HUMAN ACTIVITY RECOGNITION BASED ON IMU SENSORS USING A COMBINATION OF CONVOLUTIONAL NEURAL NETWORK AND MULTI-HEAD ATTENTION

Van Khanh Tran, Duy Khanh Ninh*

The University of Danang – University of Science and Technology, Vietnam

*Corresponding author: nkduy@dut.udn.vn

(Received: May 16, 2025; Revised: June 18, 2025; Accepted: June 20, 2025)

DOI: 10.31130/ud-jst.2025.23(9C).534E

Abstract - Human activity recognition (HAR) and fall detection play crucial roles in healthcare for the elderly and remote observation. This research presents three innovative deep learning models - MSRLSTM, MSRLSTM - Refined, and MSR - MultiHeadAttention-designed for HAR and fall detection by utilizing Inertial Measurement Unit data from the UP-Fall Detection Dataset. By employing convolutional neural networks, residual learning, and multi-head attention, these models effectively capture complex temporal and spatial patterns present in multimodal sensor data. When evaluated on the UP-Fall Detection Dataset, MSRLSTM-Refined and MSR - MultiHeadAttention achieved accuracies of 93.91% and 95.49%, respectively, outpacing the baseline MSRLSTM (92.10%). The MSR-MultiHeadAttention model stands out due to its precision and temporal modeling capabilities, while MSRLSTM-Refined delivers computational efficiency suitable for wearable devices. Although there are challenges in differentiating similar motion patterns, these models demonstrate significant potential for real-time fall detection, contributing to remote healthcare monitoring solutions and related fields.

Key words - Human activity recognition; fall detection; deep learning; inertial measurement units; UP-Fall detection dataset; convolutional neural networks; multi-head attention; residual learning; elderly healthcare; wearable sensors.

1. Introduction

In recent years, with the rapid advancement of Artificial Intelligence (AI) and Deep Learning (DL), along with the growing aging population, Human Activity Recognition (HAR) has gained increasing attention and significant development, especially due to its essential applications in fields such as work safety and remote healthcare monitoring. HAR involves the automatic identification and classification of human actions using data from diverse sources, including wearable sensors, smartphones, cameras (RGB and depth), radar, and multimodal devices [1]. This field has a significant role not only in ensuring workplace safety by identifying hazardous activities but also in smart home activity monitoring, elderly healthcare, security surveillance, sports analytics, and human-robot interaction [2]. Among these, fall detection and activity classification systems are increasingly vital due to the aging global population and the growing need for remote health supervision. Additionally, falls are a leading cause of injury and mortality among the elderly, resulting in over 37.3 million severe injuries and 684,000 deaths annually worldwide, making it the second leading cause of unintentional injury death, after road traffic injuries [3].

Most of the methods and research in HAR focus on prediction based on cameras and Inertial Measurement Units (IMUs). Although showing positive results, the camera approach also has some disadvantages. Not only is the installation costly, but cameras are also ineffective in areas where visual coverage is not available. One example is the study of a real-time fall detection model using Uniformer with RGB video input, processed in small segments. By using a lightweight network and sliding window method, the model achieves good performance while maintaining low latency [5].

In contrast, IMU sensors offer a more practical alternative, as they provide a low-cost, non-intrusive, and energy-efficient solution, making them ideal for a wide range of applications. However, translating raw sensor data into meaningful insights requires sophisticated algorithms capable of learning complex temporal patterns and distinguishing subtle differences in human movements. For example, CNNs excel at feature extraction from time-series data, while RNNs, particularly Long Short-Term Memory (LSTM) networks, are adept at capturing temporal correlations [4].

This study presents a deep learning-based approach for HAR and fall detection using IMU sensor data from the UP-Fall Detection Dataset. We introduce three novel models-MSRLSTM, MSRLSTM-Refined, and MSR-MultiHeadAttention designed to leverage temporal dynamics and discriminative features in motion signals. These models address challenges in computational efficiency and generalization for low-frequency, multimodal IMU data. MSRLSTM-Refined optimizes the original architecture with a refined multilayer perceptron and dropout regularization, while MSR-MultiHeadAttention incurporates a Multi-Head Attention mechanism to enhance temporal modeling.

2. Related works

2.1. UP-Fall detection dataset: a multimodal approach

The UP-Fall Detection Dataset is a publicly available multimodal dataset designed to advance research in HAR and fall detection. Researchers collected the dataset from 17 healthy young adults (9 male and 8 female), aged 18–24 years, with an average height of 1.66 m and weight of 66.8 kg. Participants performed 11 activities, comprising 6 daily living activities (walking, standing, sitting, picking up an object, jumping, and laying) and 5 types of falls (falling forward using hands, falling forward using knees, falling backwards, falling sideward, and falling sitting in an empty chair). Each activity was repeated three times, resulting in 33 activity instances per participant [6]. The dataset is multimodal, integrating data from:

- Wearable Sensors: Five IMUs positioned at the left wrist, under the neck, right pocket, middle waist, and left ankle, each providing 3-axis accelerometer and 3-axis gyroscope data.

- Ambient Sensors: Six infrared sensors to detect presence and motion within the environment.

- Vision Devices: Two cameras (RGB and depth) to capture visual information of the activities.

The dataset totals approximately 850 GB, encompassing 296,364 samples collected at an average sampling rate of 18.4 Hz. It provides both raw sensor data and pre-extracted features, making it versatile for various machine learning and deep learning approaches [6]. The authors conducted extensive experiments on various data fusion strategies, employing different models with three different window sizes: 1 second, 2 seconds, and 3 seconds. These experiments provide a comprehensive overview of the importance of selecting appropriate input data, machine learning models, and window sizes. Notably, using only IMU data still yielded promising results, achieving an F1-score of up to 70.31 ± 1.48 (%).

2.2. Combining residual and LSTM recurrent networks for transportation mode detection using multimodal sensors integrated in smartphones

Yu et al. [7] proposed the Multimodal Sensor Residual LSTM (MSRLSTM) model for transportation mode detection using multimodal sensor data from smartphones, including accelerometers, gyroscopes, magnetometers, and pressure. The MSRLSTM model combines residual learning with LSTM recurrent networks to effectively capture spatial and temporal dependencies in sensor data. The residual blocks extract high-level spatial features from raw sensor inputs, while the LSTM layers model temporal sequences, enabling the identification of complex patterns in activities such as walking, cycling, and driving. The designed MSRLSTM architecture is shown in Figure 1.

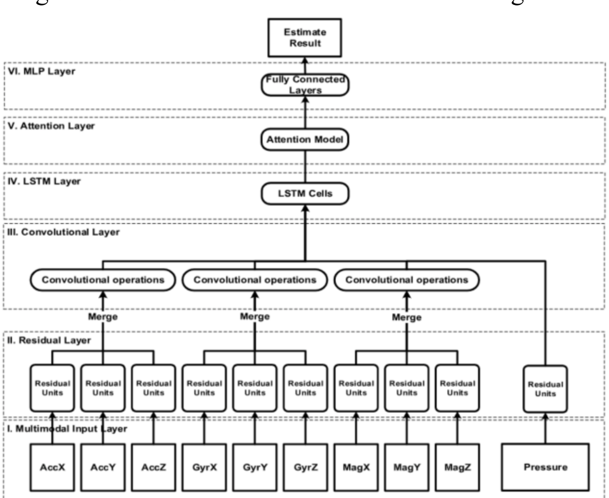

Figure 1. *The architecture of the MSRLSTM model [7]*

Evaluated on the SHL dataset, the model achieves an accuracy of 98.27%, outperforming standalone CNN and LSTM models by leveraging the complementary strengths of both architectures. The study emphasizes the importance of multimodal sensor fusion for robust activity recognition but does not address fall detection explicitly. While the MSRLSTM model is computationally intensive, its architecture inspires our proposed work, which adapts similar residual and recurrent structures for fall detection using IMU data from the UP-Fall dataset, optimizing for lower computational overhead.

The MSRLSTM model's success in transportation mode detection inspires our proposed models, which adapt its architecture for fall detection using the UP-Fall Detection Dataset, as described in the following sections.

3. Methodology

3.1. MSRLSTM-Refined model

The MSRLSTM-Refined model inherits the residual block structure and LSTM layers from Yu et al. [7]. However, we introduce several novel modifications: (1) an expanded input layer to accommodate five IMU sensors, increasing input dimensions five-fold (Ankle, Right Pocket, Belt, Neck, and Wrist); (2) a refined MLP with decreasing unit sizes to reduce computational complexity; and (3) a dropout rate of 0.3 to prevent overfitting. These changes address the lower sampling rate (18.4 Hz) and multimodal nature of the UP-Fall Detection Dataset. Unlike the 100 Hz SHL dataset used by Yu et al. for transportation mode detection, the MSRLSTM-Refined model is optimized for human activity recognition and fall detection using the UP-Fall Detection Dataset. The model architecture was refined as follows:

- Input Layer: The input layer processes data from five IMU sensors, each providing 3-axis accelerometer and 3-axis gyroscope signals, resulting in a five-fold increase in input dimensions compared to the original MSRLSTM model designed for the 100 Hz SHL dataset. For a window of 100 samples (approximately 5 seconds at 18.4 Hz), the input shape is $[100{,}5 \times 6]$, where 5 represents the number of sensors and 6 represents the accelerometer and gyroscope axes.

- Residual Blocks and Convolutional Layer: The architecture comprises four residual blocks with filter sizes ([64, 128, 128, 128]) and kernel sizes ([3, 2, 2, 4]). Each block uses a pooling size of 2, a filter size of 3, and a stride of 4 to extract robust spatial features while mitigating vanishing gradients. The residual connection is defined as:

$$y = F(x, W_i) + x \quad , \quad (1)$$

where x is the input to the block, $F(x, W_i)$ is the residual function (two convolutional layers with ReLU activation), and y is the output. Each convolutional layer within a block is computed as:

$$z = \sigma(W * x + b), \quad (2)$$

where W is the convolutional filter, b is the bias, and σ is the ReLU activation function ($\sigma(z) = \max(0, z)$). A subsequent convolutional layer with 32 filters and a kernel size of 3 combines spatial features (see Figure 2 and Figure 3).

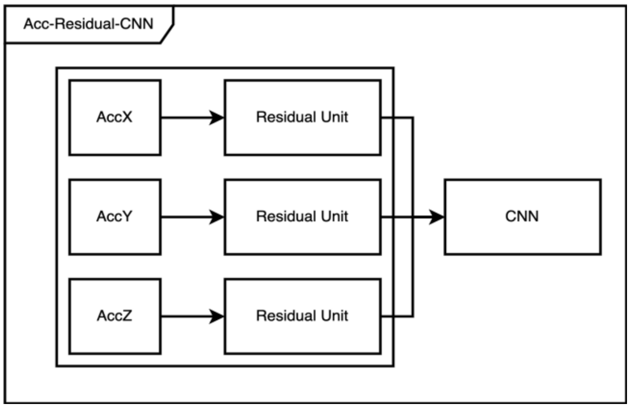

Figure 2. *Acc-Residual-CNN block*

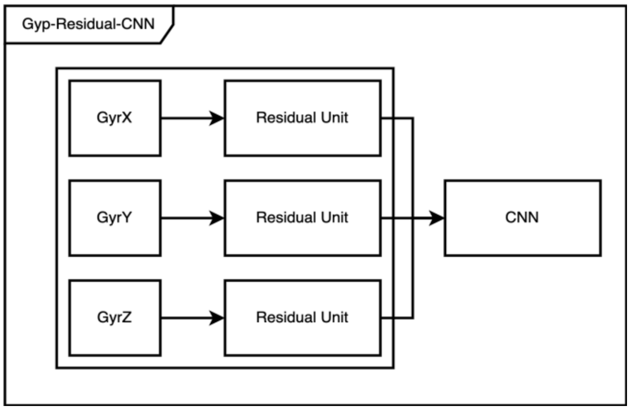

Figure 3. *Gyp-Residual-CNN block*

- LSTM Layers: Three LSTM layers with units ([256, 36, 128]) model temporal dependencies. The increased unit count in the first layer accommodates the expanded input dimensions. The LSTM cell updates are defined as:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right), \tag{3}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{4}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \tag{5}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh(W_c \cdot [h_{t-1}, x_t] + b_c), \tag{6}$$

$$h_t = o_t \cdot \tanh(c_t), \tag{7}$$

where f_t, i_t, and o_t are the forget, input, and output gates, c_t is the cell state, h_t is the hidden state, x_t is the input at time t, (W) and (b) are weights and biases, and σ is the sigmoid function. These layers capture temporal patterns in the IMU data.

- The MLP consists of dense layers with units ([256, 128, 64, 7]), replacing the original increasing unit structure ([128, 256, 512, 1024, 8]) to reduce computational complexity while maintaining discriminative power for the seven activity classes (six daily activities and one fall type). The output is computed as:

$$\hat{y} = \text{softmax}(W \cdot h + b), \tag{8}$$

where h is the output from the final LSTM layer, W and b are the weights and bias of the final dense layer, and $\hat{y}$ is the predicted probability distribution over the seven classes. A dropout rate of 0.3 is applied to the dense layers to prevent overfitting.

3.2. MSR-MultiHeadAttention model

The MSR-MultiHeadAttention model builds on the MSRLSTM-Refined architecture by replacing LSTM layers with a Multi-Head Attention mechanism (Figure 4), a novel contribution inspired by Transformer architectures. This change enhances the model's ability to capture long-range temporal dependencies, addressing the limitations of LSTM layers for low-frequency, multimodal IMU data. This model is specifically designed to handle the characteristics of the UP-Fall Detection Dataset, which differs from the original dataset used for the original MSRLSTM model.

The original MSRLSTM model processed IMU sensor data sampled at 100Hz, whereas the UP-Fall dataset provides data at a lower and less stable sampling frequency of approximately 18-21Hz. Additionally, integrating data from five IMU sensors (Ankle, Right Pocket, Belt, Neck, and Wrist) poses challenges for LSTM layers, which struggle to model long-range dependencies across large, multimodal datasets efficiently. To overcome these issues, we adopt a Multi-Head Attention mechanism to ensure robust temporal sequence learning. The Multi-Head Attention mechanism uses 8 heads, each with a dimension of 64, to capture diverse temporal patterns across the five IMU sensors. This configuration was selected based on empirical experiments showing improved accuracy over 4 or 16 heads. The attention mechanism's scaling factor normalizes the dot product to prevent large values, ensuring stable training. The attention mechanism for a single head is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{9}$$

where Q, K, and V are the query, key, and value matrices derived from the input sequence, and $d_k = 64$ is the dimension of each head. The scaling factor $\sqrt{d_k}$ normalizes the dot product to ensure stable training. For multiple heads, the mechanism is:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_8)W^O, \tag{10}$$

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \tag{11}$$

where W_i^Q, W_i^K, and W_i^V are the weight matrices for the i_{th} head, and W^O is the output projection matrix. This configuration allows the model to focus on critical temporal features, improving discrimination of activities like "falling" and "laying".

4. Experimental results

4.1. Dataset and parameter setting

In this paper, we utilize the UP-Fall Detection Dataset, a comprehensive multimodal dataset designed to facilitate research in HAR and fall detection [6]. This dataset is particularly suitable for our study due to its inclusion of data from wearable IMUs, which align with our focus on developing efficient, real-time fall detection systems using IMU sensor data.

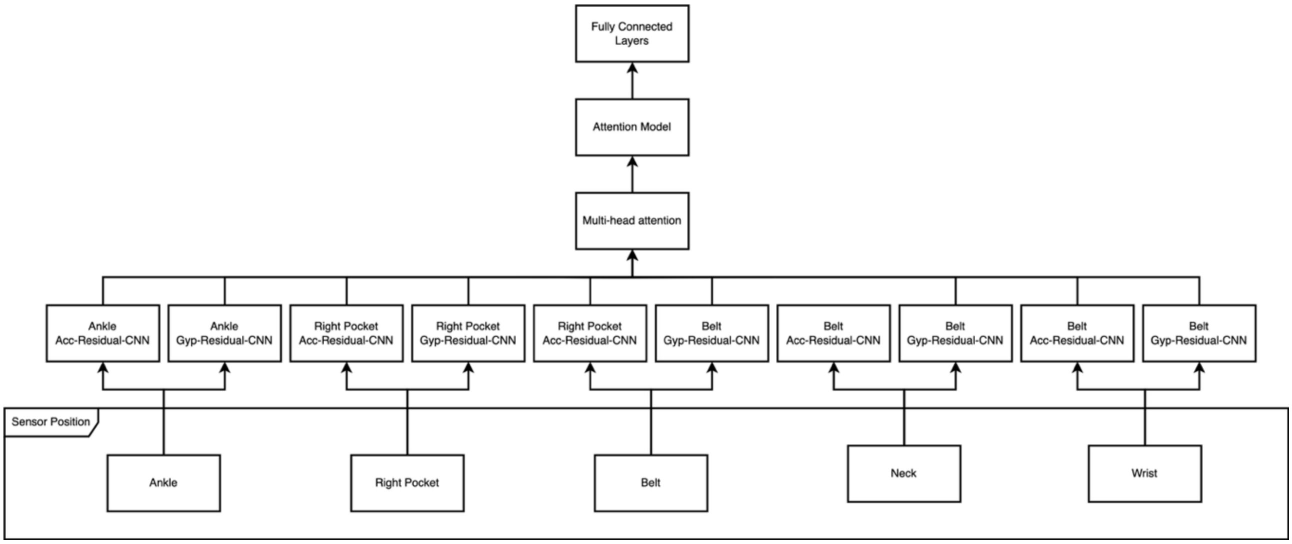

Figure 4. *The architecture of our proposed MSR-MultiHeadAttention model*

For this study, we focus exclusively on the IMU sensor data from the five wearable sensors, as they are most relevant to our proposed models and are intended to develop lightweight, wearable-based solutions. To streamline the classification task and focus on distinguishing falls from non-fall activities, we consolidate the five fall-related activities into a single *"falling"* class. The resulting activity classes are presented in the following Table 1.

Table 1. *Activities performed by the subject after combining all the falling classes*

Activity ID	Description	Duration (s)
1	Falling	10
2	Walking	60
3	Standing	60
4	Sitting	60
5	Picking up an object	10
6	Jumping	30
7	Laying	60

For model evaluation, we adopt the dataset's original train-test split to ensure consistency with prior studies and fair performance comparisons. The split is as follows:

- Train Set: Data from subjects 1, 3, 4, 7, and 10–14, comprising 70% of the dataset.

- Test Set: Data from subjects 15–17, comprising the remaining 30%.

The dataset authors trained the MSRLSTM model and evaluated the detection results by using windows of 1-second duration without overlapping [6]. However, we conducted a detailed analysis of the dataset and determined that a 1-second window is insufficient to capture the full context of certain activities, particularly falls, which typically occur within 2–4 seconds. Using a short window risk mislabeling segments (e.g., initial standing segments in a 10-second fall sequence being classified as *"falling"*). To address this, we select a window size of 100 samples (approximately 5 seconds at 18.4 Hz) with a 50% overlap for the training data to ensure sufficient temporal context. For the test data, we are also using a window size of 100 samples, but without overlap.

4.2. Comparative results

To evaluate the effectiveness of our proposed models, we compare the performance of the original MSRLSTM model, the MSRLSTM-Refined model, and the MSR-MultiHeadAttention model on the UP-Fall Detection Dataset. The original MSRLSTM model, inspired by Yu et al. [7], serves as a baseline to highlight the improvements introduced by our architectural modifications. The MSR-LSTM-Refined model optimizes the original architecture for the UP-Fall dataset's multimodal IMU data, while the MSR-MultiHeadAttention model incorporates a Multi-Head Attention mechanism to enhance temporal sequence modeling. These comparisons demonstrate the necessity of tailored model designs to address the challenges of low sampling rates (approximately 18.4 Hz) and the integration of data from five IMU sensors (left wrist, under the neck, right pocket, middle waist, and left ankle).

The models are evaluated on the test set (subjects 15–17), using a window size of 100 samples (approximately 5 seconds) with non-overlapping windows. For training, a 50% overlap is applied to capture sufficient temporal context, as discussed in Section 4.1. Performance is assessed using accuracy, precision, recall, and F1-score. The performance metrics for the three models are presented in Table 2, demonstrating the improvements achieved by our proposed architectures over the baseline. The MSRLSTM-Refined model enhances computational efficiency, while the MSR-MultiHeadAttention model leverages attention mechanisms to improve temporal modeling, resulting in superior accuracy and precision for fall detection tasks. Additionally, the models are trained using the categorical cross-entropy loss function, suitable for multi-class classification:

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y_i}), \qquad (12)$$

where C is the number of classes (7 in this study), y_i is the true label for class i, and $\hat{y_i}$ is the predicted probability for class (i). Dropout regularization (rate of 0.3) is applied to prevent overfitting, particularly in the MSRLSTM-Refined and MSR-MultiHeadAttention models.

4.3. Analysis of comparative results

The results in Table 2 indicate that both the MSRLSTM-Refined and MSR-MultiHeadAttention models outperform the original MSRLSTM model across most metrics, validating the effectiveness of our proposed optimizations. Below is the detailed analysis of the results.

MSRLSTM-Refined: The MSRLSTM-Refined model, optimized for the UP-Fall dataset, shows a notable improvement with an accuracy of 93.91%. The modifications, including an expanded input layer to handle five IMU sensors, a refined multilayer perceptron (MLP) with decreasing unit sizes, and a dropout rate of 0.3, enhance the model's ability to process multimodal data while reducing computational overhead. The slight decrease in precision (90.19%) compared to the baseline suggests a trade-off in favor of improved recall, indicating better detection of fall events. This model demonstrates that targeted architectural adjustments can significantly improve performance on specific datasets without requiring entirely new designs.

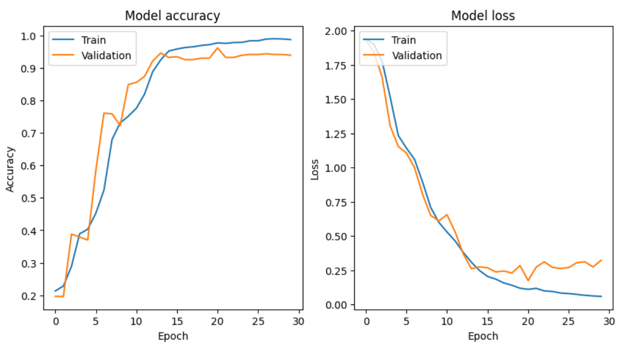

Figure 5. *Training/validation accuracy and loss for MSRLSTM-Refined model*

The loss function trend (Figure 5) shows that the MSRLSTM-Refined model converges steadily over 30 epochs, with no significant signs of overfitting. Both training and validation losses decrease consistently, indicating robust learning and generalization. This stability is attributed to the refined MLP structure and dropout regularization, which mitigate the overfitting issues observed in the baseline model.

The confusion matrix (Figure 6) reveals improved performance over the baseline, particularly in distinguishing *"falling"* and *"walking"* activities, with classification accuracies approaching 99-100%. However, the model still exhibits confusion between *"picking up an object"* and *"standing"*, as well as occasional misclassification of *"laying"* as *"falling"*. These errors are understandable, as *"picking up an object"* and *"standing"* involve similar upper-body movements, while *"laying"* and *"falling"* share transient motion patterns during the initial descent. These findings suggest that while the MSRLSTM-Refined model improves overall performance, further enhancements in feature extraction are needed to address activities with overlapping motion characteristics.

MSR-MultiHeadAttention: The MSR-MultiHeadAttention model achieves the highest performance, with an accuracy of 95.49%, a precision of 96.00%, and an F1 score of 91.08%. Based on MSRLSTM-Refined and by replacing LSTM layers with a Multi-Head Attention mechanism, inspired by Transformer architectures, the model effectively captures long-range temporal dependencies and focuses on critical motion patterns. This is particularly beneficial for the UP-Fall dataset's lower sampling rate and multimodal inputs, which challenge traditional recurrent architectures. The model's ability to prioritize relevant time steps contributes to its superior precision, reducing false positives in fall detection. However, its slightly lower recall (89.63%) compared to MSRLSTM-Refined suggests that it may miss some events, though its overall balance of metrics indicates robust performance. With approximately 22 million parameters, the model is computationally intensive but converges faster, requiring only 20 epochs due to the efficiency of the attention mechanism.

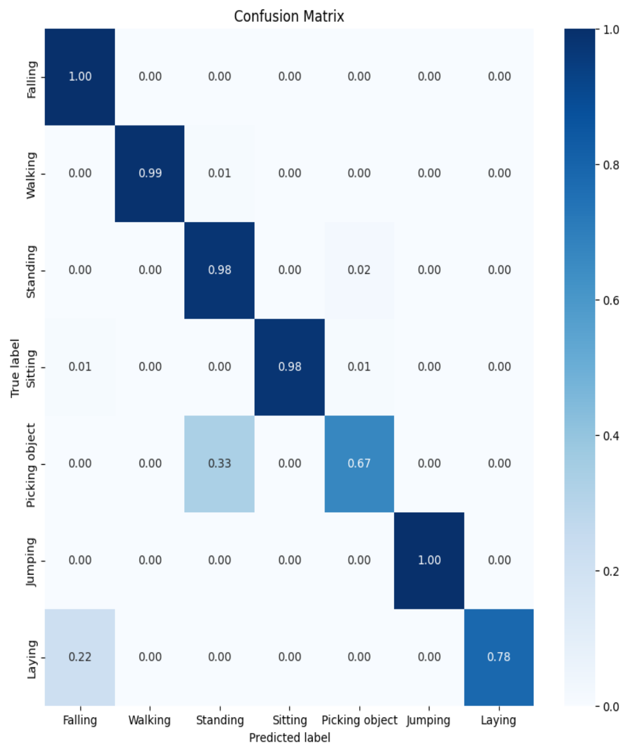

Figure 6. *Confusion matrix for MSRLSTM-Refined model*

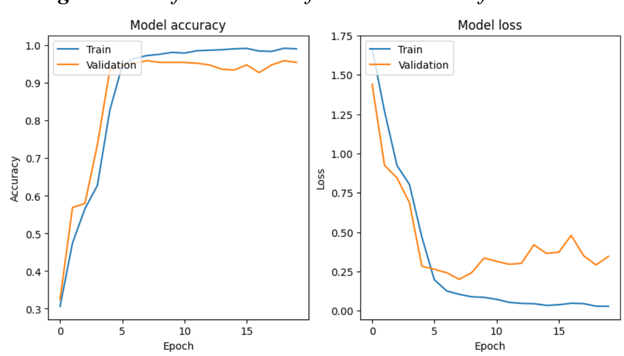

Figure 7. *Training/validation accuracy and loss for MSR-MultiHeadAttention model*

The loss function trend (Figure 7) indicates early convergence at epoch 6, with both training and validation losses stabilizing thereafter. The rapid convergence and lack of overfitting highlight the Multi-Head Attention mechanism's ability to efficiently learn complex patterns in the dataset. Model performance is consistent across epochs, indicating strong test generalization.

Table 2. *Comparison of model performance*

Model	Epoch	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
MSRLSTM	30	92.10	91.97	89.96	90.38
MSRLSTM-Refined	30	93.91	90.19	91.33	90.17
MSR-MultiHeadAttention	20	95.49	96.00	89.63	91.08

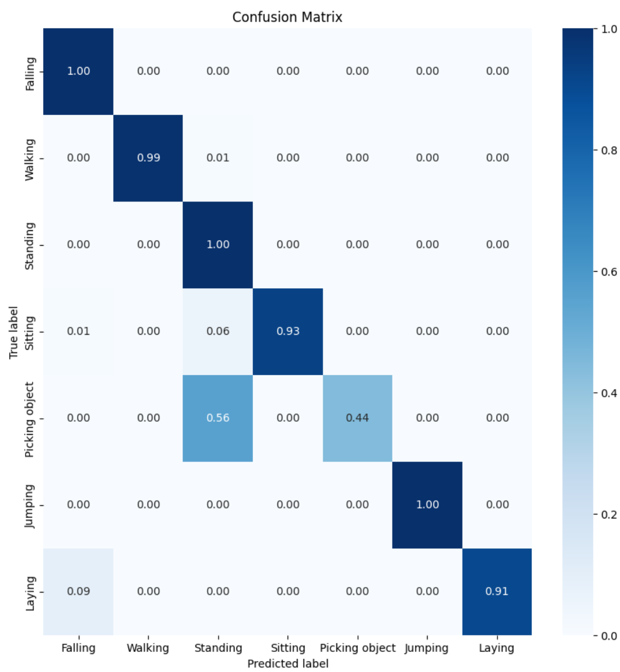

Figure 8. *Confusion matrix for MSR-MultiHeadAttention model*

The confusion matrix (Figure 8) shows that the MSR-MultiHeadAttention model significantly improves the differentiation between *"falling"* and *"laying"* compared to the MSRLSTM and MSRLSTM-Refined models. This improvement is attributed to the attention mechanism's ability to focus on key temporal features, such as the rapid acceleration changes during falls. This improvement is attributed to the attention mechanism's ability to focus on key temporal features, such as the rapid acceleration changes during falls. However, the model exhibits a trade-off, with increased misclassification of *"picking up an object"* as *"standing"* compared to the MSRLSTM and MSRLSTM-Refined models. This suggests that while the attention mechanism enhances the detection of dynamic activities, it may overemphasize certain motion patterns, leading to errors in activities with subtle differences.

5. Conclusion

This study presents a comprehensive evaluation of three deep learning models-MSRLSTM, MSRLSTM-Refined, and MSR-MultiHeadAttention for HAR and fall detection using IMU sensor data from the UP-Fall Detection Dataset. The results demonstrate that our proposed models, MSRLSTM-Refined and MSR-MultiHeadAttention, significantly outperform the baseline MSRLSTM model, achieving accuracies of 93.91% and 95.49%, respectively, compared to 92.10% for the baseline. These improvements validate the effectiveness of our architectural optimizations, including the refined MLP structure and dropout regularization in MSRLSTM-Refined, and the Multi-Head Attention mechanism in MSR-MultiHeadAttention.

The MSR-MultiHeadAttention model stands out for its superior accuracy and precision, driven by capturing long-range dependencies and emphasizing key motion. This makes it particularly suitable for fall detection, where distinguishing falls from similar activities like *"laying"* is crucial. However, both models face challenges in distinguishing activities with similar motion patterns, such as *"picking up an object"* and *"standing"*, highlighting the need for enhanced feature extraction techniques.

This work advances wearable-based HAR and fall detection systems for remote health monitoring. Despite the UP-Fall Detection Dataset using data from young adults, future research will explore frequency-domain features and additional sensor modalities to improve discrimination of similar activities and ensure robustness across diverse populations.

Acknowledgments: This work was supported by The University of Danang – University of Science and Technology under the Project No. T2024-02-58.

REFERENCES

[1] R. Liu, A. A. Ramli, H. Zhang, E. Henricson, and X. Liu, "An Overview of Human Activity Recognition Using Wearable Sensors: Healthcare and Artificial Intelligence", In: Tekinerdogan, B., Wang, Y., Zhang, LJ. (eds) Internet of Things – ICIOT 2021, *Lecture Notes in Computer Science*, vol. 12993, pp. 1–14, 2022. https://doi.org/10.1007/978-3-030-96068-1_1

[2] M. A. Hossen and P. E. Abas, "Machine Learning for Human Activity Recognition: State-of-the-Art Techniques and Emerging Trends", *Journal of Imaging*, vol. 11, no. 3, 91, 2025. https://doi.org/10.3390/jimaging11030091

[3] A. Joseph, D. Kumar, and M. Bagavandas, "A Review of Epidemiology of Fall among Elderly in India", *Indian Journal of Community Medicine*, vol. 44, no. 2, pp. 166–168, 2019. https://doi.org/10.4103/ijcm.IJCM_201_18

[4] S. W. Pienaar and R. Malekian, "Human Activity Recognition using LSTM-RNN Deep Neural Network Architecture", in *2019 IEEE 2nd Wireless Africa Conference (WAC)*, Pretoria, South Africa, 2019, IEEE, pp. 1–5. https://doi.org/10.1109/AFRICA.2019.8843403

[5] A. Núñez-Marcos and I. Arganda-Carreras, "Transformer-Based Fall Detection in Videos", *Engineering Applications of Artificial Intelligence*, vol. 132, 107937, 2024. https://doi.org/10.1016/j.engappai.2024.107937

[6] L. Martínez-Villaseñor, H. Ponce, J. Brieva, E. Moya-Albor, J. Núñez-Martínez, and C. Peñafort-Asturiano, "UP-Fall Detection Dataset: A Multimodal Approach", *Sensors*, vol. 19, no. 9, 1988, 2019. https://doi.org/10.3390/s19091988

[7] C. Wang, H. Luo, F. Zhao, and Y. Qin, "Combining Residual and LSTM Recurrent Networks for Transportation Mode Detection Using Multimodal Sensors Integrated in Smartphones", *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 9, pp. 5473–5485, 2021. https://doi.org/10.1109/TITS.2020.2987598