Proceedings of the 4th International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2026.36966

# *Hierarchical Early-Exit Inference for Open-World Human Activity Recognition on Microcontrollers*

**Jiahao Zhang**

*Headquarters (Beijing), China Ship Research Academy, Beijing, China*
*rara481846778@gmail.com*

***Abstract.*** Human activity recognition on microcontrollers requires reliable classification under strict computation, memory, latency, and power constraints, while real deployment may also contain activities that are absent from model training. This study develops a hierarchical early-exit framework for open-world human activity recognition using six-axis wrist inertial signals. A three-stage depthwise-separable temporal network combines fine-activity classification, hierarchical supervision, class-prototype modeling, energy-based unknown detection, and deep-to-shallow knowledge distillation. Intermediate predictions are accepted only when class confidence, hierarchy consistency, energy, and prototype distance jointly satisfy calibrated criteria. The network is quantized to signed INT8 and deployed on an Arduino Nano 33 BLE Sense Rev2 using TensorFlow Lite for Microcontrollers and CMSIS-NN. Subject-independent evaluation produced a macro-F1 of 0.918, an unknown-class AUROC of 0.938, and an open-set classification rate of 0.884. About 78% of known windows terminated before the final temporal block, reducing average computation by approximately 52%. Median embedded latency decreased to 8.34 ms and measured energy consumption reached 0.31 mJ per window. These results indicate that hierarchy-aware early exiting can jointly support adaptive computation and unknown-activity rejection on resource-constrained wearable platforms.

***Keywords:*** human activity recognition, early-exit inference, open-world recognition, microcontrollers, edge intelligence

## 1. Introduction

Wearable human activity recognition translates inertial signals into estimates of posture, locomotion, exercise, and daily behavior. Executing these models locally on microcontrollers can reduce communication delay and dependence on external computing resources, but embedded processors impose strict limits on memory, arithmetic operations, latency, and energy. Tiny wearable HAR studies have demonstrated the feasibility of neural inference on highly constrained hardware while highlighting substantial trade-offs between recognition performance and computational resources [1]. Deep-learning reviews further show that wearable HAR increasingly depends on learned temporal representations rather than manually engineered features [2]. Recent self-supervised representation learning has improved the ability to extract discriminative structure from wearable sensor sequences [3].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.36966

Model compression alone cannot solve the inefficiency of processing every activity window through the same network depth. Knowledge distillation can transfer information from a large HAR model into a substantially smaller student network [4]. Early-exit architectures go further by allowing easy samples to terminate at shallow internal classifiers, and HAR-specific experiments have demonstrated favorable accuracy-speed trade-offs across several inertial benchmarks [5]. General dynamic-inference research likewise shows that intermediate representations can support selective termination instead of fixed-depth processing [6]. For TinyML platforms, uncertainty-aware early-exit systems have demonstrated that adaptive inference can reduce monitoring overhead on microcontrollers [7].

A separate problem appears when users perform activities not represented during training. Open-set recognition research has shown that high closed-set confidence does not guarantee reliable rejection of semantically unfamiliar samples [8]. Virtual outlier synthesis provides an additional mechanism for shaping boundaries around known feature distributions [9]. Recent inertial HAR research also demonstrates that lightweight post-hoc open-set scores can identify previously unseen activities without computationally expensive generative inference [10].

This study integrates these requirements within one embedded framework. Hierarchical labels, prototype familiarity, energy scoring, early exits, distillation, and integer quantization jointly determine whether a window should exit early, continue deeper, or be rejected as unknown.

## 2. Literature review

Wearable HAR increasingly adopts compact convolutional and temporal neural networks because they extract activity-specific patterns directly from inertial sequences while remaining suitable for local deployment [1, 2]. Representation quality nevertheless remains important when subject-independent signals vary across users and activities. Self-supervised approaches can improve the structure of wearable embeddings [3], while knowledge distillation reduces model complexity by transferring prediction information from a high-capacity teacher into a smaller network [4].

Adaptive-depth inference addresses a different source of inefficiency. Early-exit HAR allows easily separable windows to terminate at intermediate layers rather than executing every temporal block [5]. Class-aware early-exit methods further demonstrate that intermediate representations can support dynamic decisions while reducing unnecessary operations [6]. TinyML uncertainty research indicates that multi-exit architectures can also provide efficient confidence information under tight memory and latency constraints [7].

Open-world deployment requires additional safeguards. A classifier trained only on known activities may confidently assign unknown behavior to one of its familiar labels. Strong closed-set representation is useful but does not remove the need for explicit novelty handling [8]. Synthetic outliers can regularize uncertainty boundaries when real unknown examples are unavailable during training [9]. Lightweight inertial open-set methods have recently shown that unseen activities can be identified with limited additional computational cost [10].

The remaining challenge is to combine these ideas without making embedded inference more expensive than fixed-depth recognition. The present framework therefore treats early-exit acceptance as a structured decision. A shallow prediction must be confident, hierarchically coherent, and sufficiently close to a known class distribution. Otherwise, additional temporal processing is performed before a final known or unknown decision.

DOI: 10.54254/2753-8818/2026.36966

# 3. Method

## 3.1. Dataset processing and open-world split

The experiment uses wrist accelerometer and gyroscope channels from PAMAP2. Six-axis signals are resampled to 50 Hz, clipped at the 0.5th and 99.5th percentiles, standardized using training-subject statistics, and divided into 2.56-second windows with 50% overlap. Subjects 1–6 form the training set, subject 7 is used for validation and threshold calibration, and subjects 8–9 are reserved for subject-independent testing. Eight activities are treated as known: lying, sitting, standing, walking, Nordic walking, running, cycling, and ascending stairs. Descending stairs, vacuum cleaning, ironing, and rope jumping are excluded from supervised training and introduced only as unknown classes during final testing. Validation pseudo-unknowns are generated by manifold mixup between embeddings from different known activities. Mixup coefficients follow a Beta distribution with both shape parameters set to 0.4. This provides unfamiliar intermediate representations for calibration without exposing the network to the actual test unknown classes.

## 3.2. Hierarchical multi-exit network

The activity taxonomy contains three levels. Motion state separates static and dynamic behavior. Activity family distinguishes posture, ambulatory movement, and cyclic exercise. Fine classification predicts the eight known activities. The backbone begins with a 24-channel one-dimensional convolution followed by three depthwise-separable temporal blocks containing 32, 48, and 64 channels. Each block uses kernel size 5, batch normalization, ReLU6 activation, and stride-2 max pooling. An exit head is attached after every block and consists of global average pooling, a 64-unit dense bottleneck, fine and hierarchical classifiers, and a normalized 32-dimensional embedding. Depthwise-separable operations reduce multiply-accumulate requirements, while progressively deeper blocks expand temporal context for ambiguous movements.

## 3.3. Hierarchical early-exit and unknown rejection

For each known activity and exit depth, a prototype is calculated from normalized embeddings of correctly classified training windows. Each exit produces fine-class probabilities, motion-state probabilities, family probabilities, an energy score, and distance from the predicted class prototype.

An intermediate exit is accepted only when all reliability conditions are satisfied as shown in Equation (1):

$$\mathscr{E}_j(x) = I[p_j^*(x) \geq \tau_j] I[h_j^m(x) = H^m(\hat{y}_j)] I[h_j^f(x) \tag{1}$$

$$= H^f(\hat{y}_j)] I[E_j(x) \leq \epsilon_j] I[d(z_j(x), \mu_{\hat{y}_j,j})| \leq r_{\hat{y}_j,j}$$

Here, $p_j^*(x)$ represents maximum fine-class confidence, $h_j^m$ and $h_j^f$ are predicted hierarchical labels, $H^m$ and $H^f$ map the fine class to its expected hierarchy, $E_j$ is the energy score, and d measures embedding distance from prototype $\mu$. Thresholds are calibrated to retain 95% of correctly classified known validation windows. Windows failing an intermediate rule continue to deeper blocks. At the final exit, excessive energy or prototype distance produces an unknown label.

DOI: 10.54254/2753-8818/2026.36966

### 3.4. Training, quantization, and embedded execution

The training objective jointly optimizes fine-activity classification, hierarchical motion-state and activity-family supervision, prototype compactness, energy-based unknown separation, and deep-to-shallow knowledge distillation. The final exit acts as the teacher for earlier exits, while prototype and energy constraints improve the separation between known and unfamiliar activities. This joint optimization enables shallow exits to retain reliable discrimination while supporting open-world rejection under limited microcontroller resources.

AdamW uses an initial learning rate of 0.001 with cosine decay, batch size 128, and 80 training epochs. Quantization-aware training continues for another 20 epochs before signed INT8 export. TensorFlow Lite for Microcontrollers and CMSIS-NN execute the final network on an Arduino Nano 33 BLE Sense Rev2. All exits share one tensor arena. Latency is measured over 2,000 test windows following 200 warm-up runs, and energy is measured at 3.3 V with a Nordic Power Profiler Kit II. As shown in Figure 1, the proposed architecture combines hierarchical activity modeling, multi-stage temporal inference, joint early-exit reliability checks, open-world rejection, and INT8 microcontroller deployment within a unified HAR framework.

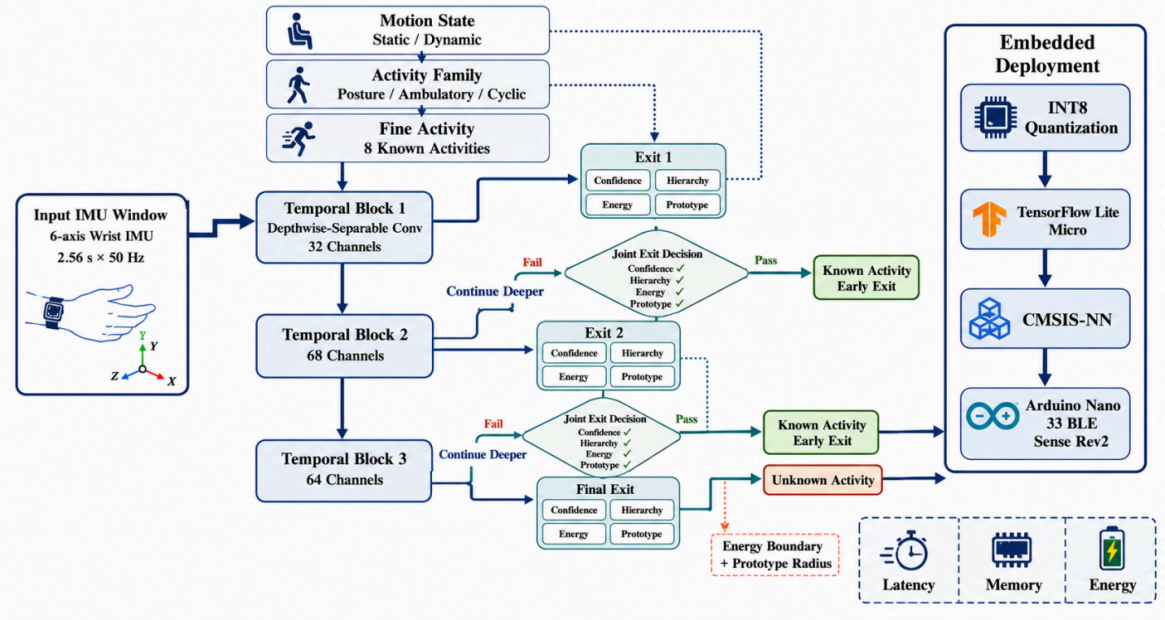



Figure 1. Hierarchical early-exit and open-world recognition architecture for microcontroller-based HAR

## 4. Experimental design

### 4.1. Comparative models and metrics

The proposed architecture is compared with a matched full-depth CNN, a confidence-only early-exit model, and a full-depth open-set CNN under identical data splits and backbone widths. Experiments use random seeds 17, 42, and 89. Known-class evaluation includes macro-F1 and balanced accuracy. Open-world evaluation includes unknown AUROC, FPR at 95% true-positive rate, open-set classification rate, and expected calibration error. Embedded evaluation measures flash size, peak RAM, average operations, exit depth, median latency, 95th-percentile latency, and energy per window. Confidence intervals are obtained from 1,000 subject-stratified bootstrap samples.

DOI: 10.54254/2753-8818/2026.36966

### 4.2. Ablation and robustness tests

Ablation removes hierarchical supervision, prototype distance, energy-margin training, distillation, and quantization-aware training individually. Sensor robustness is evaluated by adding accelerometer noise at standard deviations of 0.02 g and 0.05 g and gyroscope noise at 1 and 3 degrees per second. Additional tests mask 5% and 10% of temporal samples before linear interpolation. All variants use the same unseen subjects and held-out unknown activities.

## 5. Results

### 5.1. Known-activity recognition and unknown rejection

The proposed network achieved a macro-F1 of 0.918 and balanced accuracy of 0.921. Unknown-class AUROC reached 0.938, while FPR95 declined to 0.171. The open-set classification rate was 0.884 and calibration error was 0.028. The full-depth open-set baseline produced similar known-class recognition but required every sample to execute the complete backbone Isee Table 1).

Table 1. Recognition and open-world performance

| Model | Macro-F1 | Balanced Accuracy | Unknown AUROC | FPR95 | OSCR | ECE |
|---|---|---|---|---|---|---|
| Full-depth CNN | 0.924 ± 0.005 | 0.926 ± 0.005 | 0.651 ± 0.021 | 0.493 ± 0.028 | 0.711 ± 0.018 | 0.081 ± 0.009 |
| Confidence-only early exit | 0.910 ± 0.007 | 0.914 ± 0.006 | 0.704 ± 0.019 | 0.421 ± 0.025 | 0.748 ± 0.016 | 0.069 ± 0.007 |
| Full-depth open-set CNN | 0.922 ± 0.006 | 0.924 ± 0.005 | 0.931 ± 0.010 | 0.184 ± 0.016 | 0.881 ± 0.011 | 0.031 ± 0.005 |
| Proposed model | 0.918 ± 0.006 | 0.921 ± 0.005 | 0.938 ± 0.009* | 0.171 ± 0.014* | 0.884 ± 0.010* | 0.028 ± 0.004* |

*Note: **p < .001 relative to confidence-only early exiting.*

### 5.2. Early-exit allocation and component ablation

Approximately 46.8% of known windows terminated at the first exit and another 31.6% terminated at the second, leaving only 21.6% for full-depth processing. Removing hierarchy supervision reduced the open-set classification rate to 0.852. Removing prototype-distance rejection increased FPR95 to 0.219, while removing energy-margin training increased it to 0.241. Without distillation, shallow-exit macro-F1 declined by about 2.1 percentage points and almost 30% of windows required full-depth inference. Under the strongest combined sensor-noise condition, macro-F1 remained 0.873 and unknown AUROC remained 0.901.

### 5.3. Microcontroller latency, memory, and energy

The adaptive model reduced average computation from approximately 1.64 million operations to 0.79 million. Median inference latency fell from 15.92 ms for full-depth execution to 8.34 ms, while measured energy decreased from 0.56 mJ to 0.31 mJ per window. The additional hierarchy and

DOI: 10.54254/2753-8818/2026.36966

prototype constants increased flash and RAM slightly, but the deployment remained within the target microcontroller budget.

## 6. Conclusion

The proposed framework combines hierarchical prediction, open-world rejection, and adaptive-depth inference in a microcontroller-ready HAR model. Confidence alone is insufficient for early termination because unfamiliar activity windows may still produce strong softmax responses. Requiring hierarchy agreement, acceptable energy, and proximity to learned class prototypes makes intermediate decisions more conservative. Distillation improves shallow-exit discrimination, while INT8 quantization and depthwise-separable blocks control embedded cost. The resulting implementation substantially reduces average computation, latency, and energy while preserving known-activity performance and robustly rejecting unknown activities. Ablation results further show that hierarchy, prototype distance, energy regularization, and distillation contribute complementary functions rather than redundant improvements. The framework therefore provides a practical approach for wearable HAR systems that adapt computation to input difficulty while retaining an explicit unknown decision under microcontroller constraints.

## References

[1] Lattanzi, E., Donati, M., & Freschi, V. (2022). Exploring artificial neural networks efficiency in tiny wearable devices for human activity recognition. *Sensors, 22*(7), 2637.
[2] Zhang, S., Li, Y., Zhang, S., Shahabi, F., Xia, S., Deng, Y., & Alshurafa, N. (2022). Deep learning in human activity recognition with wearable sensors: A review on advances. *Sensors, 22*(4), 1476.
[3] Haresamudram, H., Essa, I., & Plötz, T. (2024). Towards learning discrete representations via self-supervision for wearables-based human activity recognition. *Sensors, 24*(4), 1238.
[4] Gonçalves, P. H. N., Bragança, H., & Souto, E. (2024). Efficient human activity recognition on wearable devices using knowledge distillation techniques. *Electronics, 13*(18), 3612.
[5] Yu, J., Zhang, L., Cheng, D., Huang, W., Wu, H., & Song, A. (2024). Improving human activity recognition with wearable sensors through BEE: Leveraging early exit and gradient boosting. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 32*, 3452–3464.
[6] Wang, J., Li, B., & Zhang, G. L. (2024). Early-exit with class exclusion for efficient inference of neural networks. *2024 IEEE 6th International Conference on AI Circuits and Systems*, 263–267.
[7] Ghanathe, N. P., & Wilton, S. J. E. (2025). QUTE: Quantifying uncertainty in TinyML models with early-exit-assisted ensembles for model-monitoring. *Proceedings of the 42nd International Conference on Machine Learning, 267*, 19286–19306.
[8] Vaze, S., Han, K., Vedaldi, A., & Zisserman, A. (2022). Open-set recognition: A good closed-set classifier is all you need? *International Conference on Learning Representations*.
[9] Du, X., Wang, Z., Cai, M., & Li, Y. (2022). VOS: Learning what you don't know by virtual outlier synthesis. *International Conference on Learning Representations*.
[10] Cortese, A., Solbiati, S., Scandelli, A., Giudici, A., Antonello, N., Trojaniello, D., Boracchi, G., & Caiani, E. G. (2026). Open-set recognition of human activities from head-mounted inertial sensor. *Sensors, 26*(3), 1079.