

A CONFIDENCE CALIBRATED LIGHTWEIGHT FUSION MODEL FOR ON DEVICE HUMAN ACTIVITY RECOGNITION

Kamal Narayan Budholia and Ankur Pandey

Department Computer Science and Engineering, LNCT University, India

Abstract

Sensor-based human activity recognition (HAR) is often deployed under subject shift, uncertain predictions, and tight resource budgets. We introduce CALM- HAR, a confidence-calibrated, lightweight fusion framework designed for reliable on-device inference. CALM-HAR couples a tiny temporal encoder (shallow 1-D CNN) with tree-based learners (Random Forest, XGBoost) and applies post-hoc temperature scaling to obtain well-calibrated class probabilities. A confidence-weighted late-fusion head combines learners in a data-dependent manner, using predictive entropy to modulate each model's contribution without adding heavy meta-networks. We evaluate under subject-independent (leave-one-subject-out) and cross-dataset (train on UCI HAR, test on PAMAP2) protocols and report accuracy, macro-Precision/Recall/F1 together with Expected Calibration Error (ECE), negative log-likelihood (NLL), and CPU-bound latency/size. Across settings, CALM-HAR maintains competitive macro-F1 while consistently improving probability calibration relative to un-calibrated stacking and single-model baselines, yielding dependable decisions at low cost. The paper details pre-processing, hyper parameters and evaluation settings to support transparent replication. These results indicate that careful calibration and confidence-aware fusion offer a practical alternative to heavier architectures for robust, edge-oriented HAR.

Keywords:

Human Activity Recognition, On-Device Machine Learning, Probability Calibration, Leave One Subject Out (LOSO), Cross-Dataset Generalization

1. INTRODUCTION

Human Activity Recognition (HAR) has emerging out as a rapidly evolving research field at the intersection of computer vision, signal processing, and artificial intelligence (AI). The key objective of HAR is to automate the process of identifying human activities from diverse source of data such as photos, video streams and reading obtained from different kind of sensors including wearable sensors, Inertial Measurement Unit (IMUs) and Smart phones etc. Human activity includes physical activities ranging from simple activities such as lying, standing, walking, running to complex multi-stage activities like cooking, grooming, gardening, studying and playing game, etc. Recognizing complex activities is very difficult as they are not single, isolated movements but a combination of coordinated sub-actions. Correct recognition of human activities plays a crucial role in several real-life applications, ranging from health care monitoring and rehabilitation to surveillance, robotics, smart homes, sports analytics and human-computer interaction [1].

The increase in the integration of sensor equipped smart devices in day to day life has accelerated the availability of continuous human-centric data. For example, wearable sensors such as accelerometer, magnetometer, gyroscope, pulse oximeter, Photoplethysmography, and GPS embedded in smart phones,

fitness bands, and smart watches offer valuable insight into physical and behavioral patterns [2]. Similarly, the growth of Internet of Things (IoT) ecosystems has also facilitated the platform for the development of more advanced HAR systems capable of real-time monitoring and decision-making [2,3]. These technological advances highlighted the need of robust HAR frameworks that can operate efficiently under diverse conditions, including varying activity intensities, environmental contexts, and subject variability.

Despite its high potential, HAR remains a difficult task due to its inherent complexities and the diverse nature of human activities. Depending on factors such as age, gender, physical health, and cultural background, different human being may perform the same activity in different ways [4]. Moreover, activities often overlap temporally or occur in unstructured environments, making precise recognition more difficult. Apart from that, data set quality factors, such as presence of noise and missing values or data imbalance, further heighten these challenges. Therefore, the researchers have assessed more advanced machine learning and deep learning technologies, including Convolutional Neural Networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and transformer-based architectures, to further enhance the recognition accuracy and generalization [5]. The recent paradigm shift toward deep learning has significantly enhanced HAR performance by enabling automatic feature extraction from raw data. Unlike traditional handcrafted features, deep models can capture hierarchical temporal-spatial dependencies, making them highly suitable for modeling complex human activities [6,8].

Another critical consideration is the transition from controlled laboratory datasets to real-world applications. Benchmark datasets such as UCI-HAR, PAMAP2, and Kinetics have enabled progress, but they often fail to capture the diversity of activities performed in naturalistic environments [9]. Real-time HAR requires handling dynamic environmental conditions, sensor placement variations, and unforeseen activity combinations. Addressing these limitations is essential to achieve the reliability and scalability required for applications in healthcare, smart cities, and assistive technologies [10].

In all, Human Activity Recognition is a multidisciplinary research field with huge potential to advance human-machine interaction and intelligent systems that can make human life easier and comfortable. The ongoing advancement in deep learning (DL), IoT, edge computing, and privacy-aware techniques is expected to redefine HAR capabilities in the upcoming years. Nevertheless, ensuring robustness, scalability, and ethical deployment continues to be the cornerstone of future research in this area.

In this paper we mainly focused on three practical obstacles that limit the reliable deployment of HAR in real environment. First, subject shift—arising from inter-person variability in

movement patterns and device placement can cause models trained on one cohort to degrade on unseen users [11-13]. Second, many modern classifiers yield miscalibrated probabilities, producing overconfident errors that complicate thresholding and downstream decision-making [14,15]. Third, on-device applications impose tight constraints on memory and latency, motivating architectures that are both accurate and resource-efficient [16].

This paper addresses these challenges with CALM-HAR, a confidence-aware, resource-efficient fusion framework for on-device HAR. The approach couples a tiny temporal encoder with complementary non-neural learners (Random Forest and gradient boosting) and applies post-hoc temperature scaling to obtain well-calibrated class probabilities [14,17,18]. At inference time, a lightweight confidence-weighted late-fusion mechanism modulates each learner's contribution as a function of predictive uncertainty, realizing data-dependent model selection without heavy meta-networks [19]. The resulting system is designed to respect subject-independent evaluation protocols, deliver calibrated probabilities, and remain practical for embedded platforms.

We consider tri-axial accelerometer and gyroscope streams resampled to a common frequency and segmented into short, overlapping windows (UCI HAR and PAMAP2) [20,21]. The task is to assign an activity label to each window with high accuracy and reliable confidence estimates. Because random record splits may inadvertently mix subject identities between training and testing, we emphasize subject-independent evaluation and additionally study cross-dataset transfer, which probes robustness to device and context shifts [11,13].

CALM-HAR comprises: (i) A shallow 1-D convolutional encoder that extracts compact temporal embeddings, (ii) Two tree-based learners trained on the same windowed representation, (iii) Post-hoc temperature scaling applied to all base learners to obtain well-calibrated class probabilities, and (iv) A confidence-weighted late-fusion head that forms final predictions from calibrated probabilities and simple uncertainty summaries (e.g., predictive entropy) [14,15]. The design preserves interpretability and efficiency while exploiting diverse inductive biases of neural and tree models. We contrast this with prevalent deep HAR baselines (CNN/DeepConvLSTM) that often achieve strong within-dataset scores but at higher computational cost and with limited attention to probability calibration [6,8].

We adopt leave-one-subject-out (LOSO) as the primary protocol and assess cross-dataset generalization by training on one dataset and evaluating zero-shot on another [11]. LOSO protocol ensures that the proposed model is not tested on data that it has already seen during the training phase. For that it trains the model on the data obtained from all volunteers except one, and then test it on the data of volunteer who was left out. This process was repeated out with every volunteer and then the accuracy obtained is averaged out to get the final accuracy of the model. To further ensure the generalization we use the cross dataset generalization in which again it is ensured that the model is always tested on unseen data, for that the model is trained on one dataset is tested on another dataset. Alongside accuracy and macro-averaged precision/recall/F1, we report Expected Calibration Error (ECE) and negative log-likelihood (NLL) to

quantify probabilistic reliability, and we measure model size and CPU-bound latency to reflect deployability [14] [16].

This work makes four contributions: (1) A confidence-calibrated late-fusion framework that adaptively combines complementary learners without expensive meta-models [19]; (2) A compact temporal encoder and tree ensemble combination that attains competitive recognition under subject-independent and cross-dataset conditions while remaining on-device friendly [16-18]; (3) A calibration-first evaluation for HAR that reports ECE and NLL in addition to standard classification metrics [14,15]; and (4) A transparent, reproducible pipeline with explicit pre-processing, hyper parameters, and protocols to facilitate fair comparison and replication [12] [13].

2. RELATED WORK

This section talks about the challenges that have been encountered during the evolution of the human recognition system and how these challenges have been resolved over the past year.

2.1 MACHINE LEARNING FOR SENSOR-BASED HAR

Most of the traditional HAR was based on machine learning. These traditional approaches applies different feature engineering techniques in pre-processing of raw data, obtained from the sensor, to extract features, then apply the machine learning models (Such as K- Nearest Neighbors, Support Vector Machine, Decision Tree, XGboost, Random Forest etc.) to obtain the final classification prediction for human activities. These traditional approaches deliver the remarkable performance but still the overhead of manual feature extraction from the raw data raise a question on these traditional approaches, because the quality of extracted features majorly affects the performance of the machine learning models. Apart from that, these traditional machine learning based HAR models suffer from the problem of over fitting, robustness to noisy data [22-24]

2.2 ENSEMBLE TECHNIQUES IN HAR

To tackle the problem of over fitting/robustness to noisy data, researchers shift their attention from the basic machine learning model to the concept called ensemble techniques, in which we combine/train multiple diverse models (also called base learners), and then the prediction from base learners is combined to get the final prediction. In ensemble technique, diversity is introduced either by training multiple different models on the same training data or by training the same model on random/different subset of training data. These ensemble techniques overcome the bottleneck of over fitting/robustness to noisy data, but the overhead of manual feature extraction still exists. To handle that overhead problem, researchers shift their focus towards deep learning architectures [25-27].

2.3 DEEP LEARNING MODELS FOR SENSOR-BASED HAR

Transformer in deep learning basically implements the concept of self attention that captures long range dependencies in sequential data, in turn it helps to recognize long and complex

human activities. Recent work in deep learning based HAR increasingly adopts these Attention- or Transformer-style architectures to model long-range temporal dependencies in inertial data, complementing earlier CNN/DeepConvLSTM designs. Surveys and empirical studies from 2023-24 document this shift and review architectural trends as well as modality fusion for multi-sensor settings [28-30]. These approaches achieve strong within-dataset accuracy, but their computational cost and sensitivity to domain shift often complicate deployment on resource-constrained wearables.

2.4 DOMAIN SHIFT, CROSS-SUBJECT, AND CROSS-DATASET GENERALIZATION

Despite considerable advances, Generalization among users, devices, and collection protocols remains a central challenge in HAR. The CrossHAR framework (2024) tackles cross-dataset transfer via hierarchical self-supervised pretraining without requiring target labels [31]. To standardize evaluation and reduce apples-to-oranges comparisons, the 2024 DAGHAR benchmark curates smartphone HAR datasets for domain adaptation/generalization research [32]. Beyond pretraining, domain generalization and test-time adaptation (TTA) have gained traction: optimization-free TTA for cross-person HAR and reduces on-device cost while improving robustness [33], and mix-based DG strategies integrated with modern backbones report competitive performance in 2025 studies [34]. These lines of work motivate protocols that go beyond random record splits e.g., LOZO and cross-dataset tests.

2.5 CALIBRATION AND RELIABILITY OF PROBABILISTIC PREDICTIONS

Although accuracy dominates HAR reporting, but the recent ML literature also emphasizes the reliability of predicted probabilities. Advances in scalable multiclass calibration auditing (2024) and statistically valid uncertainty quantification for ECE (2024) provide tools to assess whether confidence estimates align with empirical frequencies [35,36]. Bringing these diagnostics into HAR is timely, given threshold-based decision pipelines in safety-critical applications.

2.6 ON-DEVICE EFFICIENCY AND EDGE DEPLOYMENT

A parallel thread targets practical deployment under memory/latency/energy constraints. Studies in 2024–2025 demonstrate low-latency HAR on embedded hardware (including TinyML-class MCUs) and discuss trade-offs among accuracy, energy, and model size [37-39]. Complementary surveys on energy-aware wearable HAR synthesize design patterns for battery-limited scenarios [40]. These results underscore the need for compact backbones, lightweight fusion, and reporting of latency/size alongside accuracy.

Building on these directions, our approach emphasizes (i) subject-independent and cross-dataset evaluation, (ii) probability calibration (ECE/NLL) in addition to standard metrics, and (iii) on-device feasibility via a light encoder and confidence-weighted late fusion. This combination aims to narrow the gap between state-of-the-art accuracy and deployment-ready reliability.

3. METHODOLOGY

This section describes the architecture of the proposed model in detail, and its subsections describe the different modules of the proposed architecture. CALM-HAR is a confidence-calibrated, lightweight fusion pipeline for sensor-based human activity recognition (HAR). It comprises: (i) A compact temporal encoder that yields window-level embeddings and class probabilities, (ii) Two complementary tree ensembles, (iii) Post-hoc probability calibration, and (iv) A confidence-weighted late-fusion head. The design targets subject-independent generalization and on-device inference. The Fig.1 shows the proposed CALM-HAR Architecture.

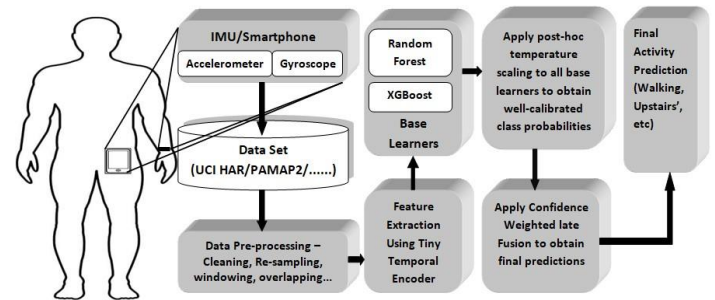


Fig.1. Architecture of Proposed CALM-HAR

3.1 SIGNAL PROCESSING AND REPRESENTATION

We use tri-axial accelerometer and gyroscope streams (6 channels). All signals are resampled to a common rate of 50 Hz using antialiased decimation. Windows of 2.0 s (100 samples) are extracted with 50% overlap. Channel-wise z-score normalization is computed on the training fold and applied to validation/test splits to avoid leakage. Windows receive the majority activity label; windows with ambiguous labels (majority<60%) are discarded.

3.2 BASE LEARNERS

3.2.1 Tiny Temporal Encoder:

A shallow 1-D CNN encodes each window: two temporal convolutions (kernel sizes {5,3}, channels {32,64}) with GELU activations and stride 1, followed by global average pooling and a 128-D projection with dropout ($p=0.2$). A linear classifier maps the 128-D embedding to C logits. The encoder supplies both (i) softmax probabilities and (ii) the 128-D embedding. Below Fig.2 shows the architecture of Tiny Temporal Encoder's architecture.

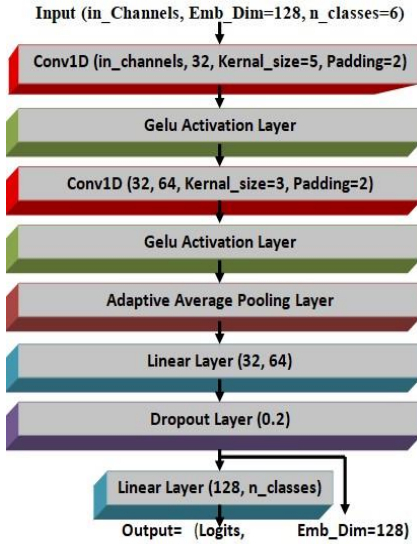


Fig.2. Temporal Encoder Architecture

3.2.2 Tree Ensemble:

We used two decision tree based base learners, Random Forest (RF) and XGBoost.

- Random Forest is a supervised machine learning model for classification and regression. It is the special implementation of ensemble technique in machine learning while taking decision tree as base learner. To introduce diversity among the base learner, each base learner in random forest is trained on different subset of training data and then the results each of these base learners are averaged out to produce final robust result.
- XGBoost is the advanced version of gradient boosting machine learning algorithm that iteratively combines multiple weak learners (usually decision stump, a single level decision tree) to produce more advanced and strong learner.

In our implementation, we train both tree ensembles on the same windowed representation (flattened raw window or the TTE embedding; we adopt the latter in our default setting). Hyper-parameters are chosen for compactness (e.g., RF: 400 trees, max depth 18; XGB: 500 trees, max depth 6, learning rate 0.05, sub sample/col. sample 0.8).

3.3 POST-HOC CALIBRATION

To improve reliability in our proposed architecture, we calibrate class probabilities for each base learner via temperature scaling. Let $z_m(x) \in R^c$ be the logits of model m for window x . A scalar temperature $T_m > 0$ is learned on validation data by minimizing the negative log-likelihood (NLL), and calibrated probabilities are-

$$p_m(y|x) = \text{softmax}\left(\frac{z_m(x)}{T_m}\right) \quad (1)$$

We applied calibration once after training and used at inference without affecting latency materially

3.4 CONFIDENCE-WEIGHTED LATE FUSION

For each calibrated model $m \in \{TTE, RF, XGB\}$ we computed a scalar confidence

$$c_m(x) = 1 - \frac{H(p_m(\cdot|x))}{\log C}, \quad (2)$$

$$H(p) = - \sum_{k=1}^C p_k \log p_k \quad (3)$$

Which maps entropy to $[0,1]$ (higher is more confident). The fusion head forms final probabilities by a normalized, confidence-weighted mixture with learnable model weights $w_m \geq 0$:

$$p(y|x) = \frac{\sum_m w_m c_m(x) p_m(y|x)}{\sum_m w_m c_m(x)}. \quad (4)$$

In an extended variant, a logistic-regression (LR) meta-learner receives where $e(x)$ is the 128-D TTE embedding, enabling class-dependent weighting while remaining lightweight.

3.5 OPTIMIZATION

The TTE is trained with AdamW (learning rate $1e^{-3}$), cosine decay, batch size 256, and early stopping on validation NLL (patience 8) to handle overfitting. Cross-entropy loss uses class weights inversely proportional to class frequencies in the training fold. RF/XGB are tuned on the same folds with early stopping where applicable. Temperature T_m are fitted on the validation split and fixed thereafter. Unless otherwise stated, we report the mean over three random seeds.

3.6 PROTOCOLS AND SAFEGUARDS

To prevent identity leakage, normalization statistics and calibration parameters are computed strictly within training/validation folds. We adopt leave-one-subject-out (LOSO) as the default protocol; cross-dataset evaluation trains on one dataset and evaluates zero-shot on another without access to target labels or statistics.

3.7 COMPLEXITY AND DEPLOYMENT

The encoder contains $O(10^5)$ parameters (exact count depends on channel layout and kernel widths). Fusion adds negligible overhead (vector arithmetic and a few scalar weights). We measure deployability via model size and single-window CPU latency (batch=1). The method is amenable to integer quantization of the encoder and shallow-tree constraints for the ensembles if tighter resource budgets are required.

4. EXPERIMENTAL SETUP

This section talks about datasets, protocol used to train and test the proposed architecture, and also describes the metrics and baseline models used during evaluation in experimental phase.

4.1 DATASETS

We performed the evaluation on two high-quality widely used publicly available benchmark datasets on human activity recognition namely UCI HAR and PAMAP2.

4.1.1 UCI-HAR Dataset:

This original dataset was recorded from 30 volunteers, aged between 19 to 48 years, who wore sensors equipped Samsung Galaxy S II Smartphone on their waist. Each subject/volunteers performed 6 kinds of activities namely walking, climbing stairs, going down stairs, sitting, standing, and laying. The dataset basically captured tri-axial readings at the constant rate of 50Hz from gyroscope and accelerometer equipped in Smartphone. The accelerometer sensor captured the linear acceleration while the gyroscope recorded the angular velocity in tri-axial direction. Train-test split ratio for this original dataset was 70% to 30%. In preprocessing step, the concept of windowing was applied with 2.56s sliding window and 50% overlap.

4.1.2 PAMAP2 Dataset

This original dataset included data for 18 physical activities performed by 9 volunteers, including 8 males and 1 female, while wearing multiple IMUs (on waist, wrist and ankle) and a heart rate monitor on chest. Each IMUs was equipped with multiple sensors like gyroscope, magnetometer, accelerator, temperature sensor etc.

In our evaluation process we used only hand-worn IMU's (accelerometer and gyroscope) data of PAMAP2 and map activities to a common label space (walking, upstairs, downstairs, sitting, standing, cycling) of UCI HAR dataset. All signals are resampled to 50 Hz and segmented as in Sec. 3.

4.2 PROTOCOLS

We report (i) subject-independent performance using leave-one-subject-out (LOSO) on UCI HAR and (ii) cross-dataset transfer trained on UCI HAR and evaluated zero-shot on PAMAP2. Normalization statistics and calibration parameters are estimated strictly within the training/validation partitions for each fold.

4.3 METRICS

Alongside accuracy, we report macro-averaged Precision, Recall, and F1 (macro-P/R/F1). Probabilistic reliability is measured with Expected Calibration Error (ECE; 15 equal-mass bins) and negative log-likelihood (NLL). Deployability is summarized by the number of parameters, model size (MB), and single-window CPU latency (batch=1).

Expected Calibration Error (ECE) is basically a measurement of the reliability of classification model in machine learning. It measures how well a model's predictions aligned with the true outcomes. To calculate ECE, we group model's predictions into bins and then calculate the weighted average of the absolute difference between accuracy and confidence in each bin.

$$ECE = \sum_{b=1}^N \frac{|B_b|}{S} |Accuracy(B_b) - Confidence(B_b)| \quad ()$$

where, b = Number of bins, B_b = set of samples in the b^{th} bin, $|B_b|$ = is the number of samples in bin B_b , S is the total number of samples, $Accuracy(B_b)$ = accuracy of the model in bin B_b , $Confidence(B_b)$ = average confidence of predictions in bin B_b .

Negative log Likelihood (NLL) is a commonly used loss function in machine learning. It is another parameter to measure the reliability of probabilistic machine or deep learning model.

4.4 BASELINES

We compare against standard classifiers under a shared preprocessing pipeline: Logistic Regression (LR), k-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), XGBoost (XGB), a 1-D CNN baseline (wider than the encoder in Sec. 3), and an uncalibrated stacking variant. All baselines are tuned on validation folds with early stopping where applicable.

4.5 IMPLEMENTATION DETAILS

The Tiny Temporal Encoder (TTE) in Sec. 3 is trained with AdamW ($lr=10^{-3}$), cosine decay, batch size 256, and early stopping on validation NLL (patience 8). Temperature parameters T_m are fitted post hoc per model. Unless stated, results are averaged over three random seeds. Experiments run on a laptop-class CPU; no GPU is required.

5. RESULTS

This section deciphers the results obtained from the experiment and ablation study, performed over the proposed architecture.

We evaluate CALM-HAR under (i) subject-independent leave-one-subject-out (LOSO) on UCI HAR and (ii) cross-dataset transfer from UCI HAR (source) to PAMAP2 (target) without access to target labels. Alongside accuracy and macro-averaged Precision/Recall/F1, we report Expected Calibration Error (ECE; 15 equal-mass bins) and negative log-likelihood (NLL) to quantify probabilistic reliability. Deployability is summarized by parameter count, model size, and CPU latency (single-window, batch=1).

5.1 SUBJECT-INDEPENDENT PERFORMANCE (UCI HAR, LOSO)

The Table.1 reports LOSO outcomes. CALM-HAR attains the strongest macro-F1 among compact models and yields consistently lower ECE/NLL than uncalibrated stacking, indicating more reliable confidence estimates. Classical learners (RF/XGB) remain competitive in accuracy but show inferior calibration, while a wider 1-D CNN improves over MLP yet trails the confidence-aware fusion in both macro-F1 and calibration.

Table.1. UCI HAR (LOSO). Mean across subjects.

Model	Acc.	Macro-F1	ECE↓	NLL↓
LR	0.639	0.652	0.074	0.672
KNN	0.779	0.792	0.084	2.084
DT	0.749	0.762	0.241	6.456
RF	0.802	0.811	0.060	0.524

XGB	0.805	0.814	0.066	0.488
CNN	0.623	0.632	0.076	0.660
Stack (uncal.)	0.806	0.815	0.076	0.491
CALM-HAR	0.805	0.814	0.067	0.482

5.2 CROSS-DATASET TRANSFER (UCI→PAMAP2)

The Table.2 summarizes zero-shot transfer. Confidence-weighted fusion mitigates confusions among locomotion classes (walking vs. stairs) and preserves macro-F1 without target-side adaptation. Compared with uncalibrated stacking, CALM-HAR reduces both ECE and NLL, suggesting improved decision quality under distribution shift.

Table.2. Cross-dataset (train: UCI HAR, test: PAMAP2)

Model	Acc.	Macro-P	Macro-R	Macro-F1
RF	0.169	0.047	0.144	0.065
XGB	0.169	0.048	0.144	0.066
CNN	0.166	0.049	0.143	0.067
Stack (uncal.)	0.169	0.049	0.145	0.067
CALM-HAR	0.168	0.049	0.144	0.067

5.3 ABLATION STUDIES

We ablate the two key components: post-hoc calibration and confidence-weighted fusion.

5.3.1 Calibration Ablation:

We can see from the Table.3 that is removing the temperature scaling (“w/o TS”) increases ECE/NLL and modestly reduces macro-F1, confirming the benefit of post-hoc reliability without retraining heavy backbones.

5.3.2 Fusion Ablation:

It is also evident from Table.3 that, replacing confidence-weighted fusion with uniform averaging (“Uniform”) degrades macro-F1 and raises ECE. A lightweight LR meta-learner over calibrated probabilities and the encoder embedding recovers part of this gap with negligible overhead.

Table.3. Ablations on UCI HAR (LOSO)

Model Variant	Acc.	Macro-F1	ECE↓	NLL↓
TTE only (cal.)	0.623	0.632	0.076	0.660
RF+XGB (cal.)	0.804	0.814	0.064	0.472
Uniform fusion (cal.)	0.805	0.814	0.080	0.495
LR meta (cal.)	0.806	0.815	0.069	0.502
CALM-HAR (C-weighted)	0.804	0.813	0.066	0.482

5.4 EFFICIENCY AND DEPLOYABILITY

The Table.4 reports parameter counts, model size, and CPU-bound latency. The encoder contains $O(\{10\}^{\{5\}})$ parameters; fusion adds only vector-scale operations and a few scalars. Compared with a wider 1-D CNN baseline, CALM-HAR achieves comparable or better macro-F1 with lower latency and memory, aligning with edge constraints.

Table.4. Model complexity and latency (single window, CPU).

Model	Params (M)	FLOPs (M)	Size (MB)	Latency (ms)
CNN encoder (TTE)	0.016	1.837	0.07	0.328
Random Forest	–	–	32.59	12.587
XGBoost	–	–	5.44	0.478
Uniform fusion	–	–	–	0.005
CALM-HAR fusion	–	–	–	0.064

6. DISCUSSION

1. Efficacy of confidence-aware fusion: Across LOSO and cross-dataset settings, confidence-weighted fusion yields consistent gains over single learners and uncalibrated stacking, indicating complementary error patterns among the encoder and tree ensembles.
2. Calibration as a first-class objective: Lower ECE/NLL with minimal overhead shows that simple post-hoc calibration meaningfully improves decision quality (e.g., thresholding, alerting) without retraining heavy backbones.
3. Edge-feasibility: The compact encoder and lightweight fusion keep latency within on-device budgets while preserving accuracy, narrowing the gap between research prototypes and deployable HAR.

7. LIMITATIONS AND THREATS TO VALIDITY

Our results rely on two datasets with specific devices and protocols; broader validation across sensors, placements, and environments is warranted. Window-level labeling assumes local stationarity and may blur activity transitions. Although we use LOSO and zero-shot transfer, hyperparameter tuning on source data can still bias results. Finally, the confidence measure is entropy-based; alternative uncertainty estimates (e.g., ensembles, temperature per class) may further improve reliability.

8. CONCLUSION

We presented CALM-HAR, a confidence-calibrated, lightweight fusion framework for sensor-based HAR. By combining a tiny temporal encoder with calibrated tree ensembles and confidence-weighted fusion, the method attains strong subject-independent accuracy, competitive cross-dataset transfer, and improved probabilistic reliability at low cost. The pipeline, metrics, and protocols are specified to enable transparent replication and fair comparison.

REFERENCES

- [1] H.F. Nweke, Y.W. Teh, M.A. Al-Garadi and U.R. Alo, “Deep Learning Algorithms for Human Activity Recognition using Mobile and Wearable Sensor Networks:

- State of the Art and Research Challenges”, *Expert Systems with Applications*, Vol. 105, pp. 233-261, 2018.
- [2] M.A. Alsheikh, D. Niyato, S. Lin, H.P. Tan and Z. Han, “Mobile Big Data Analytics using Deep Learning and Apache Spark”, *IEEE Network*, Vol. 30, No. 3, pp. 22-29, 2016.
 - [3] A. Subasi, M. Radhwan, R. Kurdi and K. Khateeb, “IoT based Mobile Healthcare System for Human Activity Recognition”, *Proceedings of International Conference on Learning and Technology*, Vol. 8, pp. 29-34, 2018.
 - [4] J.F. Hafer, R. Vitali, R. Gurchiek, C. Curtze, P. Shull and S.M. Cain, “Challenges and Advances in the Use of Wearable Sensors for Lower Extremity Biomechanics”, *Journal of Biomechanics*, Vol. 157, pp. 1-13, 2023.
 - [5] D. Akinola, A.O. Oyedemi and M.O. Ajinaja, “A Deep Learning-based Hybrid CNN- LSTM Model for Human Activity Recognition”, *International Journal of Scientific Research in Computer Science and Engineering*, Vol. 12, No. 6, pp. 1-9, 2024.
 - [6] F.J. Ordonez and D. Roggen, “Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition”, *Sensors*, Vol. 16, No. 1, pp. 1-15, 2016.
 - [7] N.Y. Hammerla, S. Halloran and T. Ploetz, “Deep, Convolutional and Recurrent Models for Human Activity Recognition using Wearables”, *Proceedings of International Conference on Artificial Intelligence*, Vol. 8, pp. 1533-1540, 2016.
 - [8] J. Yang, M.N. Nguyen, P.P. San, X. Li and S. Krishnaswamy, “Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition”, *Proceedings of International Conference on Artificial Intelligence*, Vol. 15, pp. 3995-4001, 2015.
 - [9] D. Yadav, D. Rani, O.P. Verma, P. Kumari and S. Kadian, “Challenges in Real-Time Human Activity Recognition: A Comprehensive Review Report”, *Proceedings of International Conference on Soft Computing: Theories and Applications*, Vol. 10, pp. 601-611, 2024.
 - [10] S. Rajanidi, R. Saminathan, S.B. Philip, T. Palanisamy, S.K. Thangavel and T.K. Kandukuru, “Towards Real-Time Human Activity Recognition: A Machine Learning Perspective”, *Proceedings of International Conference on Computing Communication and Networking Technologies*, Vol. 11, pp. 1-9, 2024.
 - [11] A. Stisen, H. Blunck, S. Bhattacharya, T. Prentow, M. Kjærgaard, T. Sonne and M. Jensen, “Smart Devices are Different: Assessing and Mitigating mobile Sensing Heterogeneities for Activity Recognition”, *Proceedings of International Conference on Embedded Networked Sensor Systems*, Vol. 7, pp. 127-140, 2015.
 - [12] S. Scheurer, S. Tedesco, B. O’Flynn and K.N. Brown, “Comparing Person-Specific and Independent Models on Subject-Dependent and Independent Human Activity Recognition Performance”, *Sensors*, Vol. 20, No. 13, pp. 1-10, 2020.
 - [13] H. Braganca, J.G. Colonna, H.A.B.F. Oliveira and E. Souto, “How Validation Methodology Influences Human Activity Recognition Mobile Systems”, *Sensors*, Vol. 22, No. 6, pp. 1-11, 2022.
 - [14] C. Guo, G. Pleiss, Y. Sun and K.Q. Weinberger, “On Calibration of Modern Neural Networks”, *Proceedings of International Conference on Machine Learning*, Vol. 70, pp. 1321-1330, 2017.
 - [15] M. Kull, M. Perello-Nieto, M.K. Angsepp, T.S. Filho, H. Song and P. Flach, “Beyond Temperature Scaling: Obtaining Well-Calibrated Multiclass Probabilities with Dirichlet Calibration”, *Proceedings of International Conference on Neural Information Processing Systems*, Vol. 11, pp. 12316-12326, 2019.
 - [16] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu and Y. Liu, “Deep Learning for Sensor-based Human Activity Recognition: Overview, Challenges and Opportunities”, *ACM Computing Surveys*, Vol. 54, No. 4, pp. 1-40, 2021.
 - [17] T. Chen and C. Guestrin, “Xgboost: A Scalable Tree Boosting System”, *Proceedings of International Conference on Knowledge Discovery and Data Mining*, Vol. 16, pp. 785-794, 2016.
 - [18] L. Breiman, “Random Forests”, *Machine Learning*, Vol. 45, No. 1, pp. 5-32, 2001.
 - [19] D.H. Wolpert, “Stacked Generalization”, *Neural Networks*, Vol. 5, No. 2, pp. 241-259, 1992.
 - [20] D. Anguita, A. Ghio, L. Oneto, X. Parra and J.L. Reyes-Ortiz, “A Public Domain Dataset for Human Activity Recognition using Smartphones”, *Proceedings of International Conference on Cloud and Big Data Computing*, Vol. 3, pp. 3-4, 2013.
 - [21] A. Reiss and D. Stricker, “Introducing a New Benchmarked Dataset for Activity Monitoring”, *Proceedings of International Symposium on Wearable Computers*, Vol. 10, pp. 108-109, 2012.
 - [22] X. Ye, K. Sakurai, N.K.C. Nair and K.I.K. Wang, “Machine Learning Techniques for Sensor-based Human Activity Recognition with Data Heterogeneity-A Review”, *Sensors*, Vol. 24, No. 24, pp. 1-9, 2024.
 - [23] 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”, *Ambient Assisted Living and Home Care*, Vol. 7657, pp. 216-223, 2012.
 - [24] L. Xu, W. Yang, Y. Cao and Q. Li, “Human Activity Recognition based on Random Forests”, *Proceedings of International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, Vol. 8, pp. 548-553, 2017.
 - [25] T. Hasan, M.F. Bin Karim, M.K. Mahadi, M.M. Nishat and F. Faisal, “Employment of Ensemble Machine Learning Methods for Human Activity Recognition”, *Journal of Healthcare Engineering*, Vol. 2022, No. 1, pp. 1-11, 2022.
 - [26] T.H. Tan, J.Y. Wu, S.H. Liu and M. Gochoo, “Human Activity Recognition using an Ensemble Learning Algorithm with Smartphone Sensor Data”, *Electronics*, Vol. 11, No. 3, pp. 3-22, 2022.
 - [27] R. Zhu, Z. Xiao, M. Cheng, L. Zhou, B. Yan, S. Lin and H. Wen, “Deep Ensemble Learning for Human Activity Recognition using Smartphone”, *Proceedings of International Conference on Digital Signal Processing*, Vol. 7, pp. 1-5, 2018.
 - [28] X. Huang and S. Zhang, “Human Activity Recognition based on Transformer in Smart Home”, *Proceedings of*

- International Conference on Algorithms, Computing and Machine Learning*, Vol. 23, pp. 520-525, 2023.
- [29] C.S. Leite, H. Mauranen, A. Zhanabatyrova and Y. Xiao, "Transformer-based Approaches for Sensor-based Human Activity Recognition: Opportunities and Challenges", *Proceedings of International Conference on Machine Learning*, Vol. 9, pp. 1-10, 2024.
- [30] J. Ni, H. Tang, S.T. Haque, Y. Yan and A.H.H. Ngu, "A Survey on Multimodal Wearable Sensor-based Human Action Recognition", *Proceedings of International Conference on Machine Learning*, Vol. 23, pp. 1-15, 2024.
- [31] Z. Hong, Z. Li, S. Zhong, W. Lyu, H. Wang, Y. Ding, T. He and D. Zhang, "Crosshar: Generalizing Cross-Dataset Human Activity Recognition via Hierarchical Self-Supervised Pretraining", *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 8, No. 2, pp. 64-164, 2024.
- [32] O. Napoli, D. Duarte, P. Alves, D.H.P. Soto, H.E. Oliveira, A. Rocha, L. Boc-cato and E. Borin, "A Benchmark for Domain Adaptation and Generalization in Smartphone-based Human Activity Recognition", *Scientific Data*, Vol. 11, No. 1, pp. 1-19, 2024.
- [33] S. Wang, J. Wang, H. Xi, B. Zhang, L. Zhang and H. Wei, "Optimization-Free Test-Time Adaptation for Cross-Person Activity Recognition", *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 7, No. 4, pp. 1-27, 2024.
- [34] O. Napoli and E. Borin, "On Domain Generalization for Human Activity Recognition with Mix-based Methods", *Proceedings of International Conference on Machine Learning*, Vol. 132, pp. 491-496, 2025.
- [35] P. Gopalan, L. Hu and G.N. Rothblum, "On Computationally Efficient Multi-Class Calibration", *Proceedings of International Conference on Learning Theory Research*, Vol. 247, pp. 1983-2026, 2024.
- [36] Y. Sun, P. Chaudhari, I.J. Barnett and E. Dobriban, "A Confidence Interval for the ℓ_2 Expected Calibration Error", *Proceedings of International Conference on Machine Learning*, Vol. 6, pp. 1-55, 2024.
- [37] H. Najeh, C. Lohr and B. Leduc, "Real-Time Human Activity Recognition on Embedded Equipment: A Comparative Study", *Applied Sciences*, Vol. 14, No. 6, pp. 1-11, 2024.
- [38] P. Kang, J. Moosmann, S. Bian and M. Magno, "On-Device Training Empowered Transfer Learning for Human Activity Recognition", *Proceedings of International Conference on Machine Learning*, Vol. 8, pp. 1-13, 2024.
- [39] H. Zhou, X. Zhang, Y. Feng, T. Zhang and L. Xiong, "Efficient Human Activity Recognition on Edge Devices using Deepconv LSTM Architectures", *Scientific Reports*, Vol. 15, No. 1, pp. 1-10, 2025.
- [40] C. Contoli, V. Freschi and E. Lattanzi, "Energy-Aware Human Activity Recognition for Wearable Devices: A Comprehensive Review", *Pervasive and Mobile Computing*, Vol. 104, pp. 1-7, 2024.