao et al., "Image expression of time series data of wearable IMU sensor and fusion classification of gymnastics action," Expert Syst. Appl., vol. 238, p. 121978, Mar. 2024, doi: 10.1016/J.ESWA.2023.121978.

[5] I. Priyadarshini, R. Sharma, D. Bhatt, and M. Al-Numay, "Human activity recognition in cyber-physical systems using optimized machine learning techniques," Cluster Comput., vol. 26, no. 4, pp. 2199–2215, Aug. 2023, doi: 10.1007/S10586-022-03662-8/METRICS.

[6] F. Amjad, M. H. Khan, M. A. Nisar, M. S. Farid, and M. Grzegorzek, "A Comparative Study of Feature Selection Approaches for Human Activity Recognition Using Multimodal Sensory Data," Sensors 2021, Vol. 21, Page 2368, vol. 21, no. 7, p. 2368, Mar. 2021, doi: 10.3390/S21072368.

[7] S. Waghchaware and R. Joshi, "Machine learning and deep learning models for human activity recognition in security and surveillance: a review," Knowl. Inf. Syst., vol. 66, no. 8, pp. 4405–4436, Aug. 2024, doi: 10.1007/S10115-024-02122-6/METRICS.

[8] F. Li, K. Shirahama, M. A. Nisar, L. Köping, and M. Grzegorzek, "Comparison of Feature Learning Methods for Human Activity Recognition Using Wearable Sensors," Sensors 2018, Vol. 18, Page 679, vol. 18, no. 2, p. 679, Feb. 2018, doi: 10.3390/S18020679.

[9] G. M. Weiss, K. Yoneda, and T. Hayajneh, "Smartphone and Smartwatch-Based Biometrics Using Activities of Daily Living," IEEE Access, vol. 7, pp. 133190–133202, 2019, doi: 10.1109/ACCESS.2019.2940729.

[10] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human Activity Recognition on Smartphones Using a Multiclass Hardware-Friendly Support Vector Machine," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7657 LNCS, pp. 216–223, 2012, doi: 10.1007/978-3-642-35395-6_30.

[11] W. N. Ismail, H. A. Alsalamah, M. M. Hassan, and E. Mohamed, "AUTO-HAR: An adaptive human activity recognition framework using an automated CNN architecture design," Heliyon, vol. 9, no. 2, Feb. 2023, doi: 10.1016/J.HELIYON.2023.E13636/ASSET/8182D440-0B5D-4D21-9255-

91C8EA2B55F4/MAIN.ASSETS/GR016.JPG.

[12] M. Bock, A. Hölzemann, M. Moeller, and K. Van Laerhoven, "Improving Deep Learning for HAR with Shallow LSTMs," Proc. - Int. Symp. Wearable Comput. ISWC, pp. 7–12, Sep. 2020, doi: 10.1145/3460421.3480419.

[13] N. Rashid, B. U. Demirel, and M. Abdullah Al Faruque, "AHAR: Adaptive CNN for Energy-Efficient Human Activity Recognition in Low-Power Edge Devices," IEEE Internet Things J., vol. 9, no. 15, pp. 13041–13051, Aug. 2022, doi: 10.1109/JIOT.2022.3140465.

[14] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, "A new hybrid deep learning model for human action recognition," J. King Saud Univ. - Comput. Inf. Sci., vol. 32, no. 4, pp. 447–453, May 2020, doi: 10.1016/J.JKSUCI.2019.09.004.

[15] R. Mutegeki and D. S. Han, "A CNN-LSTM Approach to Human Activity Recognition," 2020 Int. Conf. Artif. Intell. Inf. Commun. ICAIIC 2020, pp. 362–366, Feb. 2020, doi: 10.1109/ICAIIC48513.2020.9065078.

[16] Z. Chen, L. Zhang, Z. Cao, and J. Guo, "Distilling the Knowledge from Handcrafted Features for Human Activity Recognition," IEEE Trans. Ind. Informatics, vol. 14, no. 10, pp. 4334–4342, Oct. 2018, doi: 10.1109/TII.2018.2789925.

[17] R. Fatima, M. H. Khan, M. A. Nisar, R. Doniec, M. S. Farid, and M. Grzegorzek, "A Systematic Evaluation of Feature Encoding Techniques for Gait Analysis Using Multimodal Sensory Data," Sensors 2024, Vol. 24, Page 75, vol. 24, no. 1, p. 75, Dec. 2023, doi: 10.3390/S24010075.

[18] M. H. Khan, M. S. Farid, and M. Grzegorzek, "A comprehensive study on codebook-based feature fusion for gait recognition," Inf. Fusion, vol. 92, pp. 216–230, Apr. 2023, doi: 10.1016/J.INFFUS.2022.12.001.

[19] M. H. Khan, M. S. Farid, and M. Grzegorzek, "Using a generic model for codebook-based gait recognition algorithms," IWBF 2018 - Proc. 2018 6th Int. Work. Biometrics Forensics, pp. 1–7, Jun. 2018, doi: 10.1109/IWBF.2018.8401551.

[20] K. Shirahama and M. Grzegorzek, "On the Generality of Codebook Approach for Sensor-Based Human Activity Recognition," Electron. 2017, Vol. 6, Page 44, vol. 6, no. 2, p. 44, Jun. 2017, doi: 10.3390/ELECTRONICS6020044.

[21] L. Köping, K. Shirahama, and M. Grzegorzek, "A general framework for sensor-based human activity recognition," Comput. Biol. Med., vol. 95, pp. 248–260, Apr. 2018, doi: 10.1016/J.COMPBIOMED.2017.12.025.

[22] M. H. Khan, F. Li, M. S. Farid, and M. Grzegorzek, "Gait Recognition Using Motion Trajectory Analysis," Adv. Intell. Syst. Comput., vol. 578, pp. 73–82, 2018, doi: 10.1007/978-3-319-59162-9_8.

[23] R. HECHT-NIELSEN, "Theory of the Backpropagation Neural Network," Neural Networks Percept., pp. 65–93, Jan. 1992, doi: 10.1016/B978-0-12-741252-8.50010-8.

[24] S. Chung, J. Lim, K. J. Noh, G. Kim, and H. Jeong, "Sensor Data Acquisition and Multimodal Sensor Fusion for Human Activity Recognition Using Deep Learning," Sensors 2019, Vol. 19, Page 1716, vol. 19, no. 7, p. 1716, Apr. 2019, doi: 10.3390/S19071716.

[25] S. Mohsen, "Recognition of human activity using GRU deep learning algorithm," Multimed. Tools Appl., vol. 82, no. 30, pp. 47733–47749, Dec. 2023, doi: 10.1007/S11042-023-15571-Y/FIGURES/11.

[26] M. N. Haque, M. T. H. Tonmoy, S. Mahmud, A. A. Ali, M. A. H. Khan, and M. Shoyaib, "GRU-based Attention Mechanism for Human Activity Recognition," 1st Int. Conf. Adv. Sci. Eng. Robot. Technol. 2019, ICASERT 2019, May 2019, doi: 10.1109/ICASERT.2019.8934659.

[27] N. Dua, S. N. Singh, and V. B. Semwal, "Multi-input CNN-GRU based human activity

recognition using wearable sensors," Computing, vol. 103, no. 7, pp. 1461–1478, Jul. 2021, doi: 10.1007/S00607-021-00928-8/METRICS.

[28] G. Khodabandelou, H. Moon, Y. Amirat, and S. Mohammed, "A fuzzy convolutional attention-based GRU network for human activity recognition," Eng. Appl. Artif. Intell., vol. 118, p. 105702, Feb. 2023, doi: 10.1016/J.ENGAPPAI.2022.105702.

[29] S. Mekruksavanich and A. Jitpattanakul, "Deep Convolutional Neural Network with RNNs for Complex Activity Recognition Using Wrist-Worn Wearable Sensor Data," Electron. 2021, Vol. 10, Page 1685, vol. 10, no. 14, p. 1685, Jul. 2021, doi: 10.3390/ELECTRONICS10141685.

[30] L. Tong, H. Ma, Q. Lin, J. He, and L. Peng, "A Novel Deep Learning Bi-GRU-I Model for Real-Time Human Activity Recognition Using Inertial Sensors," IEEE Sens. J., vol. 22, no. 6, pp. 6164–6174, Mar. 2022, doi: 10.1109/JSEN.2022.3148431.

[31] S. Grossberg, "Recurrent neural networks," Scholarpedia, vol. 8, no. 2, p. 1888, 2013, doi: 10.4249/SCHOLARPEDIA.1888.

[32] S. Batool, M. H. Khan, and M. S. Farid, "An ensemble deep learning model for human activity analysis using wearable sensory data," Appl. Soft Comput., vol. 159, p. 111599, Jul. 2024, doi: 10.1016/J.ASOC.2024.111599.

[33] M. A. Khatun et al., "Deep CNN-LSTM With Self-Attention Model for Human Activity Recognition Using Wearable Sensor," IEEE J. Transl. Eng. Heal. Med., vol. 10, 2022, doi: 10.1109/JTEHM.2022.3177710.

[34] S. Perez-Gamboa, Q. Sun, and Y. Zhang, "Improved Sensor Based Human Activity Recognition via Hybrid Convolutional and Recurrent Neural Networks," Inert. 2021 - 8th IEEE Int. Symp. Inert. Sensors Syst. Proc., Mar. 2021, doi: 10.1109/INERTIAL51137.2021.9430460.

[35] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," Pattern Recognit. Lett., vol. 119, pp. 3–11, Mar. 2019, doi: 10.1016/J.PATREC.2018.02.010.

[36] J. Wu, "Introduction to Convolutional Neural Networks," 2017.

[37] H. Hewamalage, C. Bergmeir, and K. Bandara, "Recurrent Neural Networks for Time Series Forecasting: Current status and future directions," Int. J. Forecast., vol. 37, no. 1, pp. 388–427, Jan. 2021, doi: 10.1016/J.IJFORECAST.2020.06.008.

[38] S. Wager, S. Wang, and P. Liang, "Dropout Training as Adaptive Regularization," Adv. Neural Inf. Process. Syst., Jul. 2013, Accessed: Oct. 22, 2024. [Online]. Available: https://arxiv.org/abs/1307.1493v2

[39] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human Activity Recognition on Smartphones Using a Multiclass Hardware-Friendly Support Vector Machine," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7657 LNCS, pp. 216–223, 2012, doi: 10.1007/978-3-642-35395-6_30.

[40] "Deep learning for computer vision: image classification, object detection, and face recognition in python - Google Search." Accessed: Oct. 22, 2024. [Online]. Available: https://www.google.com/search?q=Deep+learning+for+computer+vision%3A+image+classification%2C+object+detection%2C+and+face+recognition+in+python&oq=Deep+learning+for+computer+vision%3A+image+classification%2C+object+detection%2C+and+face+recognition+in+python&gs_lcrp=EgZjaHJvbWUyBggAEEUYOdIBBzk1OGowajSoAgCwAgE&sourceid=chrome&ie=UTF-8

[41] Y. Bengio, "Practical Recommendations for Gradient-Based Training of Deep Architectures," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7700 LECTURE NO, pp. 437–478, 2012, doi: 10.1007/978-3-642-35289-8_26.

[42] N. Ashfaq, M. H. Khan, and M. A. Nisar, "Identification of Optimal Data Augmentation Techniques for Multimodal Time-Series Sensory Data: A Framework," Inf. 2024, Vol. 15, Page 343, vol. 15, no. 6, p. 343, Jun. 2024, doi: 10.3390/INFO15060343.

[43] Y. Wang et al., "A Novel Deep Multifeature Extraction Framework Based on Attention Mechanism Using Wearable Sensor Data for Human Activity Recognition," IEEE Sens. J., vol. 23, no. 7, pp. 7188–7198, Apr. 2023, doi: 10.1109/JSEN.2023.3242603.

[44] S. Gupta, "Deep learning based human activity recognition (HAR) using wearable sensor data," Int. J. Inf. Manag. Data Insights, vol. 1, no. 2, p. 100046, Nov. 2021, doi: 10.1016/J.JJIMEI.2021.100046.

[45] V. Soni, H. Yadav, V. B. Semwal, B. Roy, D. K. Choubey, and D. K. Mallick, "A Novel Smartphone-Based Human Activity Recognition Using Deep Learning in Health care," Lect. Notes Electr. Eng., vol. 946, pp. 493–503, 2023, doi: 10.1007/978-981-19-5868-7_36.

[46] Z. N. Khan and J. Ahmad, "Attention induced multi-head convolutional neural network for human activity recognition," Appl. Soft Comput., vol. 110, p. 107671, Oct. 2021, doi: 10.1016/J.ASOC.2021.107671.

[47] "Human Activity Recognition using WISDM Datasets." Accessed: Oct. 22, 2024. [Online]. Available: https://www.researchgate.net/publication/372420736_Human_Activity_Recognition_using_WISDM_Datasets

[48] M. M. Afsar et al., "Body-Worn Sensors for Recognizing Physical Sports Activities in Exergaming via Deep Learning Model," IEEE Access, vol. 11, pp. 12460–12473, 2023, doi: 10.1109/ACCESS.2023.3239692.

[49] M. K. R. Al-juaifari and A. A. Athari, "A Novel Framework for Future Human Activity Prediction Using Sensor-Based Data," Int. J. Intell. Eng. Syst., vol. 16, no. 6, pp. 981–991, 2023, doi: 10.22266/ijies2023.1231.81.

[50] V. B. Semwal, N. Gaud, P. Lalwani, V. Bijalwan, and A. K. Alok, "Pattern identification of different human joints for different human walking styles using inertial measurement unit (IMU) sensor," Artif. Intell. Rev., vol. 55, no. 2, pp. 1149–1169, Feb. 2022, doi: 10.1007/S10462-021-09979-X/METRICS.

[51] P. Duan, C. Li, J. Li, X. Chen, C. Wang, and E. Wang, "WISDOM: Wi-Fi-Based Contactless Multiuser Activity Recognition," IEEE Internet Things J., vol. 10, no. 2, pp. 1876–1886, Jan. 2023, doi: 10.1109/JIOT.2022.3210131.