



Hybrid Deep Learning Models for Free-Living Imbalanced Human Activity Recognition: Comparative Study

Aditya Heru Prathama^{1*}, Joy Milliaan², Ghandy³

^{1,2}Department of Intelligent Robotics and Electronic Systems, Faculty of Science and Technology, Calvin Institute of Technology, Jakarta, Indonesia

³Department of IT and Big Data Analytics, Faculty of Science and Technology, Calvin Institute of Technology, Jakarta, Indonesia

*aditya.prathama@calvin.ac.id

Abstract

This study presents a comparative evaluation of three hybrid deep learning models for human activity recognition (HAR) in free-living and highly imbalanced conditions: 1DCNN-ResBLSTM-Attention (Model A), Attention-Mechanism-Based Deep Learning Feature Combination (Model B), and Time-Reversal-1DCNN-ResLSTM-Attention (Model C). Each architecture integrates convolutional layers for feature extraction, recurrent networks for temporal modeling, and attention mechanisms to enhance relevant representations. The HARTH v2.0 dataset, comprising 31 subjects and 15 activity classes under strong class imbalance, is used for evaluation. Results show that soft labeling consistently improves performance by better capturing transitional uncertainty in windowed sensor data. Model A achieves the highest accuracy (96.21%) and macro-averaged F1-score (88.17%), followed by Model C with comparable performance at lower computational cost, while Model B underperforms on minority classes due to limitations of spectrogram-based representations. Across all models, persistent confusion is observed among activities with similar motion patterns, such as walking, standing, and shuffling, indicating intrinsic ambiguity in sensor signals. This study provides a controlled and standardized comparison of hybrid architectures under realistic conditions, revealing both performance trade-offs and shared limitations. The findings highlight the importance of modeling uncertainty and temporal context for improving robustness, particularly transitional and underrepresented activities.

Keywords: attention mechanism; deep learning; HARTH; human activity recognition; imbalanced dataset; wearable sensors

How to Cite: A. H. Prathama, Joy Milliaan, and Ghandy, "Hybrid Deep Learning Models for Free-Living Imbalanced Human Activity Recognition: Comparative Study", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 2, pp. 289 - 297, Apr. 2026.

Permalink/DOI: <https://doi.org/10.29207/resti.v10i2.6936>

Received: July 9, 2025

Accepted: March 3, 2026

Available Online: April 6, 2026

This is an open-access article under the CC BY 4.0 License

Published by Ikatan Ahli Informatika Indonesia

1. Introduction

In recent years, the field of human activity recognition (HAR) has attracted significant research interest due to its broad range of applications, including elderly and remote healthcare, robotics, augmented and virtual reality, human-computer interaction, anomaly and fall detection, and surveillance [1]. HAR refers to the automated identification and classification of physical activities performed by individuals, based on visual (vision-based) or sensor data (sensor-based) [2]. Among these, sensor-based HAR is often preferred for real-world applications owing to its lower power consumption and bandwidth requirements, improved privacy and user mobility, and increased robustness to variations in environmental conditions [3].

Despite significant progress, a major challenge remains, i.e. achieving high accuracy in various complex real-world scenarios [4]. To address this, researchers have explored a wide array of modeling approaches, including machine learning, deep learning, attention-based, and transformer-based methods to enhance the accuracy and robustness of HAR systems. Traditional machine learning models typically rely on handcrafted features that depend on the prior knowledge and domain expertise of human designers [5]. While these models perform well in controlled settings, they often struggle to generalize in complex or dynamic real-world environments [6]. On the other hand, deep learning approaches—especially convolutional neural networks (CNNs) and recurrent neural networks (RNNs)—have shown significant advantages by learning multilevel representations directly from raw sensor signals [7].

CNNs effectively capture spatial patterns, whereas RNNs are well-suited for modeling sequential characteristics in time-series data [8].

More recently, transformer architectures have gained increasing attention for their ability to model long-range temporal dependencies and capture contextual information effectively [9]. However, they are computationally expensive and require large volumes of data to perform well. Some researchers leverage transfer learning methods for feature extraction, due to deep learning advancement and robustness [10]. Time-series signals are converted into spectrogram images, which are then fed into EfficientNet_b0 and Mobilenet_v3 transfer learning models for feature extraction. Finally, a wrapper function over a filter-based methodology with binary bat algorithm is applied for optimal feature selection and higher classification performance. Another method proposed in [11] is to pre-train—up to 100,000 hours of data—a transformer encoder network to extract smaller labelled data downstream, to overcome the issue of limited training data in some classes. Researchers in [12] implement the Zero-Shot Learning method—a learning paradigm where a model is trained to recognize classes it has never seen during training—by leveraging data augmentation using sensor context features. This results in better generalizability across benchmark datasets.

Interestingly, hybrid models that combine the strengths of these approaches have been emerging as a promising trend [13]. These architectures combine CNNs for extracting spatial features with RNNs or attention mechanisms for modeling temporal dynamics, allowing them to capture both localized and long-range dependencies in the data [14]. Such systems often demonstrate improved generalization and robustness, especially in complex and dynamic HAR scenarios [15]. As the field continues to evolve, exploring and optimizing these integrated models remains a compelling direction for achieving higher performance and adaptability in diverse real-world applications [16].

In this study, we investigate the performance of three hybrid deep learning models for sensor-based HAR: 1DCNN-ResBLSTM-Attention (hereafter referred to as Model A) [17], Attention-Mechanism-Based Deep Learning Feature Combination (Model B) [18], and Time-Reversal-1DCNN-ResLSTM-Attention (Model C) [19]. All three models share a hybrid architecture design that integrates convolutional and recurrent layers with attention mechanisms, yet differ in how they structure and integrate these components.

While numerous sensor-based datasets are widely used in the literature, such as WISDM [20], KU-HAR [21], and UCI-HAR [22], this study focuses on the HARTH dataset, which was developed by researchers at the Norwegian University of Science and Technology [23]. The dataset was initially released in 2021 (v1.0), containing data from 22 individuals performing activities for about 2 hours in free-living environments. Activities are classified into 12 classes including

walking, running, standing, sitting, lying, and cycling. This version has been extensively utilized in prior sensor-based HAR research [24] – [29]. The latest release, HARTH v2.0, published in late 2024, expands the dataset with 9 new subjects (bringing the total to 31) and 3 new activity classes (totaling 15). To the best of our knowledge, this study is among the earliest to utilize the updated version of HARTH dataset. The increased subject diversity and expanded activity set make it a valuable resource for evaluating deep learning models in multi-class HAR tasks, particularly under realistic conditions and class imbalance [30]. Furthermore, the dataset offers three key strengths: professional activity annotation by human experts, fixed sensor placement and orientation, and use of high-quality wearable sensors [23]. Through this comparative study, we aim to assess the effectiveness and trade-offs of each model design in handling complex, imbalanced, and realistic activity recognition tasks on the HARTH v2.0 dataset. However, comparative accuracy alone does not reveal how architectural decisions influence reliability under imbalance.

Rather than emphasizing incremental performance gains, this study seeks methodological insight through a controlled evaluation of hybrid architectures in realistic and highly imbalanced free-living conditions. Using the recently expanded HARTH v2.0 dataset, this work provides one of the earliest standardized evaluations of CNN–RNN–attention hybrid architectures under consistent experimental settings, ensuring fair comparability. It further investigates the role of soft labeling in modeling transitional activities, demonstrating its impact on generalization and training stability. In addition, the study analyzes per-class reliability, with particular emphasis on minority and ambiguous activities, revealing systematic weaknesses shared across architectures. Moreover, common failure patterns across independently developed models are identified, suggesting limitations imposed by intrinsic signal similarity. Finally, the trade-off between recognition performance and parameter complexity is examined, offering practical insights for deployment in resource-constrained environments.

The remainder of the paper is organized as follows. Section 2 describes the methodology and experimental design, Section 3 presents and interprets the results, and Section 4 concludes the study and outlines directions for future work.

2. Methods

2.1. Model A: 1DCNN-ResBLSTM-Attention

Model A, illustrated in Figure 1, integrates a one-dimensional convolutional neural network (1DCNN), a residual bidirectional long short-term memory (ResBLSTM) module, and an attention mechanism. The 1DCNN extracts local temporal features using four convolutional blocks consisting of convolution, batch normalization, activation, pooling, and dropout layers. Temporal dependencies are modeled using a residual

bidirectional LSTM (ResBLSTM), which combines bidirectional sequence learning with residual connections.

The residual mapping is defined in Equation 1.

$$x^{i+1} = x^i + F(x^i, W_i), \quad (1)$$

x^i and $F(x^i, W_i)$ denote the direct mapping and residual components, respectively. The recursive formulation of the ResBLSTM is given in Equations 2a-2c.

$$x_t^{f(i+1)} = LN[x_t^{f(i)} + L(x_t^{f(i)}, W_i)], \quad (2a)$$

$$x_t^{b(i+1)} = LN[x_t^{b(i)} + L(x_t^{b(i)}, W_i)], \quad (2b)$$

$$y_t = Concat(x_t^f, x_t^b), \quad (2c)$$

$L(\cdot)$ denotes the LSTM transformation with weights W_i , LN represents layer normalization, and t indexes time. The superscripts f and b indicate forward and backward sequences, respectively, while y_t is the concatenated output feature vector.

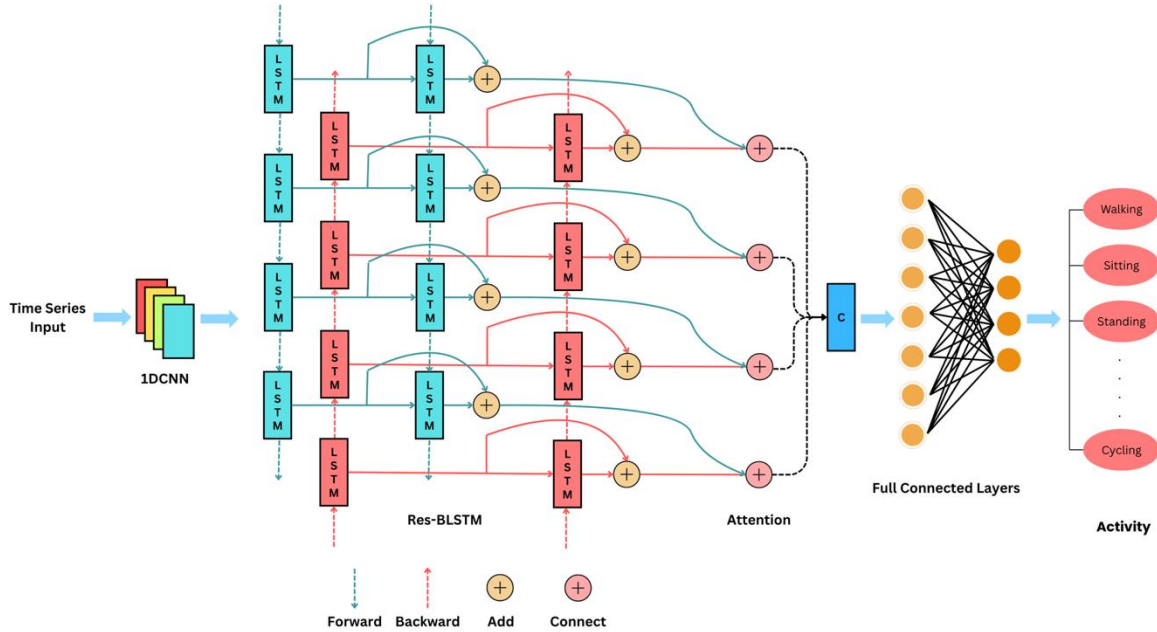


Figure 1. 1DCNN-ResBLSTM-Attention model structure (adapted from [17]).

An attention mechanism is applied to assign adaptive weights to temporal features. The context vector is computed as defined in Equation 3.

$$e_t = a(y_t), \alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^T \exp(e_k)}, c = \sum_{t=1}^T \alpha_t y_t, \quad (3)$$

$a(\cdot)$ is a learnable function, e_t denotes the attention score, and α_t represents the weight associated with each time step t . The resulting features are passed through fully connected layers followed by SoftMax classification.

2.2. Model B: Attention-Mechanism-Based Deep Learning Feature Combination

Model B, referred to as the Attention-Mechanism-Based Deep Learning Feature Combination (AM-DLFC), adopts a spectrogram-based approach and is illustrated in Figure 2. Raw time-series signals are transformed into two-dimensional inputs using short-time Fourier transform (STFT). The signals are segmented using a 2.56-second window with 50% overlap, corresponding to 128 time steps at a sampling rate of 50 Hz, resulting in spectrograms of size 8 (height) x 129 (width) x 6 (channels).

The architecture consists of six convolutional blocks, each comprising a convolutional layer followed by a Convolutional Block Attention Module (CBAM). The

CBAM sequentially applies channel and spatial attention mechanisms to refine feature representations, as defined in Equations 4 and 5,

$$Y = C(X) \otimes X, \quad (4)$$

$$Z = S(Y) \otimes Y, \quad (5)$$

X, Y , and Z denote the input feature map, intermediate output, and final output, respectively, $C(\cdot)$ and $S(\cdot)$ represent the channel and spatial attention functions, and $\otimes$ denotes element-wise multiplication. The extracted feature maps are processed through fully connected layers with progressively reduced dimensions. Dropout is applied to mitigate overfitting, while selective pooling preserves feature resolution. The final output layer consists of 15 neurons corresponding to the activity classes in the HARTH dataset, followed by SoftMax classification.

2.3. Model C: Time-Reversal-1DCNN-ResLSTM-Attention

Model C, referred to as the Time-Reversal-1DCNN-ResLSTM-Attention architecture, is illustrated in Figure 3. Rather than utilizing external data augmentation, it approximates bidirectional temporal modeling by employing an internal sequence-flipping mechanism alongside a residual LSTM (ResLSTM).

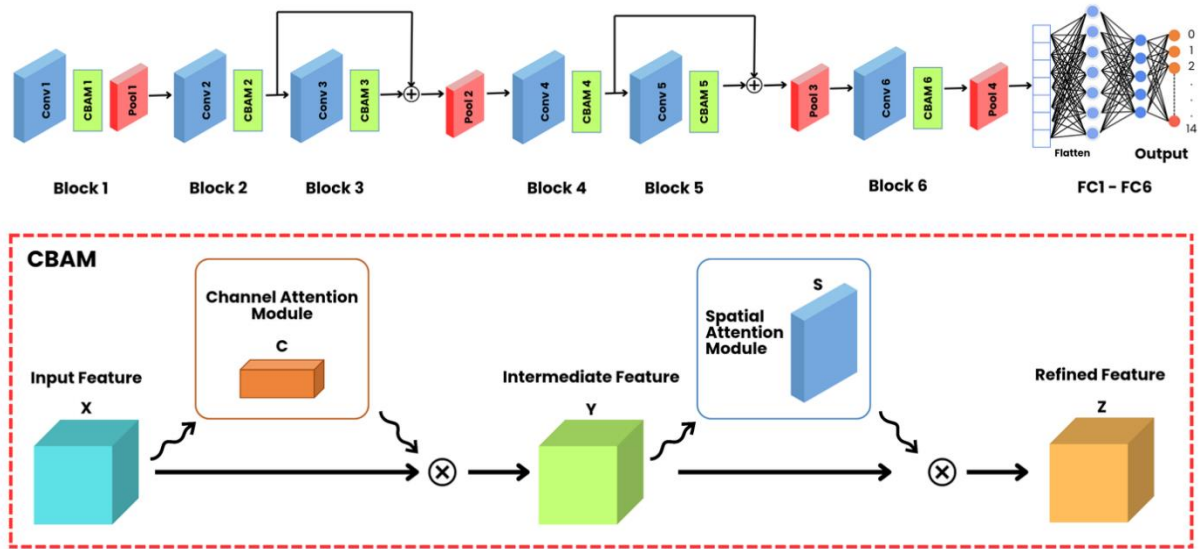


Figure 2. The architecture of the AM-DLFC model (top), and detailed schematic of convolutional block attention module (CBAM) (bottom) (adapted from [18]).

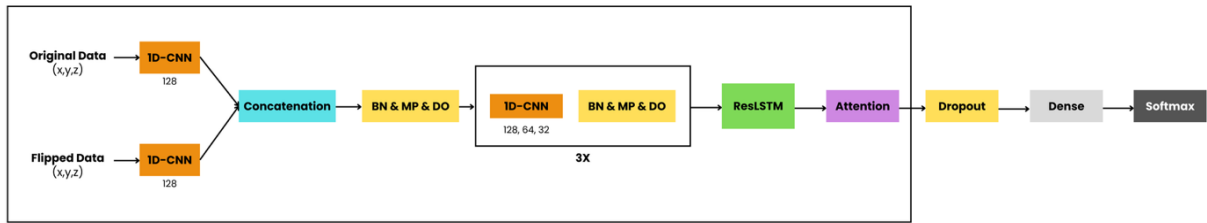


Figure 3. Time-Reversal-1DCNN-ResLSTM-Attention model (adapted from [19]).

The input sequence is processed in both original and reversed forms, as defined in Equations 6a and 6b.

$$X_{flipped}(t) = x_{end}, x_{end-1}, \dots, x_{start+1}, \quad (6a)$$

$$start = \max\left(1, \left\lceil t - \frac{w}{2} \right\rceil\right), end = \min\left(n, \left\lfloor t + \frac{w}{2} \right\rfloor\right) \quad (6b)$$

w denotes the window size, t is the current time index ($1 \leq t \leq n$), with the window centered at t . Two parallel 1DCNN branches extract features from the original and time-reversed sequences. The resulting feature maps are concatenated and passed through batch normalization, max-pooling, and dropout layers, followed by additional convolutional blocks for further feature extraction.

Temporal dependencies are modeled using a ResLSTM module, where residual connections are implemented via point-wise addition as defined in Equation 7.

$$h_{resLSTM} = LSTM(h_{do}) + h_{do}. \quad (7)$$

An attention mechanism subsequently assigns weights to the extracted features, as defined in Equation 8.

$$\alpha_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)}, \quad (8)$$

α_i is the attention weight of feature i , $e_i = \tanh(W_h h_i + b_h)$ is the attention score with W_h and b_h being learnable parameters. The weighted sum of features is computed as given in Equation 9.

$$h_{final} = \sum_i \alpha_i h_i. \quad (9)$$

The resulting representation is processed through dropout and fully connected layers, followed by SoftMax classification.

2.4. Dataset Description and Preprocessing Stage

The study utilizes the HARTH v2.0 dataset [23], comprising 31 individuals performing 15 activity classes in free-living environments for approximately 2 hours. Data are collected using two 3-axis Axivity AX3 accelerometers placed on the lower back and right thigh at a sampling rate of 50 Hz [31], producing six features corresponding to x-, y-, and z-axis accelerations. Activity labels are professionally obtained via synchronized chest-mounted camera recordings.

Activities are labeled with class IDs 1–8, 10, 11, 13, 14, 17, 130, and 140, with no official descriptions for labels 10, 11, and 17. Figure 4 displays the class distribution in minutes, indicating strong class imbalance. No data augmentation is applied in order to evaluate model performance under the original imbalanced conditions.

In the preprocessing stage, data are segmented using a sliding window of 128 time steps with 50% overlap (2.56 seconds per window), resulting in 129,996 samples. The dataset is then split into training, validation, and test sets with an 80:10:10 stratified ratio to preserve class distribution. Furthermore, Z-score normalization is performed on the training set to avoid information leakage and subsequently applied to the validation and test sets.

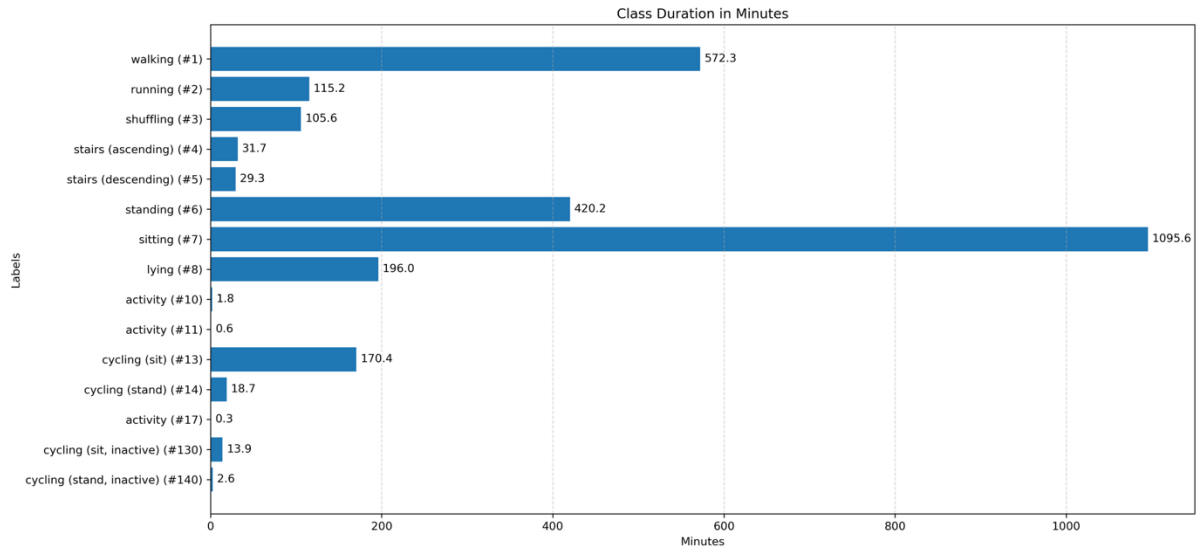


Figure 4. Class distribution in terms of number of minutes for the HARTH dataset.

2.5. Implementation Details

Building upon the preprocessing pipeline described in Section 2.4, all architectures are trained and evaluated under identical experimental conditions to ensure fair comparison. The same inputs, optimization settings, training schedules, and evaluation metrics are used across all models, without architecture-specific tuning, resampling, or augmentation.

TensorFlow and Keras Python libraries are employed to implement the model architectures. Experiments are conducted on a Windows 11 system with an Intel Core i5-13400 CPU, 64 GB of RAM, and an NVIDIA GeForce RTX 4070 GPU. Nadam optimizer is used for Models A and C, while Model B utilizes Adam. The maximum number of epochs is 150, the batch size is 64, the learning rate is 0.001, and the loss function is cross-entropy. This study investigates hard and soft labeling strategies during training. Hard labeling assigns each instance a single class using one-hot encoding, whereas soft labeling represents each instance as a probability distribution over classes. During testing, predictions are assigned to the class with the highest probability.

3. Results and Discussions

Model performance is evaluated using accuracy, precision, recall, and F1-score under both macro- and weighted-averaging schemes to account for class imbalance. Accuracy represents the proportion of correctly classified samples, while precision and recall measure the correctness and completeness of predictions, respectively. The F1-score summarizes the trade-off between precision and recall. Macro-averaged metrics treat all classes equally and emphasize minority class performance, whereas weighted-averaged metrics reflect overall performance by incorporating class distribution. In addition, loss is used to monitor training convergence, and confusion matrices are employed to

analyze class-wise prediction behavior and misclassification patterns.

3.1. Hard and Soft Labeling

Table 1 presents a comparison of the three hybrid models (Models A, B, and C) in terms of computational complexity and classification performance. Complexity is measured by the number of trainable parameters (#Params), while performance is evaluated using Accuracy (Acc.), Precision (Prec.), Recall (Rec.), and F1-score (F1) as defined earlier. To account for class imbalance in the dataset, both Macro-averaged (M) and Weighted-averaged (W) metrics are reported.

The results demonstrate that soft labeling consistently outperforms hard labeling across all architectures in terms of accuracy, indicating improved generalization. In HAR, each windowed sample may capture transitions between activities, which are more appropriately represented by probability distributions rather than single discrete labels. During backpropagation, soft labeling results in smoother gradients, leading to more stable and efficient convergence. This property is particularly beneficial in free-living recordings, where activity boundaries are often ambiguous and multiple activities may coexist within a single window. The effectiveness of soft labeling in modeling uncertainty and class overlap has also been supported by recent studies [13], [15], [24]. Importantly, the observed improvement is not merely numerical but reflects a more faithful representation of real-world activity transitions, where strict class boundaries rarely exist. Therefore, the following discussion focuses on the results obtained using the soft labeling approach.

A closer inspection of Table 1 reveals clear performance differences between the architectures. Under soft labeling, Model A achieves the highest

macro-averaged F1-score (88.17%), exceeding Model C by 3.12 percentage points and Model B by more than 25 percentage points. A similar trend is observed in accuracy, where Model A surpasses Model C by 0.63%. This comparison highlights a trade-off between recognition capability and computational efficiency. The relatively small accuracy gap (0.63%) between Models A and C, despite a substantial difference in parameter count, suggests diminishing returns from increased architectural complexity beyond a certain point.

Table 1. Models' no. of parameters and performance metrics (in %).

Model ID	A		B		C	
#param.	286,903		294,923		195,127	
Labeling	Hard	Soft	Hard	Soft	Hard	Soft
Acc.	95.88	96.21	90.34	90.63	95.20	95.58
Prec. (M)	85.59	87.32	67.05	65.47	86.50	84.70
Prec. (W)	95.74	96.06	90.09	90.44	94.99	95.52
Rec. (M)	87.25	90.35	66.83	62.30	80.36	86.53
Rec. (W)	95.88	96.21	90.34	90.63	95.20	95.58
F1 (M)	85.89	88.17	66.81	63.11	82.49	85.05
F1 (W)	95.77	96.08	90.18	90.49	95.05	95.53

Gray cells show best results. Acc. = accuracy, Prec. = precision, Rec. = Recall, F1 = F1-score, M/W = macro/weighted averaging.

3.2. Model Performance Comparison

The results reveal consistent behavioral differences across architectures, reflecting how representation choice and temporal modeling influence performance under class imbalance. Table 1 shows that LSTM-based Models A and C outperform the spectrogram-based Model B across all performance metrics. For further analysis, Figures 5–7 display the confusion matrices for the three models. Model B, which employs spectrogram inputs, performs reasonably well on dominant activity classes but fails to correctly identify rare classes 10, 11, 17, and 140. This behavior indicates that spectrogram-based representations may compress temporal dynamics into frequency patterns, obscuring short-duration or transitional activities that are critical for minority class recognition, and consequently leading to lower macro-averaged precision, recall, and F1-score.

Similar observations have been reported in prior studies, where spectrogram-based models emphasize broader frequency trends while neglecting fine-grained temporal details essential for rare activity recognition [6], [19]. Under severe class imbalance, such models are also prone to overfitting dominant classes, further limiting generalization to underrepresented activities [4]. Consequently, despite achieving satisfactory overall accuracy, Model B underperforms on macro-level metrics, highlighting a known limitation of spectrogram-based approaches in imbalanced HAR settings. The consistency of these trends across multiple metrics indicates that the observed differences are not artifacts of a particular evaluation measure, but reflect inherent differences in how each architecture captures temporal dependencies and class variability, confirming that architectural design plays a dominant role in minority-class sensitivity.

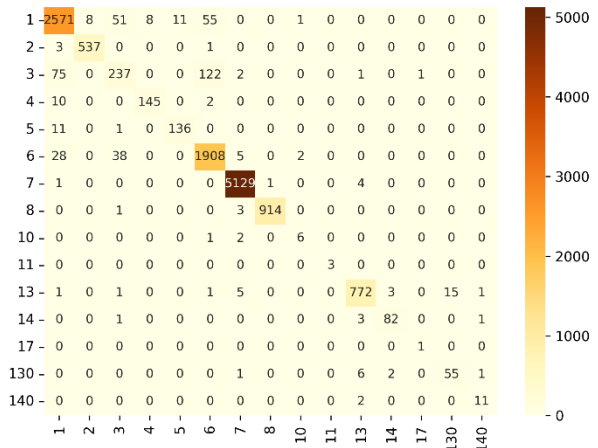


Figure 5. Confusion matrix for the 1DCNN-ResBLSTM-Attention (Model A) with soft labeling approach. Vertical and horizontal axes are true and predicted labels, respectively.

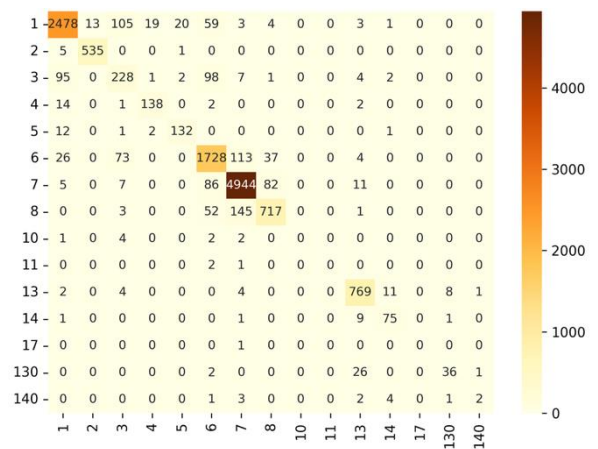


Figure 6. Confusion matrix for the AM-DLFC (Model B) with soft labeling approach. Vertical and horizontal axes are true and predicted labels, respectively.

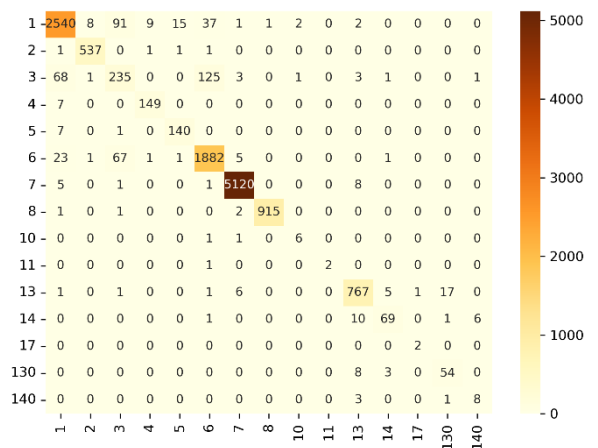


Figure 7. Confusion matrix for the Time-Reversal-1DCNN-ResLSTM-Attention (Model C) with soft labeling approach. Vertical and horizontal axes are true and predicted labels, respectively.

Among the evaluated models, Model A with soft labeling delivers the strongest overall results, achieving the highest accuracy of 96.21%, followed by Model C with 95.58% accuracy. These results highlight the effectiveness of hybrid architectures that integrate residual structures and attention mechanisms into LSTM-based networks. In imbalanced settings, metrics

such as the F1-score provide a more reliable evaluation than accuracy—particularly given the more severe class imbalance in HARTH v2.0 compared to its v1.0 counterpart. Models A and C achieve macro-averaged F1-scores of 88.17% and 85.05%, and weighted-averaged F1-scores of 96.06% and 95.53%, respectively. Model A's superior performance comes at the cost of increased computational complexity, requiring 286,903 parameters compared to Model C's 195,127 (approximately 47% more). Consequently, in resource-constrained environments, Model C may provide a more favorable balance between performance and computational demand.

Further evaluation using confusion matrices reveals that all models consistently struggle to distinguish between walking (activity #1), shuffling/standing with leg movement (activity #3), and standing (activity #6). In particular, the classification performance for shuffling, often misclassified as standing, is notably poor, as evidenced by the low precision, recall, and F1-scores shown in the per-class evaluation results (Figures 8–10). These activities exhibit highly similar motion signatures and often occur in rapid succession, with transitions occurring in as little as 60 ms. As a result, overlapping temporal patterns within a single window can obscure class boundaries, making reliable discrimination difficult [32]. The resulting temporal ambiguity leads to systematic misclassification across models.

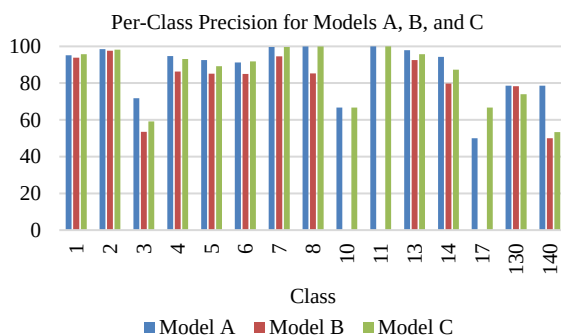


Figure 8. Per-Class Precision for Models A, B, and C.

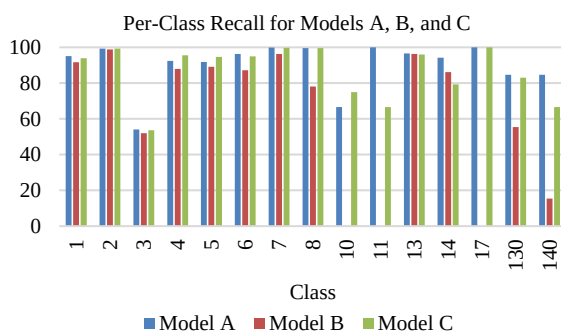


Figure 9. Per-Class Recall for Models A, B, and C.

The performance of Models A and C in classifying rare activity classes—specifically labels 10, 11, 17, and 140—demonstrates notable variability. Model A, in particular, exhibits inconsistent behavior: while it achieves perfect scores (100%) on certain metrics for

some rare classes, its precision for class 17 drops as low as 50%. This variability highlights the instability of decision boundaries for underrepresented classes, where limited training samples lead to high sensitivity to feature representation and model initialization. As a result, although the model can occasionally capture meaningful patterns for rare activities, its predictions are not consistently reliable across evaluation metrics. This indicates that successful recognition of underrepresented behaviors remains fragile and highly dependent on subtle feature distinctions learned during training. These findings underscore a fundamental challenge in imbalanced classification and point to the need for more robust strategies to improve rare class recognition [33]. Consistent with the confusion matrix analysis, misclassification remains prevalent among activities with similar motion patterns, such as walking, standing, and shuffling. This recurring pattern across independently developed architectures suggests that the limitation is fundamentally linked to intrinsic similarity in sensor signals and the granularity of activity annotations.

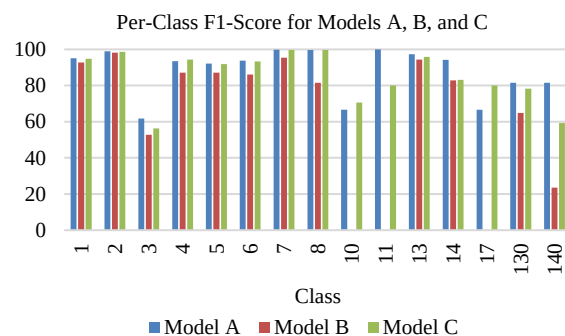


Figure 10. Per-Class F1-score for Models A, B, and C.

3.3 Limitations and Future Work

This study evaluates three hybrid deep learning architectures for activity recognition under realistic and highly imbalanced free-living conditions without applying synthetic data augmentation. While this setup provides a faithful representation of real-world scenarios, the window-based formulation inevitably mixes transitional behaviors, which can obscure class boundaries and contribute to systematic misclassification patterns observed across all models.

Future work may focus on refining model architectures and tuning hyperparameters to improve generalization. However, the findings suggest that architectural scaling alone may not fully resolve ambiguities between closely related motion patterns. Data augmentation techniques could be explored to enhance minority class recognition, provided that generated samples preserve realistic temporal dynamics. In addition, transfer learning from larger or related datasets may further improve performance for underrepresented activity classes. Incorporating boundary-aware learning or sequence-level context modeling may also help capture rapid transitions and reduce ambiguity in activity

recognition, particularly in transitional and underrepresented activities.

4. Conclusions

This study evaluates three hybrid deep learning architectures for human activity recognition in free-living and highly imbalanced conditions using the HARTH v2.0 dataset. The results reveal three key findings. First, hybrid architectures that combine convolutional, recurrent, and attention mechanisms achieve strong overall performance, with Model A attaining the highest accuracy (96.21%) and macro-averaged F1-score (88.17%), followed by Model C (95.58%, 85.05%), while Model B shows substantially lower macro-level performance despite comparable complexity.

Second, soft labeling consistently improves performance across all architectures, yielding modest gains in accuracy (approximately 0.3 percentage points) and more pronounced improvements in macro-averaged metrics, indicating enhanced robustness for minority classes and transitional activities. This confirms its effectiveness in representing uncertainty within windowed sensor data.

Third, all models exhibit persistent confusion among activities with similar motion patterns, particularly walking, standing, and shuffling. This behavior, observed consistently across architectures, suggests that limitations arise not only from model design but also from inherent similarities in sensor signals and the ambiguity of transitional movements. The instability observed in rare-class recognition further highlights the challenges of learning reliable decision boundaries under severe class imbalance.

In addition, a trade-off between recognition performance and computational complexity is evident. While Model A achieves the best performance, it requires substantially more parameters (286,903) than Model C (195,127), which provides competitive results with lower computational demand. This trade-off is particularly relevant for deployment in resource-constrained environments.

Overall, this study demonstrates that standardized evaluation of hybrid architectures not only identifies performance differences but also reveals systematic limitations shared across models. The findings highlight the strengths of temporal-spatial attention-based designs while indicating that architectural scaling alone is insufficient to resolve ambiguities inherent in free-living activity recognition. Future progress may therefore depend on improved modeling of transition boundaries, incorporation of richer temporal context, and robust strategies for handling class imbalance. Ultimately, advancing free-living HAR requires combining architectural innovation with a deeper understanding of uncertainty, context, and the intrinsic limits of sensor-based observation.

Acknowledgements

The authors declare no conflict of interest. This research was supported by a research grant from Calvin Institute of Technology [LPPM-R-DR-2022-07-02-007-2].

References

- [1] A. O. Ige and M. H. M. Noor, "A Survey on Unsupervised Learning for Wearable Sensor-Based Activity Recognition," *Applied Soft Computing*, vol. 127, p. 109363, 2022, doi: [10.1016/j.asoc.2022.109363](https://doi.org/10.1016/j.asoc.2022.109363).
- [2] M. A. A. Al-qaness, A. M. Helmi, A. Dahou, and M. A. Elaziz, "The Applications of Metaheuristics for Human Activity Recognition and Fall Detection Using Wearable Sensors: A Comprehensive Analysis," *Biosensors*, vol. 12, no. 10, p. 821, 2022, doi: [10.3390/bios12100821](https://doi.org/10.3390/bios12100821).
- [3] M. S. Irwanto, F. A. Bachtiar, and N. Yudistira, "Klasifikasi Aktivitas Manusia Menggunakan Algoritme Computed Input Weight Extreme Learning Machine dengan Reduksi Dimensi Principal Component Analysis," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK)*, vol. 9, no. 6, pp. 1195–1202, 2022, doi: [10.25126/jtik.2022965504](https://doi.org/10.25126/jtik.2022965504).
- [4] M. Karim, S. Khalid, A. Aleryani, J. Khan, I. Ullah, and Z. Ali, "Human Action Recognition Systems: A Review of the Trends and State-of-the-Art," *IEEE Access*, vol. 12, pp. 36372–36390, 2024, doi: [10.1109/ACCESS.2024.3373199](https://doi.org/10.1109/ACCESS.2024.3373199).
- [5] M. H. Abdullah and M. A. Ahmed, "Human Activity Recognition using Accelerometer & Gyroscope Smartphone Sensor by Extract Statistical Features," *Journal of Robotics and Control (JRC)*, vol. 5, no. 5, pp. 1390–1398, 2024, doi: [10.18196/jrc.v5i5.22443](https://doi.org/10.18196/jrc.v5i5.22443).
- [6] S. Zhang, Y. Li, S. Zhang, F. Shahabi, S. Xia, Y. Deng, and N. Alshurafa, "Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances," *Sensors*, vol. 22, no. 4, p. 1476, 2022, doi: [10.3390/s22041476](https://doi.org/10.3390/s22041476).
- [7] M. El Ghazi and N. Akin, "A Systematic Review of Non-Intrusive Human Activity Recognition in Smart Homes Using Deep Learning," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 3, pp. 3188–3202, 2024, doi: [10.11591/ijai.v13.i3.pp3188-3202](https://doi.org/10.11591/ijai.v13.i3.pp3188-3202).
- [8] M. Zaher, A. S. Ghoneim, L. Abdelhamid, and A. Atia, "Unlocking the Potential of RNN and CNN Models for Accurate Rehabilitation Exercise Classification on Multi-Datasets," *Multimedia Tools and Applications*, vol. 84, no. 3, pp. 1261–1301, 2024, doi: [10.1007/s11042-024-19092-0](https://doi.org/10.1007/s11042-024-19092-0).
- [9] I. D. Luptáková, M. Kubovčík, and J. Pospíchal, "Wearable Sensor-Based Human Activity Recognition with Transformer Model," *Sensors*, vol. 22, no. 5, p. 1911, 2022, doi: [10.3390/s22051911](https://doi.org/10.3390/s22051911).
- [10] K. K. Sahoo, R. Ghosh, S. Mallik, A. Roy, P. K. Singh, and Z. Zhao, "Wrapper-Based Deep Feature Optimization for Activity Recognition in the Wearable Sensor Networks of Healthcare Systems," *Scientific Reports*, vol. 13, no. 1, p. 965, 2023, doi: [10.1038/s41598-022-27192-w](https://doi.org/10.1038/s41598-022-27192-w).
- [11] A. Logacjov, S. Herland, A. Ustad, and K. Bach, "SelfPAB: Large-Scale Pre-Training on Accelerometer Data for Human Activity Recognition," *Applied Intelligence*, vol. 54, no. 6, pp. 4545–4563, 2024, doi: [10.1007/s10489-024-05322-3](https://doi.org/10.1007/s10489-024-05322-3).
- [12] R. R. Chowdhury, R. Kapila, A. Panse, X. Zhang, D. Teng, R. Kulkarni, D. Hong, R. K. Gupta, and J. Shang, "ZeroHAR: Sensor Context Augments Zero-Shot Wearable Action Recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 15, pp. 16046–16054, 2025, doi: [10.1609/aaai.v39i15.33762](https://doi.org/10.1609/aaai.v39i15.33762).
- [13] F. J. Ordóñez and D. Roggen, "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016, doi: [10.3390/s16010115](https://doi.org/10.3390/s16010115).
- [14] M. H. M. Noor, S. Y. Tan, and M. N. Ab Wahab, "Deep Temporal Conv-LSTM for Activity Recognition," *Neural Processing Letters*, vol. 54, no. 5, pp. 4027–4049, 2022, doi: [10.1007/s11063-022-10799-5](https://doi.org/10.1007/s11063-022-10799-5).
- [15] S. P. Singh, A. Lay-Ekuakille, D. Gangwar, M. K. Sharma, and S. Gupta, "Deep ConvLSTM with Self-Attention for Human

- Activity Decoding Using Wearables,” 2020, doi: [10.48550/ARXIV.2005.00698](https://doi.org/10.48550/ARXIV.2005.00698).
- [16] Y. Li and L. Wang, “Human Activity Recognition Based on Residual Network and BiLSTM,” *Sensors*, vol. 22, no. 2, p. 635, 2022, doi: [10.3390/s22020635](https://doi.org/10.3390/s22020635).
- [17] J. Zhang, Y. Liu, and H. Yuan, “Attention-Based Residual BiLSTM Networks for Human Activity Recognition,” *IEEE Access*, vol. 11, pp. 94173–94187, 2023, doi: [10.1109/ACCESS.2023.3310269](https://doi.org/10.1109/ACCESS.2023.3310269).
- [18] M. Akter, S. Ansary, M. A.-M. Khan, and D. Kim, “Human Activity Recognition Using Attention-Mechanism-Based Deep Learning Feature Combination,” *Sensors*, vol. 23, no. 12, p. 5715, 2023, doi: [10.3390/s23125715](https://doi.org/10.3390/s23125715).
- [19] S. AlMuhaideb, L. AlAbdulkarim, D. M. AlShahrani, H. AlDhubaib, and D. E. AlSadoun, “Achieving More with Less: A Lightweight Deep Learning Solution for Advanced Human Activity Recognition (HAR),” *Sensors*, vol. 24, no. 16, p. 5436, 2024, doi: [10.3390/s24165436](https://doi.org/10.3390/s24165436).
- [20] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, “Activity Recognition Using Cell Phone Accelerometers,” *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011, doi: [10.1145/1964897.1964918](https://doi.org/10.1145/1964897.1964918).
- [21] N. Sikder and A.-A. Nahid, “KU-HAR: An Open Dataset for Heterogeneous Human Activity Recognition,” *Pattern Recognition Letters*, vol. 146, pp. 46–54, 2021, doi: [10.1016/j.patrec.2021.02.024](https://doi.org/10.1016/j.patrec.2021.02.024).
- [22] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, “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)*, 2013, pp. 437–442.
- [23] A. Logacjov, K. Bach, A. Kongsvald, H. B. Bårdstu, and P. J. Mork, “HARTH: A Human Activity Recognition Dataset for Machine Learning,” *Sensors*, vol. 21, no. 23, p. 7853, 2021, doi: [10.3390/s21237853](https://doi.org/10.3390/s21237853).
- [24] T. Gong, J. Jeong, T. Kim, Y. Kim, J. Shin, and S.-J. Lee, “NOTE: Robust Continual Test-Time Adaptation Against Temporal Correlation,” in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2023.
- [25] N. Bento, J. Rebelo, M. Barandas, A. V. Carreiro, A. Campagner, F. Cabrita, and H. Gamboa, “Comparing Handcrafted Features and Deep Neural Representations for Domain Generalization in Human Activity Recognition,” *Sensors*, vol. 22, no. 19, p. 7324, 2022, doi: [10.3390/s22197324](https://doi.org/10.3390/s22197324).
- [26] A. Ustad, A. Logacjov, S. Ø. Trollebø, P. Thingstad, B. Vereijken, K. Bach, and N. S., Maroni, “Validation of an Activity Type Recognition Model Classifying Daily Physical Behavior in Older Adults: The HAR70+ Model,” *Sensors*, vol. 23, no. 5, p. 2368, 2023, doi: [10.3390/s23052368](https://doi.org/10.3390/s23052368).
- [27] A. Logacjov and K. Bach, “Self-Supervised Learning with Randomized Cross-Sensor Masked Reconstruction for Human Activity Recognition,” *Engineering Applications of Artificial Intelligence*, vol. 128, p. 107478, 2024, doi: [10.1016/j.engappai.2023.107478](https://doi.org/10.1016/j.engappai.2023.107478).
- [28] D. Jin, Y. Hu, B. Chen, G. He, J. Chen, and Z. Shen, “TIAN: A Time Series Imaging Association Network for Human Abnormal Behavior Detection,” *Information Fusion*, vol. 118, p. 102906, 2025, doi: [10.1016/j.inffus.2024.102906](https://doi.org/10.1016/j.inffus.2024.102906).
- [29] F. R. Albogamy, “Federated Learning for IoMT-Enhanced Human Activity Recognition with Hybrid LSTM-GRU Networks,” *Sensors*, vol. 25, no. 3, p. 907, 2025, doi: [10.3390/s25030907](https://doi.org/10.3390/s25030907).
- [30] A. Mulyani, S. Khoerunisa, and D. Kurniadi, “Comparison of KNN and SVM Algorithms Performance using SMOTE to Classify Diabetes,” *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 14, no. 1, pp. 25–34, 2025, doi: [10.22146/jnteti.v14i1.15198](https://doi.org/10.22146/jnteti.v14i1.15198).
- [31] R. Haddadj, A. L. Nordstoga, T. I. L. Nilsen, E. S. Skarpsno, A. Kongsvald, M. Flaaten, J. Schipperijn, K. Bach, and P. J. Mork, “Volume and Intensity of Walking and Risk of Chronic Low Back Pain,” *JAMA Network Open*, vol. 8, no. 6, p. e2515592, 2025, doi: [10.1001/jamanetworkopen.2025.15592](https://doi.org/10.1001/jamanetworkopen.2025.15592).
- [32] K. Ellis, J. Kerr, S. Godbole, and G. Lanckriet, “Multi-Sensor Physical Activity Recognition in Free-Living,” in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, 2014, pp. 431–440, doi: [10.1145/2638728.2641673](https://doi.org/10.1145/2638728.2641673).
- [33] S. Hurtado, J. García-Nieto, A. Popov, and I. Navas Delgado, “Human Activity Recognition from Sensorised Patient’s Data in Healthcare: A Streaming Deep Learning-Based Approach,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 1, pp. 23–37, 2023, doi: [10.9781/ijimai.2022.05.004](https://doi.org/10.9781/ijimai.2022.05.004).