

Article

Adaptive TinyML with Shift-Aware Routing for Human Activity Recognition Under Distribution Shifts

Hilal Akarkamçı ^{1,*}  and Mahmut Kılıçaslan ² ¹ Directorate General for Higher and Foreign Education, Ministry of National Education, 06624 Ankara, Türkiye² Department of Statistics, Vocational School of Information Technologies, Ankara University, 06590 Ankara, Türkiye; m.kilicaslan@ankara.edu.tr

* Correspondence: hilal.akarkamci@meb.gov.tr

Abstract

Human activity recognition (HAR) on wearable and mobile devices requires models that can operate with limited computational resources while remaining reliable under changing real-world sensing conditions. However, most TinyML-based HAR systems follow the same computational path for every input, regardless of whether an activity pattern is familiar or affected by changes in the user, sensor placement, or sensing configuration. This study presents a shift-aware adaptive TinyML framework that adjusts the inference effort according to the prediction confidence and feature-space distribution shift. A lightweight multi-exit architecture allows confident samples to be classified at early stages, while uncertain or shifted samples are routed to deeper exits. A source-derived feature-space shift score complements confidence-based routing by identifying distribution changes that may not be reflected by predictive uncertainty alone. Persistent severe shifts are additionally considered through a guarded label-free test-time adaptation mechanism that does not require new activity annotations or full model retraining. The framework was evaluated on three publicly available HAR benchmark datasets, REALDISP, WISDM, and PAMAP2, covering complementary sensor-displacement, sensing-domain, unseen-user, and body-location shifts. Across the three benchmarks, explicit shift-aware routing produced a stronger adaptive computational response to changing sensing conditions than confidence-only routing. In REALDISP, the computational cost remained nearly unchanged from IDEAL to SELF conditions with confidence-only routing, whereas shift-aware routing increased the normalized computation by 0.0187 ($p = 0.0253$). In WISDM and PAMAP2, the increase in computation under the distribution shift was approximately 2.88 and 2.85 times greater with shift-aware routing than with confidence-only routing, respectively. The effect on recognition performance, however, depended on the shift scenario. Shift-aware routing significantly improved the target-domain Macro-F1 in the WISDM WATCH-to-PHONE scenario (Δ Macro-F1 = +0.0096, $p = 0.0023$), whereas no significant differences were observed for REALDISP SELF or PAMAP2 ANKLE. These findings demonstrate that explicit distribution-shift information can help resource-constrained HAR models allocate additional computation when sensing conditions change, while also showing that deeper inference alone may not overcome severe domain mismatch. The proposed framework provides a practical basis for adaptive inference in TinyML-based wearable and edge sensing systems operating under variable real-world conditions.



Academic Editor: Paolo Renna

Received: 2 September 2026

Revised: 22 September 2026

Accepted: 23 September 2026

Published: 30 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.**Keywords:** TinyML; human activity recognition; adaptive inference; distribution shift; shift-aware routing; test-time adaptation; wearable sensors; edge intelligence

1. Introduction

Human activity recognition (HAR) has become an important component of wearable and mobile sensing systems, with applications ranging from health monitoring and assisted living to sports and context-aware services [1–3]. The widespread availability of accelerometers and gyroscopes in smartphones, smartwatches, and other wearable devices has increased the feasibility of continuous activity recognition in real-world settings [4,5]. However, on-device HAR is constrained by limited memory, the computational capacity, energy budgets, and latency requirements. Tiny machine learning (TinyML) provides a practical approach for deploying compact machine learning models directly on resource-constrained edge devices [6,7].

The computational efficiency alone, however, does not ensure reliable HAR under real-world conditions. Sensor data encountered during deployment may differ from the training distribution because of unseen users, variations in sensor placement, differences between sensing devices, or changes in body location [8,9]. Such distribution shifts can lead to the substantial degradation of recognition performance even when the underlying activity classes remain unchanged [10,11]. This is particularly important for wearable HAR, where sensing conditions are inherently variable and may change during everyday use [8,12]. Practical TinyML-based HAR systems therefore need to remain computationally efficient while responding effectively to changes in the sensing environment.

Conventional HAR models generally apply a fixed computational depth to input samples, although different samples may require different amounts of computation for reliable recognition [13]. Multi-exit architectures address this limitation by allowing easily recognized samples to leave the network at shallow stages, while more difficult samples proceed to deeper layers [13,14]. However, these decisions are commonly driven by the prediction confidence, which may not reliably indicate the distribution shift. Neural models can remain overconfident or provide poorly calibrated uncertainty estimates when inputs differ from the training distribution, limiting the ability of confidence-based routing to identify changes in sensing conditions [15,16]. This suggests that adaptive inference should consider not only prediction uncertainty but also changes in the learned feature distribution.

Motivated by this problem, this study investigates when a resource-constrained HAR model should allocate additional computation. A shift-aware adaptive TinyML framework is proposed that combines lightweight multi-exit inference with prediction confidence and a source-derived feature-space shift score. Familiar and confident samples can be handled by inexpensive early exits, whereas uncertain or shifted samples can be routed to deeper stages. Persistent severe shifts can additionally trigger guarded label-free test-time adaptation without requiring new activity annotations or complete model retraining. The framework is evaluated on REALDISP, WISDM, and PAMAP2 under complementary sensor-displacement, sensing-domain, unseen-user, and body-location shift scenarios [17–19].

The main contributions of this study are as follows:

1. A shift-aware adaptive TinyML framework is introduced for HAR by integrating multi-exit inference, prediction confidence, and feature-space shift detection to dynamically control computational effort.
2. The framework is evaluated using subject-independent protocols across three complementary real-world distribution-shift scenarios involving sensor displacement, sensing-domain variation, unseen users, and body-location changes.
3. Recognition and computational behavior are jointly evaluated using performance, early-exit, memory, and computational-cost measures together with subject-level statistical analyses.

- The results show that explicit shift-aware routing responds more strongly to changing sensing conditions than confidence-only routing, while also demonstrating that additional computation does not necessarily lead to proportional recognition gains under severe domain mismatch.

2. Materials and Methods

This study proposes a shift-aware adaptive TinyML framework for HAR that adjusts computational effort according to prediction confidence and changes in the learned feature distribution. The experimental framework combines a subject-independent evaluation, multi-exit inference, feature-space shift detection, adaptive routing, and an optional label-free test-time adaptation mechanism. Each dataset was evaluated independently to preserve its original sensing characteristics and distribution-shift scenario.

An overview of the proposed shift-aware adaptive TinyML framework is presented in Figure 1.

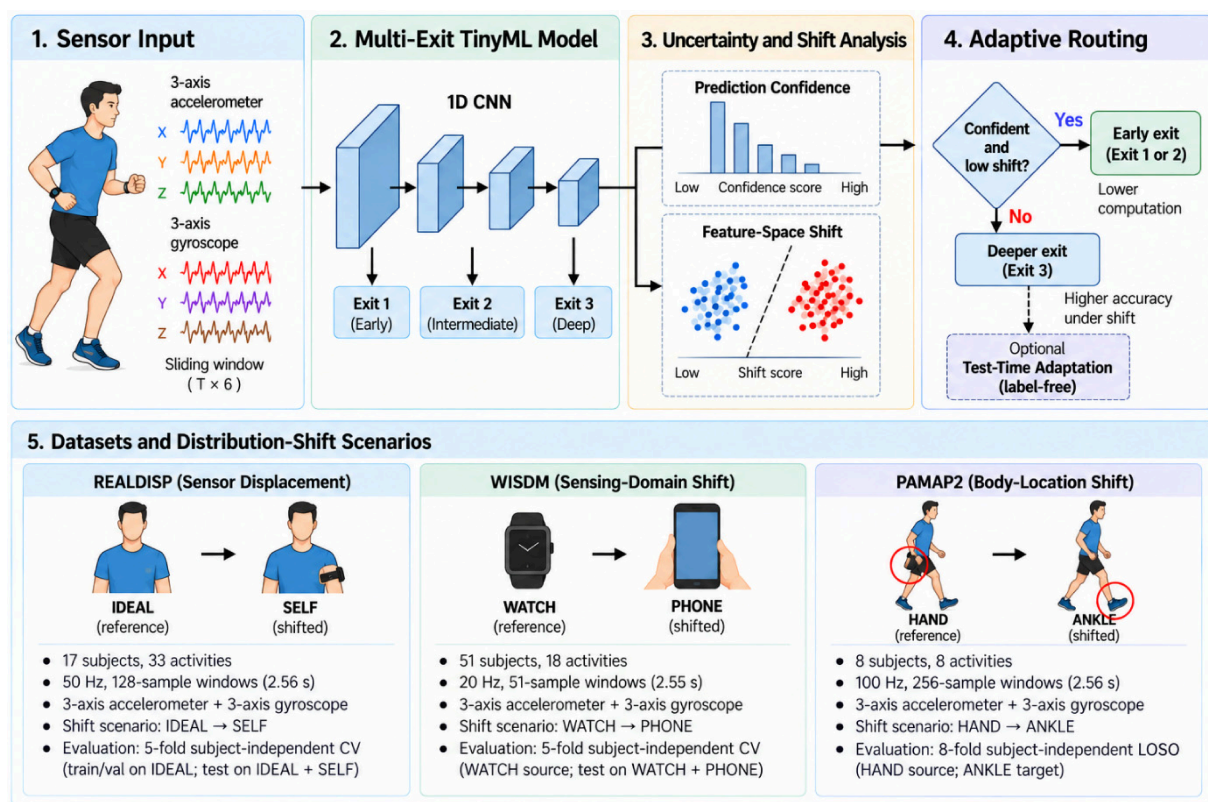


Figure 1. Proposed shift-aware adaptive TinyML framework.

2.1. Datasets and Experimental Protocol

Three public HAR datasets, REALDISP, WISDM Smartphone and Smartwatch Activity and Biometrics, and PAMAP2, were used to represent complementary real-world sensing changes [17–19]. REALDISP was used to examine sensor displacement, WISDM to evaluate sensing-domain variation under unseen-user conditions, and PAMAP2 to investigate changes in the body-worn sensor location [17–19].

For all datasets, three-axis accelerometer and gyroscope signals were used as a six-channel input. REALDISP was processed at 50 Hz using 128-sample windows, WISDM was resampled to 20 Hz and segmented into 51-sample windows, and PAMAP2 was processed at 100 Hz using 256-sample windows. Approximately 50% overlap was used in each case, resulting in comparable window durations of approximately 2.5 s. A subject-independent evaluation was adopted throughout to ensure that training and test sets contained disjoint

participants and to reduce the risk of subject-related data leakage [20]. The principal shift scenarios were IDEAL to SELF for REALDISP, WATCH to PHONE for WISDM, and HAND to ANKLE for PAMAP2. All normalization parameters were estimated exclusively from the corresponding training data. Table 1 summarizes the datasets and experimental protocols, while Figure 2 illustrates the evaluated distribution-shift scenarios.

Table 1. Dataset characteristics and experimental shift protocols.

Dataset	Subjects	Activities	Sampling	Channels	Window/Stride	Primary Shift Scenario	Evaluation Design
REALDISP	17	33	50 Hz	Acc xyz + gyro xyz	128/64 (2.56 s)	IDEAL → SELF	5-fold subject-independent CV; train/val on IDEAL; held-out IDEAL + SELF
WISDM	51	18	20 Hz	Acc xyz + gyro xyz	51/26 (~2.55 s)	WATCH → PHONE	5-fold subject-independent CV; WATCH source; held-out WATCH reference + PHONE target
PAMAP2	8	8	100 Hz	Acc xyz + gyro xyz	256/128 (2.56 s)	HAND → ANKLE	8-fold subject-independent LOSO-style evaluation; HAND source; ANKLE target

	REALDISP (Sensor Displacement)	WISDM (Sensing-Domain Shift)	PAMAP2 (Body-Location Shift)
Reference condition	IDEAL (reference)  	WATCH (reference)  	HAND (reference)  
Shifted condition	SELF (shifted)  	PHONE (shifted)  	ANKLE (shifted)  
Key difference	Same subjects and activities Different sensor placement (on the same body)		Different users Same activities Different device and sensing domain
Activities included in the dataset	33 activities: 1. Walking 2. Jogging 3. Running 4. Jump up 5. Jump front and back 6. Jump sideways 7. Jump leg/arms open/closed 8. Jump rope 9. Trunk twist (arms outstretched) 10. Trunk twist (elbows bent) 11. Waist bends forward 12. Waist rotation 13. Waist bends (reach opposite foot) 14. Reach heels backwards 15. Lateral bend 16. Lateral bend with arm up 17. Repetitive forward stretching 18. Opposite twist 19. Lateral arm elevation 20. Frontal arm elevation 21. Frontal hand claps 22. Frontal crossing of arms 23. Shoulder rotation (high) 24. Shoulder rotation (low) 25. Arms inner rotation 26. Knees to breast 27. Heels to backside 28. Crouching 29. Knees bending forward 30. Rotation on knees 31. Rowing 32. Elliptical bike 33. Cycling		6 activities: 1. Walking 2. Jogging 3. Upstairs 4. Downstairs 5. Sitting 6. Standing
			18 activities: 1. Lying 2. Sitting 3. Standing 4. Walking 5. Running 6. Cycling 7. Nordic walking 8. Watching TV 9. Computer work 10. Car driving 11. Ascending stairs 12. Descending stairs 13. Vacuum cleaning 14. Ironing 15. Ironing laundry 16. House cleaning 17. Playing soccer 18. Rope jumping

Figure 2. Experimental distribution-shift scenarios.

To facilitate reproducibility, the main training settings used for REALDISP, WISDM, and PAMAP2 are summarized in Table 2. The same core configuration was retained across the three datasets where appropriate, while a small number of dataset-specific settings was used to account for differences in their experimental protocols and training characteristics.

The main training settings were kept consistent across the three datasets, including the random seed, batch size, learning rate, weight decay, loss function, and multi-exit loss weights. Three training parameters were dataset-specific: the optimizer, maximum number of epochs, and early-stopping patience. These settings were adjusted to reflect differences in training behavior across the datasets. REALDISP used AdamW and was trained for up to 80 epochs with an early-stopping patience of 15 epochs. WISDM and PAMAP2 used Adam, with a maximum of 40 epochs and a patience of 7 epochs. In all cases, training was stopped when validation performance no longer improved. The convolutional and fully connected layers were initialized using fan-in-scaled uniform initialization, with the initialization range determined by the number of input connections to each layer. No additional custom parameter initialization was applied. The evaluation protocols were defined according to

the distribution-shift scenario considered for each dataset. WISDM followed a five-fold subject-independent protocol, PAMAP2 used an eight-fold subject-independent protocol, and REALDISP followed its sensor-displacement setting. For all three datasets, the best-performing checkpoint was selected based on validation Exit-3 Macro-F1. Normalization statistics and class weights were calculated from the corresponding training data only. The held-out test data were reserved exclusively for the final evaluation and were not used for model training or checkpoint selection.

Table 2. Training and reproducibility settings used in the experiments.

Parameter	REALDISP	WISDM	PAMAP2
Random seed	42	42	42
Batch size	512	512	512
Maximum number of epochs	80	40	40
Optimizer	AdamW	Adam	Adam
Learning rate	0.001	0.001	0.001
Weight decay	0.0001	0.0001	0.0001
Early-stopping patience	15 epochs	7 epochs	7 epochs
Checkpoint selection criterion	Validation Exit-3 Macro-F1	Validation Exit-3 Macro-F1	Validation Exit-3 Macro-F1
Weight initialization	Fan-in-scaled uniform initialization	Fan-in-scaled uniform initialization	Fan-in-scaled uniform initialization
Loss function	Class-weighted cross-entropy	Class-weighted cross-entropy	Class-weighted cross-entropy
Exit loss weights (Exit 1, Exit 2, Exit 3)	0.30, 0.30, 0.40	0.30, 0.30, 0.40	0.30, 0.30, 0.40
Normalization statistics	Training data only	Training data only	Training data only
Evaluation protocol	Shift-specific evaluation	5-fold subject-independent	8-fold subject-independent

2.2. Multi-Exit TinyML Architecture

A lightweight one-dimensional convolutional network was used as the common backbone. The architecture consists of an initial convolution followed by three depthwise-separable convolutional stages, which provide temporal feature extraction with relatively low computational complexity [21]. Classification heads were attached at three depths, denoted E_1 , E_2 , and E_3 , enabling samples to leave the network before full-depth inference [14].

The three exits were trained jointly using weighted cross-entropy (1):

$$L = 0.3L_{E1} + 0.3L_{E2} + 0.4L_{E3} \quad (1)$$

where L_{E1} , L_{E2} , and L_{E3} denote the weighted cross-entropy losses at the three exits, respectively.

The computational cost was measured using multiply–accumulate operations (MACs) and normalized relative to full E_3 inference. The same backbone was used across all datasets, with only the output layer adapted to the number of activity classes. The computational characteristics of the exits are given in Table 3.

Table 3. Three-exit TinyCNN computational characteristics.

Dataset	Classes	Parameters	FP32 Parameter Storage	Exit 1 MACs	Exit 1 Fraction	Exit 2 MACs	Exit 2 Fraction	Exit 3 MACs
REALDISP	33	12.011	~46.92 KB	150.048	0.5808	204.848	0.7929	258.368
WISDM	18	9.806	38.31 KB	60.384	0.5686	82.720	0.7790	106.192
PAMAP2	8	8.336	32.56 KB	298.240	0.5813	406.912	0.7932	513.024

2.3. Feature-Space Shift Detection and Adaptive Routing

To complement prediction confidence, the distribution shift was estimated from the pooled feature representation of $E1$. Motivated by distance-based detection in learned feature spaces [22], class-specific means and variances were calculated using source-domain training samples only. For an incoming feature vector z , the shift score was defined as follows (2):

$$S(z) = \frac{1}{d} \sum_{j=1}^d \frac{(z_j - \mu_{c,j})^2}{\sigma_{c,j}^2 + \epsilon} \quad (2)$$

where z_j is the j th feature. $\mu_{c,j}$ and $\sigma_{c,j}^2$ are the class-specific mean and variance estimated from the source training data. c is the predicted class. d is the feature dimension, and ϵ is a small constant for numerical stability.

A sequential confidence-based router was used as the adaptive-inference baseline. At each intermediate exit, samples exceeding the predefined confidence threshold were classified immediately, whereas the remaining samples were forwarded to deeper stages [23]. The proposed shift-aware router extended this strategy by incorporating the feature-space shift score into the routing decision. Samples with high shift scores could therefore be directed to deeper exits even when prediction confidence alone would favor early termination. All routing thresholds were calibrated exclusively using source-domain validation data and were fixed before evaluation under the shifted target condition. Confidence thresholds were obtained from validation-set confidence quantiles and evaluated as candidate early-exit operating points. The operating points were selected according to the computational cost required to retain a predefined proportion of the full-depth validation Macro-F1. Shift thresholds were derived independently from the source-validation shift-score distribution, with the upper quantiles representing increasingly severe feature-space deviation. Because the confidence and shift-score distributions differed across datasets and evaluation folds, the resulting thresholds were fold-specific rather than fixed globally. All thresholds were derived from the corresponding source-validation data and then kept unchanged during target-domain evaluation.

As an additional routing baseline, confidence-based routing was compared with compute-matched random routing. The random router used the same numbers of Exit 1, Exit 2, and Exit 3 decisions as the confidence-based router, while randomly assigning samples to these exits across 1,000 permutations. The analysis was also repeated at the subject level, with the same exit counts maintained separately for each subject.

To further assess the shift detector, two additional analyses were conducted using the sample-level outputs from the PAMAP2 HAND-to-ANKLE experiment. First, the feature-space shift score was compared with predictive entropy and maximum softmax probability (MSP) uncertainty as alternative uncertainty-based criteria for distinguishing the source HAND condition from the shifted ANKLE condition. Detection performance was evaluated using AUROC, AUPR, and FPR95. Second, the sensitivity of the shift score to classification errors was examined by comparing correctly classified and misclassified samples within each sensing condition. HAND and ANKLE were analyzed separately to determine whether elevated shift scores could be attributed to prediction errors rather than to the sensing-location shift itself.

2.4. Guarded Test-Time Adaptation

Persistent severe shifts were additionally considered triggers for label-free test-time adaptation [24]. Adaptation was restricted to batch normalization (BN) affine parameters and performed through entropy minimization [24]. A safety guard based on source-validation behavior rejected unstable updates and restored the previous model state when

prediction behavior exceeded predefined stability limits. Adaptation updates were not included in the inference MAC analysis. Since the frozen experimental outputs did not include separate measurements of update latency, energy consumption, memory use, or update-related MACs, adaptation activity is reported based on the update frequency rather than as a hardware-level computational cost.

2.5. Evaluation and Statistical Analysis

Recognition performance was evaluated using the accuracy, Macro-F1, and Matthews correlation coefficient (MCC) [25]. Computational behavior was assessed using the normalized MAC cost, savings relative to full-depth inference, and exit usage. Shift detection was further examined using feature-space scores and AUROC [26].

Statistical comparisons were performed at the subject level to avoid treating overlapping windows as independent observations [20]. Paired Wilcoxon signed-rank tests were used for matched comparisons [27], with $p < 0.05$ considered statistically significant. Recognition and computational responses were analyzed separately, allowing increased computational allocation to be distinguished from actual improvements in predictive performance.

Decision-path interpretability was additionally examined at the sample level to characterize how confidence and feature-space shift information influenced adaptive routing decisions. Inference paths were traced using $E1$ prediction confidence, the source-derived shift score, confidence-based and shift-aware exit decisions, and the final prediction outcome. Three representative routing patterns were considered: an early correct prediction under low shift, a shift-triggered transition to deeper inference resulting in a correct prediction, and a severe-shift case in which deeper inference remained incorrect. Examples corresponding to these patterns were selected from the frozen sample-level routing results.

3. Results

The experimental results were analyzed from three complementary perspectives: the effect of distribution shift on recognition, the ability of the feature-space score to identify changing sensing conditions, and the computational and predictive response of adaptive routing. Results were evaluated independently for REALDISP, WISDM, and PAMAP2 under their corresponding subject-independent protocols.

3.1. Distribution Shift and Feature-Space Detection

The three datasets exhibited distinct levels of performance degradation under the evaluated sensing shifts. REALDISP showed a clear reduction in recognition performance from IDEAL to SELF conditions, while the WATCH-to-PHONE transition in WISDM produced a substantially stronger domain effect. The HAND-to-ANKLE transition in PAMAP2 represented the most severe mismatch, with full-depth inference alone unable to recover the loss of recognition performance.

The proposed feature-space score consistently reflected these changes in sensing conditions. In REALDISP, the mean shift score increased by 111.75% from IDEAL to SELF across all five folds. In WISDM, PHONE samples showed a 77.70% higher mean shift score than WATCH samples, with higher target-domain scores observed for 43 of 51 subjects. The HAND-to-ANKLE transition in PAMAP2 produced an absolute mean increase of 3.583 in the shift score, with higher ANKLE scores for all eight subjects. As illustrated in Figure 3, these findings indicate that the source-derived feature representation provided a consistent signal of distribution change across different types of sensing shifts.

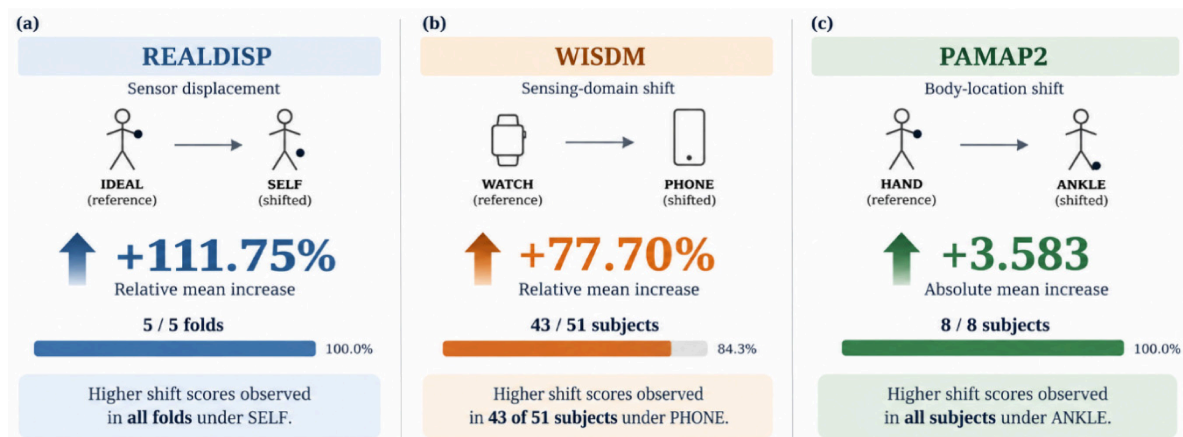


Figure 3. Feature-space shift detection across datasets. (a) REALDISP: IDEAL to SELF sensor-displacement shift, showing a 111.75% increase in the mean shift score across all five folds. (b) WISDM: WATCH to PHONE sensing-domain shift, showing a 77.70% increase in the mean shift score, with higher scores observed for 43 of 51 subjects. (c) PAMAP2: HAND to ANKLE body-location shift, showing an absolute mean shift-score increase of 3.583, with higher scores observed for all eight subjects. Higher shift scores indicate greater deviation from the source-domain feature representation.

The PAMAP2 HAND-to-ANKLE setting was examined in greater detail to assess whether the feature-space score captured the distribution shift beyond prediction uncertainty. This setting represented the most severe distribution shift and the most challenging recognition condition among the three datasets. The proposed score achieved an AUROC of 0.8821 and an AUPR of 0.8595 for distinguishing ANKLE from HAND samples, compared with AUROCs of 0.6434 for predictive entropy and 0.6148 for MSP uncertainty (Table 4). The corresponding FPR95 was 0.3578 for the proposed score, substantially lower than 0.9531 for predictive entropy and 0.9581 for MSP uncertainty. The same pattern was observed at the subject level. The proposed score achieved a mean AUROC of 0.8957 ± 0.0818 across the eight held-out subjects and yielded a higher AUROC than both uncertainty-based criteria for every subject.

Table 4. Sample-level comparison of shift-detection criteria for the PAMAP2 HAND-to-ANKLE setting, selected for detailed analysis as the most severe distribution-shift scenario.

Criterion	AUROC	AUPR	FPR95
Proposed feature-space shift score	0.8821	0.8595	0.3578
Predictive entropy	0.6434	0.6104	0.9531
MSP uncertainty	0.6148	0.5847	0.9581

The sensitivity of the class-conditioned shift score to classification errors was also examined. Under the source HAND condition, misclassified samples had higher shift scores than correctly classified samples (median 1.0174 vs. 0.5834, $p < 0.001$), indicating that prediction errors were associated with higher scores in the reference condition. Under the shifted ANKLE condition, however, the shift scores of correctly classified and misclassified samples did not differ significantly (median 2.2928 vs. 2.6794, Mann–Whitney U test, $p = 0.146$; rank-biserial correlation = -0.034). In addition, 59.17% of correctly classified ANKLE samples and 66.33% of misclassified ANKLE samples exceeded the severe-shift threshold. Thus, although classification errors can influence the class-conditioned score, the strong shift response observed for ANKLE cannot be attributed solely to incorrect class predictions.

3.2. Multi-Exit Performance and Computational Trade-Off

The three exits provided different recognition–computation trade-offs across the datasets. Deeper inference generally improved recognition, although the magnitude of

the gain varied considerably with the sensing condition. Under the shifted REALDISP SELF condition, Macro-F1 increased from 0.444 at E1 to 0.498 at E2 and 0.523 at E3. For WISDM PHONE, the corresponding values were 0.145, 0.164, and 0.170, while PAMAP2 ANKLE achieved 0.025, 0.029, and 0.045. These results show that additional computation can improve recognition in shifted conditions, but the improvement becomes limited when the source-target mismatch is severe.

The computational cost increased progressively across exits. Relative to full E3 inference, E1 required approximately 57–58% of the computation and E2 approximately 78–79%, depending on the dataset. Figure 4 summarizes this performance–computation trade-off, while Table 5 provides the recognition results for fixed-exit and adaptive routing strategies.

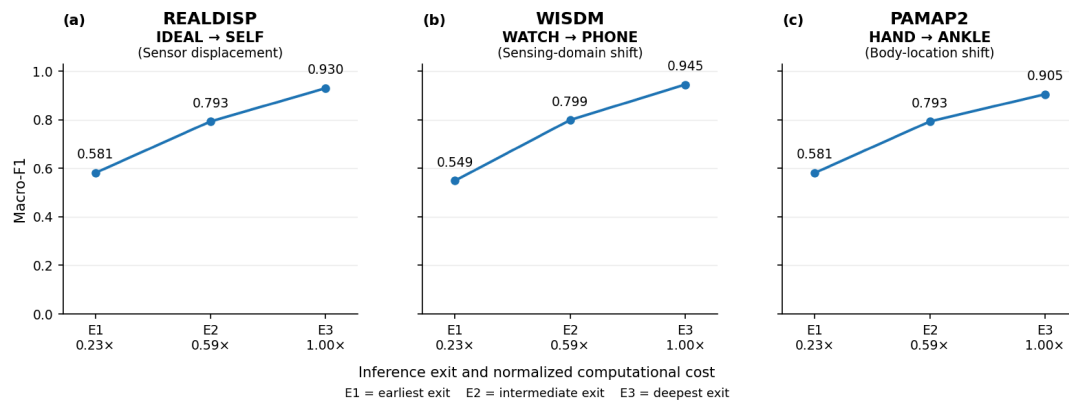


Figure 4. Multi-exit performance–compute trade-off in the principal shifted target conditions. E1 denotes the earliest exit, E2 the intermediate exit, and E3 the deepest exit (full model). Macro-F1 values are based on the same fold- or subject-level aggregation used for each dataset. Computational cost is normalized relative to full E3 inference.

Table 5. Recognition performance of fixed exits and adaptive routing in the principal shifted target conditions. Macro-F1 values are reported using the corresponding fold- or subject-level aggregation for each dataset.

Dataset/Target	E1 Macro-F1	E2 Macro-F1	E3 Macro-F1	Confidence-Based Macro-F1	Shift-Aware Macro-F1	Δ Macro-F1	<i>p</i> -Value
REALDISP/SELF	0.4440	0.4976	0.5226	0.4861	0.4852	−0.0009	0.4216
WISDM/PHONE	0.1451	0.1642	0.1701	0.1234	0.1330	+0.0096	0.0023
PAMAP2/ANKLE	0.0252	0.0293	0.0452	0.0242	0.0409	+0.0167	0.3125

Note: *p*-values correspond to paired comparisons between confidence-based and shift-aware routing.

In-domain held-out performance was also examined to distinguish the effect of the distribution shift from limitations of the underlying classifier. Based on REALDISP, the static TinyCNN achieved mean Macro-F1 scores of 0.7111 and 0.5359 under the IDEAL and SELF conditions, respectively. For the multi-exit model, using the same fold- or subject-level aggregation applied in the cross-condition analysis, the mean Exit-3 Macro-F1 decreased from 0.6785 under REALDISP IDEAL to 0.5226 under SELF, from 0.6172 based on WISDM WATCH to 0.1701 based on PHONE, and from 0.6923 based on PAMAP2 HAND to 0.0452 based on ANKLE. These results show that the performance loss was particularly pronounced under the sensing-domain and body-location shifts in WISDM and PAMAP2. They also clarify that the purpose of the adaptive multi-exit framework is not to improve recognition accuracy over a static classifier, but to adjust computational allocation according to input confidence and detected distribution shift.

3.3. Computational Response to Real Distribution Shifts

The central analysis examined whether the model increased computational allocation when sensing conditions changed. Confidence-only routing showed almost no computational response to the IDEAL-to-SELF transition in REALDISP, with a mean change in normalized computation of -0.0002 . In contrast, shift-aware routing increased computation by 0.0187 , with an increase observed for 13 of 17 subjects ($p = 0.0253$).

The same pattern was more pronounced for the other two datasets. In WISDM, normalized computation increased by 0.0395 under confidence-only routing and by 0.1137 under shift-aware routing. The shift-aware response was approximately 2.88 times greater, and the paired response advantage favored shift-aware routing for 40 of 51 subjects ($p < 0.001$). In PAMAP2, the corresponding increases were 0.0522 and 0.1488 , giving an approximately 2.85-fold stronger response. Seven of eight subjects showed a larger computational response with shift-aware routing ($p = 0.0078$).

In addition to the relative computational response, the router-level computational cost was examined using the exit-specific MAC values reported in Table 3. For REALDISP SELF, mean normalized computation was 0.8528 for confidence-only routing and 0.8775 for shift-aware routing, corresponding to savings of 14.72% and 12.25% relative to full-depth inference. Exit 1, Exit 2, and Exit 3 usage changed from 19.60% , 31.41% , and 48.99% to 13.70% , 31.41% , and 54.89% , respectively. For WISDM PHONE, mean normalized computation increased from 0.6527 to 0.7796 , with savings of 34.73% and 22.04% , while the corresponding exit distributions changed from 70.11% , 20.26% , and 9.62% to 32.33% , 36.62% , and 31.05% . For PAMAP2 ANKLE, mean normalized computation increased from 0.7249 to 0.8521 , corresponding to savings of 27.51% and 14.79% , with exit usage changing from 46.23% , 39.42% , and 14.35% to 15.87% , 39.42% , and 44.72% . Shift-aware routing allocated more samples to deeper inference under the shifted conditions while remaining less computationally demanding than full-depth inference.

Figure 5 summarizes these cross-dataset results. Explicit feature-space shift information produced a stronger computational response to changing sensing conditions than prediction confidence alone.

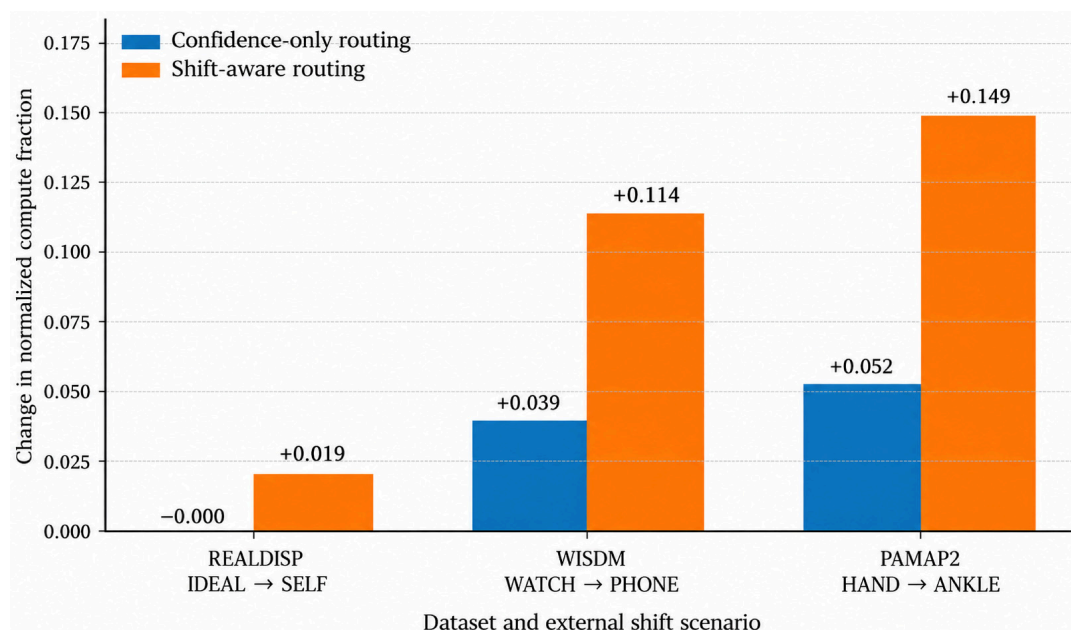


Figure 5. Computational response of confidence-based and shift-aware routing to real distribution shifts across the three datasets. Changes in normalized computation are based on subject-level comparisons between the corresponding source and shifted target conditions.

3.4. Recognition Response and Statistical Analysis

A stronger computational response did not translate uniformly into improved recognition. In REALDISP SELF, subject-level Macro-F1 remained essentially unchanged, decreasing slightly from 0.4861 with confidence-only routing to 0.4852 with shift-aware routing ($\Delta = -0.0009$, $p = 0.4216$). Thus, additional computation was allocated under sensor displacement without a significant change in recognition performance.

In WISDM PHONE, shift-aware routing increased the subject-level Macro-F1 from 0.1234 to 0.1330 ($\Delta = +0.0096$), with improvements observed for 34 of 51 subjects. This difference was statistically significant ($p = 0.0023$). PAMAP2 showed a larger numerical increase from 0.0242 to 0.0409 ($\Delta = +0.0167$), but the difference was not statistically significant ($p = 0.3125$). The very low absolute performance under HAND-to-ANKLE transfer further indicates that additional computation alone was insufficient to overcome this severe body-location mismatch.

The guarded label-free TTA mechanism was also evaluated using the frozen five-fold REALDISP outputs. Based on the IDEAL condition, Macro-F1 changed only marginally from 0.6845 without TTA to 0.6846 with either persistent or guarded BN-TTA. Under SELF displacement, the corresponding Macro-F1 values were 0.5158 without TTA, 0.5155 with persistent TTA, and 0.5157 with guarded TTA. Persistent TTA performed updates on 0.54% of IDEAL samples and 1.58% of SELF samples. Across both conditions, the guarded mechanism accepted 1.741 updates and rejected 310, corresponding to 0.91% and 0.16% of all samples, respectively. These results indicate that the guard limited adaptation activity but did not produce a meaningful recognition improvement over the non-adapted baseline.

As an additional routing baseline, confidence-based routing was compared with compute-matched random routing using the frozen exit-specific predictions. For each evaluation unit, the random router preserved exactly the same Exit 1, Exit 2, and Exit 3 usage counts as the confidence-only router, while randomly reassigning samples to those exits. Results were averaged over 1,000 random permutations and evaluated using pooled predictions across the available evaluation conditions within each dataset. Confidence-based routing achieved a higher Macro-F1 than compute-matched random routing on all three datasets: 0.5916 versus 0.5624 ± 0.0006 for REALDISP, 0.3909 versus 0.3865 ± 0.0003 for WISDM, and 0.3952 versus 0.3907 ± 0.0010 for PAMAP2. The corresponding mean exit was identical for the two strategies within each dataset (2.2864, 1.2980, and 1.5572, respectively). These results indicate that confidence-based allocation was more effective than random assignment under the same computational budget.

To determine whether this finding was also maintained at the subject level, the compute-matched random-routing analysis was repeated separately for each subject, preserving the original Exit 1, Exit 2, and Exit 3 usage counts for that subject across 1,000 random permutations. Confidence-based routing achieved a higher subject-level Macro-F1 than random routing for all 17 REALDISP subjects, with a mean paired difference of +0.0306 (one-sided paired Wilcoxon $p < 0.0001$). In WISDM, confidence-based routing was better for 33 of 51 subjects, with a mean paired difference of +0.0041 ($p = 0.0004$). In PAMAP2, confidence-based routing was better for six of eight subjects, with a mean paired difference of +0.0037, but the difference was not statistically significant ($p = 0.1250$). These findings confirm the confidence-routing advantage at the subject level for REALDISP and WISDM, while PAMAP2 showed only a numerical advantage.

A local sensitivity analysis was performed on PAMAP2 because the HAND-to-ANKLE setting represented the most severe distribution shift and the most challenging recognition condition among the three datasets. The validation-derived confidence and severe-shift thresholds were varied from 0.8 to 1.2 times their selected values, while keeping the trained models fixed. Using pooled sample-level predictions for the PAMAP2 HAND-to-ANKLE

condition, varying the confidence thresholds resulted in Macro-F1 values ranging from 0.0632 to 0.0704, with the mean exit ranging from 2.2078 to 2.3829. Varying the severe-shift threshold produced a narrower Macro-F1 range of 0.0654 to 0.0684, while the mean exit ranged from 2.1915 to 2.3987. Recognition performance remained relatively stable around the selected operating points, whereas the allocation of computation across exits was more sensitive to threshold selection. Although the same perturbation analysis was not repeated for REALDISP and WISDM, the fold-specific calibration procedure was applied consistently across all three datasets. Similar sensitivity cannot be assumed for REALDISP and WISDM because the confidence and shift-score distributions differ across datasets and folds.

A class-wise error analysis provided further insight into the severe PAMAP2 HAND-to-ANKLE condition. Using pooled sample-level predictions, shift-aware routing increased the Macro-F1 from 0.0348 for confidence-only routing to 0.0674, while full-depth Exit-3 inference achieved a Macro-F1 of only 0.0770. The errors were strongly activity-dependent. Walking was not correctly recognized in any of its 1.853 target samples, with 89.64% classified as vacuum cleaning. Similarly, 82.66% of lying samples were classified as sitting, while 80.73% of descending-stairs samples and 71.30% of ascending-stairs samples were classified as vacuum cleaning. In contrast, shift-aware routing produced clearer gains for sitting and standing, increasing their F1 scores from 0.1432 to 0.1826 and from 0.0012 to 0.1879, respectively. These results show that the severe body-location shift produced systematic activity-specific confusion rather than a uniform reduction in recognition performance. The limited performance of full-depth inference further indicates that allocating additional computation alone cannot compensate for this degree of representation mismatch.

Figure 6 illustrates the recognition response across the three target conditions, while Table 6 summarizes the subject-level statistical evidence for the main cross-dataset findings. Together, these results distinguish the primary computational effect of shift-aware routing from its dataset-dependent influence on recognition performance.

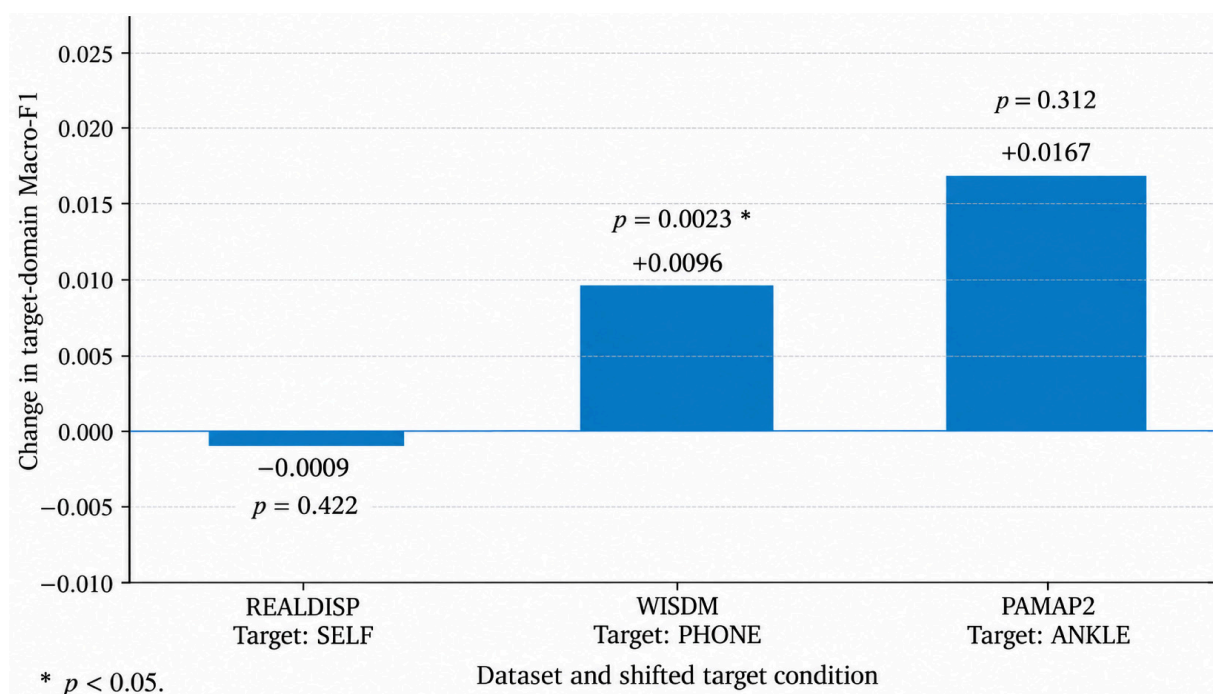


Figure 6. Recognition response to shift-aware routing in the principal shifted target conditions. Macro-F1 changes represent subject-level comparisons between confidence-based and shift-aware routing.

Table 6. Subject-level statistical comparison of computational and recognition responses to shift-aware routing across the three distribution-shift scenarios.

Dataset/Shift	Confidence Compute Response	Shift-Aware Compute Response	Δ Compute	Subject-Level Evidence	Δ Macro-F1
REALDISP IDEAL $\rightarrow$ SELF	−0.0002	+0.0187	+0.0189	13/17 increased under shift; $p = 0.0253$	−0.0009 ($p = 0.4216$)
WISDM WATCH $\rightarrow$ PHONE	+0.0395	+0.1137	+0.0742	40/51 larger shift-aware response; $p = 4.06 \times 10^{-6}$	+0.0096 ($p = 0.0023$)
PAMAP2 HAND $\rightarrow$ ANKLE	+0.0522	+0.1488	+0.0966	7/8 larger shift-aware response; $p = 0.0078$	+0.0167 ($p = 0.3125$)

Note: computational responses represent within-subject changes from the source to the corresponding shifted target condition. Macro-F1 changes compare confidence-based and shift-aware routing in the shifted target condition. Reported p -values are based on paired subject-level comparisons.

3.5. Decision-Path Interpretability

To better understand how the proposed framework makes routing decisions, representative inference paths were examined in terms of the prediction confidence, feature-space shift score, selected exit, and final recognition outcome. Rather than conventional post-hoc feature attribution, this analysis focuses on the decision process that determines how much computation is allocated to each sample.

Table 7 presents three representative decision paths from the WISDM and PAMAP2 experiments. The low-shift WISDM source sample was retained at E1, while the severely shifted WISDM target sample was redirected from E1 to E3 and correctly classified. In the PAMAP2 example, the severe shift similarly triggered deeper inference, but the final prediction remained incorrect. Together, these examples show how the detected distribution shift influences computational allocation, while highlighting that deeper inference does not necessarily resolve severe domain mismatch.

Table 7. Sample-level decision paths of shift-aware routing.

Example	Shift State	Confidence-Based Route	Selected Route	Outcome
WISDM Source WATCH	Low shift \nS = 0.702	High confidence \nE1 = 1.000	E1	Correct \nTrue: 1 Pred: 1
WISDM Target PHONE	Severe shift \nS = 17.752	E1 \nE1 = 0.475	Shift override $\rightarrow$ E3	Correct \nTrue: 1 Pred: 1
PAMAP2 Target ANKLE	Severe shift \nS = 3.791; q95 = 3.789	E1 \nE1 = 0.695	Shift override $\rightarrow$ E3	Incorrect \nTrue: 3 Pred: 6

4. Discussion

The results indicate that adaptive computation in TinyML-based HAR should not rely solely on prediction confidence. Across all three datasets, incorporating feature-space shift information produced a stronger computational response to changing sensing conditions than confidence-only routing. This was particularly clear in REALDISP, where confidence-based computation remained almost unchanged under sensor displacement, while shift-aware routing increased the use of deeper inference. Similar responses were observed for the sensing-domain shift in WISDM and the body-location shift in PAMAP2. These findings suggest that explicit distribution-shift information can complement confidence when deciding how much computation an input requires.

This distinction is important because prediction confidence may remain high even for shifted or unfamiliar inputs [28]. The source-derived feature-space score provided an additional signal for identifying such changes without requiring target labels. However, allocating more computation did not consistently translate into better recognition. Shift-aware routing produced a significant Macro-F1 improvement in WISDM, approximately

preserved the performance in REALDISP, and provided no significant improvement in PAMAP2. The severe HAND-to-ANKLE mismatch further showed that deeper inference alone may be insufficient under substantial source-target domain mismatch, consistent with previous evidence that changes in sensor placement and body position can introduce considerable distribution differences in HAR [29].

An important implication of these findings is the distinction between computational adaptation and predictive adaptation. Dynamic inference can account for differences among individual inputs rather than applying the same fixed inference process to every sample [30]. The proposed shift-aware routing strategy follows this principle by directing shifted inputs to deeper stages of the network when additional computation is needed. However, the different recognition outcomes across the three datasets show that increasing the computational effort does not necessarily lead to better recognition. This becomes particularly evident under severe distribution shifts, where allocating more computation may not improve recognition or sufficiently compensate for the mismatch between source and target domains.

The test-time adaptation experiments led to a similar conclusion. Although label-free test-time adaptation can improve robustness to changing target distributions [31], unreliable or accumulated updates during online adaptation may also degrade model performance [32]. The proposed safety guard reduced harmful adaptation behavior but did not provide a consistent recognition benefit. TTA is therefore better viewed as an optional fallback for persistent severe shifts rather than a primary mechanism for improving recognition. This interpretation is particularly relevant in light of recent HAR-specific developments. For example, Contrastive Online Adaptation for HAR (COA-HAR) uses self-supervised contrastive learning and sensor-data augmentation to adapt representations to unlabeled test data, illustrating the potential of representation-level adaptation for wearable HAR under a distribution shift [33]. Zhang et al. [34] similarly proposed a contrastive domain-adaptation framework combining time-lag attention with class-conditional teacher–student alignment for cross-domain wearable sensor time series. In contrast, the present study primarily focuses on computational adaptation, using the shift signal to determine how much inference should be allocated to each input rather than directly aligning source and target representations. Representation-adaptation methods are therefore complementary to the proposed routing framework and may become particularly useful when severe shifts persist despite deeper inference.

The PAMAP2 results also suggest when deeper inference may no longer be sufficient and representation adaptation may be needed. A severe shift can initially justify routing samples to deeper exits, but the HAND-to-ANKLE results show that additional computation provides limited benefit when the representation itself is strongly affected by the shift. In such cases, persistent severe-shift detection combined with low confidence at the deepest exit could serve as a label-free trigger for representation adaptation. This would allow the system to distinguish between samples that may benefit from deeper inference and those that require adaptation of the learned representation. Although this adaptation trigger was not implemented in the present study, evaluating such a mechanism across different types and severities of distribution shifts remains an important direction for future work.

The present study nevertheless has several limitations. It was restricted to inertial HAR datasets, and computational efficiency was assessed primarily based on the normalized MAC cost, which does not directly capture hardware-dependent costs such as latency, energy consumption, RAM usage, or the flash footprint [7,35]. Direct measurements based on representative embedded hardware are therefore needed to assess practical deployment benefits [36]. In addition, the class-conditioned shift score relies on the predicted class, making it potentially sensitive to classification errors. Although the PAMAP2 analysis showed

that the observed shift response could not be explained solely by incorrect predictions, prediction-independent alternatives such as global or class-aggregated shift measures warrant further investigation. Finally, the PAMAP2 analysis included only eight subjects. The non-significant Macro-F1 difference ($p = 0.3125$) should therefore be interpreted cautiously because of the limited statistical power, and larger subject cohorts are needed to assess whether the observed numerical difference generalizes.

5. Conclusions

This study presented a shift-aware adaptive TinyML framework for resource-constrained HAR that combines multi-exit inference with prediction confidence and feature-space shift detection. An evaluation based on REALDISP, WISDM, and PAMAP2 showed that explicit shift awareness enables the model to adapt its inference depth to changing sensing conditions rather than applying the same computational path to every input. The main contribution therefore lies in integrating feature-space shift detection with adaptive multi-exit routing within a unified computational adaptation framework, rather than in either component alone. The results also showed that deeper computation does not necessarily translate into improved recognition under all distribution shifts: significant gains were obtained based on WISDM, whereas the more severe domain mismatch in PAMAP2 limited the benefit of additional inference. The stronger computational response of shift-aware routing does not necessarily indicate improved recognition robustness, particularly under severe distribution shifts.

These findings highlight an important distinction between computational adaptivity and recognition robustness in wearable and edge-based intelligent sensing systems. By selectively allocating deeper inference to uncertain or shifted samples, the proposed framework provides a basis for input-dependent computation in resource-constrained TinyML applications. This approach avoids unnecessary full-depth processing for all inputs and is consistent with the broader principle of input-dependent computation in dynamic neural networks [37]. This makes shift-aware routing particularly relevant to embedded HAR systems operating with variable users, sensor placements, and sensing configurations. Future work will extend this framework toward learnable sample-wise routing and exit policies and will evaluate its practical efficiency through direct measurements of latency, energy consumption, and memory usage based on embedded hardware. Robustness to severe and previously unseen sensing shifts will also be investigated through selective and anti-forgetting test-time adaptation strategies, including Efficient Anti-Forgetting Test-Time Adaptation (EATA) [38] and Sharpness-Aware and Reliable Entropy Minimization (SAR) [39].

Author Contributions: Conceptualization: H.A. and M.K.; Methodology: H.A.; Software: H.A. and M.K.; Validation: H.A. and M.K.; Formal Analysis: H.A. and M.K.; Investigation: H.A.; Resources: M.K.; Writing—Original Draft Preparation: H.A. and M.K.; Writing—Review and Editing: H.A. and M.K.; Visualization: H.A. and M.K.; Supervision: H.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available from the UCI Machine Learning Repository. The REALDISP Activity Recognition Dataset is available at <https://doi.org/10.24432/C5GP6D>, the WISDM Smartphone and Smartwatch Activity and Biometrics Dataset at <https://doi.org/10.24432/C5HK59>, and the PAMAP2 Physical Activity Monitoring Dataset at <https://doi.org/10.24432/C5NW2H>.

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

References

1. Lara, O.D.; Labrador, M.A. A survey on human activity recognition using wearable sensors. *IEEE Commun. Surv. Tutor.* **2013**, *15*, 1192–1209. [\[CrossRef\]](#)
2. Nweke, H.F.; Teh, Y.W.; Al-garadi, M.A.; Alo, U.R. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Syst. Appl.* **2018**, *105*, 233–261. [\[CrossRef\]](#)
3. Wang, J.; Chen, Y.; Hao, S.; Peng, X.; Hu, L. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognit. Lett.* **2019**, *119*, 3–11. [\[CrossRef\]](#)
4. Ramanujam, E.; Perumal, T.; Padmavathi, S. Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review. *IEEE Sens. J.* **2021**, *21*, 13029–13040. [\[CrossRef\]](#)
5. Straczekiewicz, M.; James, P.; Onnela, J.-P. A systematic review of smartphone-based human activity recognition methods for health research. *Npj Digit. Med.* **2021**, *4*, 148. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Ray, P.P. A review on TinyML: State-of-the-art and prospects. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 1595–1623. [\[CrossRef\]](#)
7. Alajlan, N.N.; Ibrahim, D.M. TinyML: Enabling of inference deep learning models on ultra-low-power IoT edge devices for AI applications. *Micromachines* **2022**, *13*, 851. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Ye, X.; Sakurai, K.; Nair, N.-K.C.; Wang, K.I.-K. Machine learning techniques for sensor-based human activity recognition with data heterogeneity-A review. *Sensors* **2024**, *24*, 7975. [\[CrossRef\]](#) [\[PubMed\]](#)
9. San-Segundo, R.; Blunck, H.; Moreno-Pimentel, J.; Stisen, A.; Gil-Martín, M. Robust human activity recognition using smartwatches and smartphones. *Eng. Appl. Artif. Intell.* **2018**, *72*, 190–202. [\[CrossRef\]](#)
10. Chakma, A.; Faridee, A.Z.M.; Khan, M.A.A.H.; Roy, N. Activity recognition in wearables using adversarial multi-source domain adaptation. *Smart Health* **2021**, *19*, 100174. [\[CrossRef\]](#)
11. Xu, Q.; Wei, X.; Bai, R.; Li, S.; Meng, Z. Integration of deep adaptation transfer learning and online sequential extreme learning machine for cross-person and cross-position activity recognition. *Expert Syst. Appl.* **2023**, *212*, 118807. [\[CrossRef\]](#)
12. Kunze, K.; Lukowicz, P. Sensor placement variations in wearable activity recognition. *IEEE Pervasive Comput.* **2014**, *13*, 32–41. [\[CrossRef\]](#)
13. Yu, J.; Zhang, L.; Cheng, D.; Bu, C.; Liu, L.; Wu, H.; Song, A. Ensemble early exit network on human activity recognition using wearable sensors. *Comput. Netw.* **2025**, *269*, 111409. [\[CrossRef\]](#)
14. Teerapittayanon, S.; McDanel, B.; Kung, H.T. BranchyNet: Fast inference via early exiting from deep neural networks. In *Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR)*; IEEE: New York, NY, USA, 2016; pp. 2464–2469. [\[CrossRef\]](#)
15. Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.V.; Lakshminarayanan, B.; Snoek, J. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 13991–14002.
16. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*; PMLR: Cambridge, UK, 2017; Volume 70, pp. 1321–1330.
17. Baños, O.; Damas, M.; Pomares, H.; Rojas, I.; Tóth, M.A.; Amft, O. A benchmark dataset to evaluate sensor displacement in activity recognition. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*; Association for Computing Machinery: New York, NY, USA, 2012; pp. 1026–1035. [\[CrossRef\]](#)
18. Weiss, G. *WISDM Smartphone and Smartwatch Activity and Biometrics Dataset*; [Dataset]; UCI Machine Learning Repository: Irvine, CA, USA, 2019. [\[CrossRef\]](#)
19. Reiss, A. *PAMAP2 Physical Activity Monitoring*; [Dataset]; UCI Machine Learning Repository: Irvine, CA, USA, 2012. [\[CrossRef\]](#)
20. Dehghani, A.; Sarbishei, O.; Glatard, T.; Shihab, E. A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors. *Sensors* **2019**, *19*, 5026. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Lee, Y.-K.; Son, C.-S.; Kang, W.-S. Depthwise-separable U-Net for wearable sensor-based human activity recognition. *Appl. Sci.* **2025**, *15*, 9134. [\[CrossRef\]](#)
22. Lee, K.; Lee, K.; Lee, H.; Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 7167–7177.
23. Passalis, N.; Raitoharju, J.; Tefas, A.; Gabbouj, M. Efficient adaptive inference for deep convolutional neural networks using hierarchical early exits. *Pattern Recognit.* **2020**, *105*, 107346. [\[CrossRef\]](#)
24. Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; Darrell, T. Tent: Fully test-time adaptation by entropy minimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Virtual, 3–7 May 2021.
25. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [\[CrossRef\]](#)
27. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1945**, *1*, 80–83. [\[CrossRef\]](#)
28. Kim, T.; Lee, J.-S. Gently sloped and extended classification margin for overconfidence relaxation of out-of-distribution samples. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 12854–12865. [\[CrossRef\]](#) [\[PubMed\]](#)

29. Napoli, O.; Duarte, D.; Alves, P.; Palo Soto, D.H.; de Oliveira, H.E.; Rocha, A.; Boccato, L.; Borin, E. A benchmark for domain adaptation and generalization in smartphone-based human activity recognition. *Sci. Data* **2024**, *11*, 1192. [[CrossRef](#)] [[PubMed](#)]
30. Qian, X.; Liu, F.; Jiao, L.; Zhang, X.; Huang, X.; Li, S.; Chen, P.; Liu, X. Knowledge transfer evolutionary search for lightweight neural architecture with dynamic inference. *Pattern Recognit.* **2023**, *143*, 109790. [[CrossRef](#)]
31. Liang, J.; He, R.; Tan, T. A comprehensive survey on test-time adaptation under distribution shifts. *Int. J. Comput. Vis.* **2025**, *133*, 31–64. [[CrossRef](#)]
32. Niu, S.; Chen, G.; Chen, D.; Zhang, Y.; Wu, J.; Wen, Z.; Chen, Y.; Zhao, P.; Miao, C.; Tan, M. Adapt in the wild: Test-time entropy minimization with sharpness and feature regularization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**. Advance online publication. [[CrossRef](#)] [[PubMed](#)]
33. Fortes Rey, V.; Martelletto Bressane Rezende, P.; Zhou, B.; Suh, S.; Lukowicz, P. COA-HAR: Exploring contrastive online test-time adaptation for wearable sensor-based human activity recognition using sensor data augmentation. *Expert Syst. Appl.* **2026**, *297*, 129288. [[CrossRef](#)]
34. Zhang, G.; Zhang, H.; Jia, M.; Liu, Z. Contrastive domain adaptation with time-lag attention for wearable sensor time-series. *Expert Syst. Appl.* **2026**, *308*, 131087. [[CrossRef](#)]
35. Heydari, S.; Mahmoud, Q.H. Tiny machine learning and on-device inference: A survey of applications, challenges, and future directions. *Sensors* **2025**, *25*, 3191. [[CrossRef](#)] [[PubMed](#)]
36. Qureshi, T.S.; Shahid, M.H.; Farhan, A.A.; Alamri, S. A systematic literature review on human activity recognition using smart devices: Advances, challenges, and future directions. *Artif. Intell. Rev.* **2025**, *58*, 276. [[CrossRef](#)]
37. Han, Y.; Huang, G.; Song, S.; Yang, L.; Wang, H.; Wang, Y. Dynamic neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7436–7456. [[CrossRef](#)] [[PubMed](#)]
38. Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; Tan, M. Efficient test-time model adaptation without forgetting. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; Proceedings of Machine Learning Research. Volume 162, pp. 16888–16905.
39. Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; Tan, M. Towards stable test-time adaptation in dynamic wild world. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.