# WatchHAR: Real-time On-device Human Activity Recognition System for Smartwatches


Taeyoung Yeon
Computer Science
Northwestern University
Evanston, Illinois, USA
taeyoungyeon@northwestern.edu

Vasco Xu
Department of Computer Science
University of Chicago
Chicago, Illinois, USA
vascoxu@uchicago.edu

Henry Hoffmann
University of Chicago
Chicago, Illinois, USA
hankhoffmann@cs.uchicago.edu

Karan Ahuja
Computer Science
Northwestern University
Evanston, Illinois, USA
kahuja@northwestern.edu


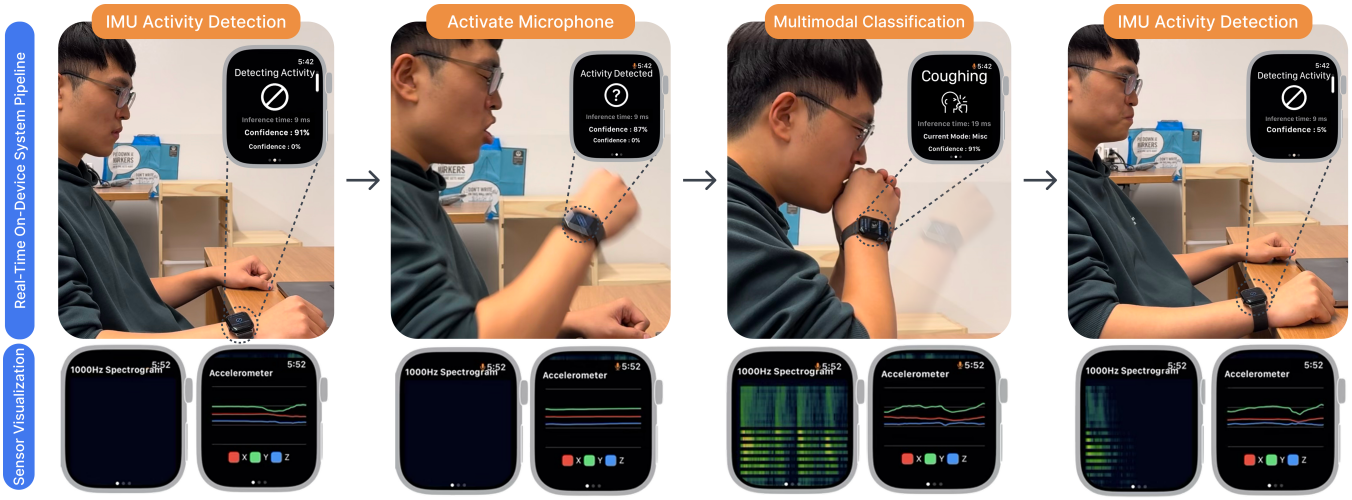


**Figure 1: Our application runs in real-time directly on a smartwatch, detecting and classifying human activities. The system first uses IMU data to detect the presence of an event. Once an event is detected, it triggers a multimodal activity classifier that activates the microphones and uses both IMU and audio data to classify the activity.**


## Abstract

Despite advances in practical and multimodal fine-grained Human Activity Recognition (HAR), a system that runs entirely on smartwatches in unconstrained environments remains elusive. We present WatchHAR, an audio and inertial-based HAR system that operates fully on smartwatches, addressing privacy and latency issues associated with external data processing. By optimizing each component of the pipeline, WatchHAR achieves compounding performance gains. We introduce a novel architecture that unifies sensor data preprocessing and inference into an end-to-end trainable module, achieving 5x faster processing while maintaining over 90% accuracy across more than 25 activity classes. WatchHAR outperforms state-of-the-art models for event detection and activity classification while running directly on the smartwatch, achieving 9.3 ms processing time for activity event detection and 11.8 ms for multimodal activity classification. This research advances on-device activity recognition, realizing smartwatches' potential as standalone, privacy-aware, and minimally-invasive continuous activity tracking devices.



## CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; • **Applied computing** → Health informatics.

## Keywords

Smartwatches, On-device processing, Real-time mobile sensing, Human activity recognition, Privacy aware sensing




*ICMI '25, Canberra, ACT, Australia*



ACM ISBN 979-8-4007-1499-3/25/10
https://doi.org/10.1145/3716553.3750775

**ACM Reference Format:**
Taeyoung Yeon, Vasco Xu, Henry Hoffmann, and Karan Ahuja. 2025. WatchHAR: Real-time On-device Human Activity Recognition System for Smartwatches. In *Proceedings of the 27th International Conference on Multimodal Interaction (ICMI '25), October 13–17, 2025, Canberra, ACT, Australia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3716553.3750775

## 1 Introduction

Human Activity Recognition (HAR) has become a cornerstone of ubiquitous computing, with applications ranging from health monitoring and context-aware services to assistive technologies for people who are deaf or hard of hearing. While significant strides have been made in developing accurate and robust HAR systems for Activities of Daily Living (ADL), a persistent challenge has been creating solutions that are both practical for everyday use and capable of operating in unconstrained environments. Smartwatches, with their array of sensors and constant proximity to users, present an ideal platform for HAR. However, most current smartwatch-based sensing systems rely on external data processing, raising concerns about privacy, latency, and the need for constant connectivity to remote machines to offload compute [2, 6, 8, 19].

On-device processing directly addresses these concerns by ensuring real-time feedback and safeguarding sensitive information. For example, workers in construction or manufacturing often operate in areas with limited connectivity; delayed notifications due to reliance on remote servers can undermine safety or productivity. Similarly, elder-care systems require continuous tracking to detect emergencies in real-time. In such cases, offloading sensitive health data to external infrastructure not only risks privacy breaches but also increases response latency.

WatchHAR addresses these challenges by introducing a novel HAR system that operates entirely on smartwatches, leveraging both audio and inertial data (Figure 2). We chose IMU and audio sensors as they are universally available on smartwatches and provide complementary information – audio provides distinctive signatures for audible activities but is power-intensive, while IMUs are lightweight and can capture fine-grained hand movements but produce less distinct signals. By eliminating the need for external data processing, WatchHAR enhances privacy and reduces latency while maintaining high accuracy across a wide range of activities. The system employs a two-stage approach: a lightweight IMU-based activity detector that triggers a more resource-intensive multimodal classifier only when necessary. This strategy optimizes power consumption without sacrificing performance (Section 3.1). This is enabled by WatchHAR's end-to-end trainable preprocessing module, which applies a Short-Time Fourier Transform (STFT) and approximates a mel-filter bank as a 1D convolutional operation that runs efficiently on mobile neural processors.

Through careful optimization of each component of the pipeline, WatchHAR achieves compounding performance gains. The system outperforms the latest models for event detection (by 5.5%, Table 1) and even achieves modest improvements in activity classification (by 0.7%) while running entirely on the smartwatch, with processing times of 9.3 ms for activity detection and 11.8 ms for multimodal classification. WatchHAR's novel architecture unifies sensor data preprocessing and inference into a single, trainable module, providing a 5x performance boost while maintaining over 90% accuracy


**Figure 2: Our WatchHAR system running in real-time on an Apple Watch Series 7 (45mm), demonstrating activity recognition across four different contexts: brushing hair, hammering, washing dishes, and clapping.**

across more than 25 activity classes. These advancements demonstrate WatchHAR's potential to revolutionize on-device activity recognition, realizing the full potential of smartwatches as standalone devices with minimally invasive activity tracking.

Our main contributions are:

(1) A complete multimodal HAR system running entirely on commodity smartwatches with real-time performance (9.3 ms event detection, 11.8 ms activity classification);

(2) An end-to-end trainable audio preprocessing module integrating STFT and mel-filter banks directly into the neural network for efficient on-device execution;

(3) Comprehensive evaluation demonstrating 5× faster processing and 5-47× lower computational cost (FLOPs) than state-of-the-art while maintaining competitive accuracy;

(4) Open-source implementation, models, and evaluation scripts to foster reproducibility and community adoption [1].

## 2 Related Work

Human Activity Recognition (HAR) has seen significant advancements in recent years, particularly in the domain of wearable technology. A wide range of wearable devices have been explored for HAR, including wrist-worn sensors [3, 4, 18, 20], smart rings [26], earbuds [28], and smartwatches [8, 9, 17, 19, 27]. These devices have proven effective in identifying various activities, from fitness exercises to daily tasks.

Smartwatches, housing a rich collection of sensors including IMUs and microphones, have proven highly effective for activity recognition in everyday settings. For example, SAMoSA [19] and Bhattacharya et al. [8] highlight the benefits of combining audio and IMU data. SAMoSA achieved 92.2% accuracy across various contexts using 1kHz audio and 50Hz IMU data, demonstrating that even lower-sampled audio can enhance activity recognition while preserving privacy. They proposed an IMU-based activity detector that activates the microphone only upon detecting an activity, enabling efficient multimodal classification. Bhattacharya et al. [8] showcased robust multimodal sensor fusion techniques,

[1] https://github.com/SPICExLAB/WatchHAR

performing well in both controlled and real-world scenarios. Despite these advancements, most systems treat smartwatches merely as data collectors and offload processing to smartphones or desktops [8, 19, 26, 27]. This approach, while computationally effective, compromises privacy and real-time responsiveness.

Some efforts have been made towards on-device processing, such as Kim et al.'s [15] exercise monitoring system using natural magnetism in exercise equipment, and Zhang et al.'s [28] cough detection system that uses IMU sensor values to activate cough detection. Kunwar et al. [17] also explored robust and deployable gesture recognition for smartwatches. However, these solutions primarily target a limited range of classes and utilize IMU data, avoiding the power-hungry and computationally intensive audio processing despite its rich contextual information.

The key challenge lies in developing a system that can leverage the rich information from both audio and IMU sensors to support the fidelity of HAR while operating entirely on resource-constrained wearable devices. This requires not only efficient algorithms but also novel approaches to sensor data processing, gating and fusion. WatchHAR overcomes these challenges by implementing a 1D convolution approach for generating log-mel spectrograms and combining it with efficient convolutional classifier architectures, allowing the model to run on smartwatch neural accelerators in real-time.

## 3 Model Architecture

Building upon prior work, SAMoSA[19], which introduced the concepts of the IMU Event Detector and Multimodal Activity Classifier, we have significantly improved the model architecture to optimize for on-device application. Our system architecture maintains these two core components while balancing computational efficiency and power consumption for real-time activity recognition on smartwatches (Figure 3).

To ensure our system runs efficiently on smartwatches, we implement several optimization techniques on the trained models. We first traced our PyTorch models using torch.jit.trace with example tensors matching our expected input shapes with batch size 1 for simulating real-time inference. This intermediate representation was then converted to CoreML [5] for optimized execution on Apple Watch hardware. We apply 16-bit float quantization, which halves the model size and improves inference speed with no impact on accuracy. All inference runs on the Apple Watch Series 7 (45mm) GPU via CoreML for optimal performance. We carefully tune the window sizes and hop lengths for both the event detector and classifier to balance between accuracy, latency, and computational load. These optimizations enable our system to run in real-time on commodity smartwatch hardware while maintaining high accuracy across a wide range of activities.

### 3.1 IMU-Only Event Detector

We use a lightweight IMU-based event detector to trigger the more resource-intensive multimodal classifier. Our detector uses a 1D depthwise Convolutional Neural Network (CNN) architecture [12], processing 3-second windows of 6-axis IMU data (3-axis accelerometer and 3-axis gyroscope) sampled at 50 Hz. The model consists of four convolutional blocks with increasing filter counts (64 to 128) and decreasing kernel sizes (10 to 5), interspersed with max pooling layers, followed by fully connected layers (512, 256, 128 nodes) and a final sigmoid output for binary event detection (activity detected or not).

To ensure rapid detection of event onset while minimizing false positives from small motions, we use a 3-second rolling window with a 20 ms hop length and apply a 2-second moving average to filter out spurious detections. Our system uses a two-stage detection process to balance performance and power consumption. The IMU-Only Event Detector continuously monitors for the presence of an event and triggers the Multimodal Activity Classifier, activating the microphone, only when an event is detected. Otherwise, the microphone remains off, conserving energy during idle periods.

### 3.2 Multimodal Activity Classification

Our Multimodal Activity Classifier (Figure 3) processes both IMU and audio data to achieve high-accuracy activity recognition. Since audio adds a significant computational overhead, we use shorter window sizes to enable faster processing and reduce latency. Both IMU and audio data use 1-second windows with a 20 ms hop length, allowing for fine-grained temporal resolution in our classifications.

We implement an end-to-end trainable audio preprocessing module directly within our neural network [11]. Our architecture consists of three main components: a Short-Time Fourier Transform (STFT) implemented using a 1D convolutional layer, a mel-filter bank designed as a trainable linear layer, and an amplitude-to-DB conversion using a logarithmic activation function. The STFT layer uses two separate convolutions for the real and imaginary parts, with kernel size corresponding to the FFT size and stride determining the hop length. The mel-filter bank layer is initialized with triangular mel filters but remains trainable, potentially learning optimized filter banks for our specific human activity recognition task. This is key, as filters suited for HAR may differ greatly from those originally designed for speech recognition tasks. Lastly, the amplitude-to-DB conversion uses a logarithmic activation function to produce a spectrogram that is then passed to the audio feature encoder.

For the audio encoder, we use a MobileNetV3 [14] backbone pretrained on the AudioSet dataset [13], without any platform-specific modifications. This choice offers a good balance between model size, computational efficiency, and accuracy. The MobileNetV3 architecture incorporates inverted residual blocks with squeeze-and-excitation modules, platform-aware neural architecture search for optimized layer design, and efficient last-stage design for classification tasks.

For the IMU encoder, we adopt the ConvBoost architecture [21]. This model uses a standard 3-layer 2D CNN structure optimized for efficient processing of multivariate time series data such as IMU signals. Each layer uses 5×1 kernels to extract temporal features, with max pooling applied after the first two convolutional layers. ReLU activations are used throughout the network for non-linearity. The classifier consists of two fully connected layers with dropout ($p = 0.5$) regularization to prevent overfitting. This architecture prioritizes simplicity and computational efficiency while maintaining effective feature extraction for human activity recognition tasks.

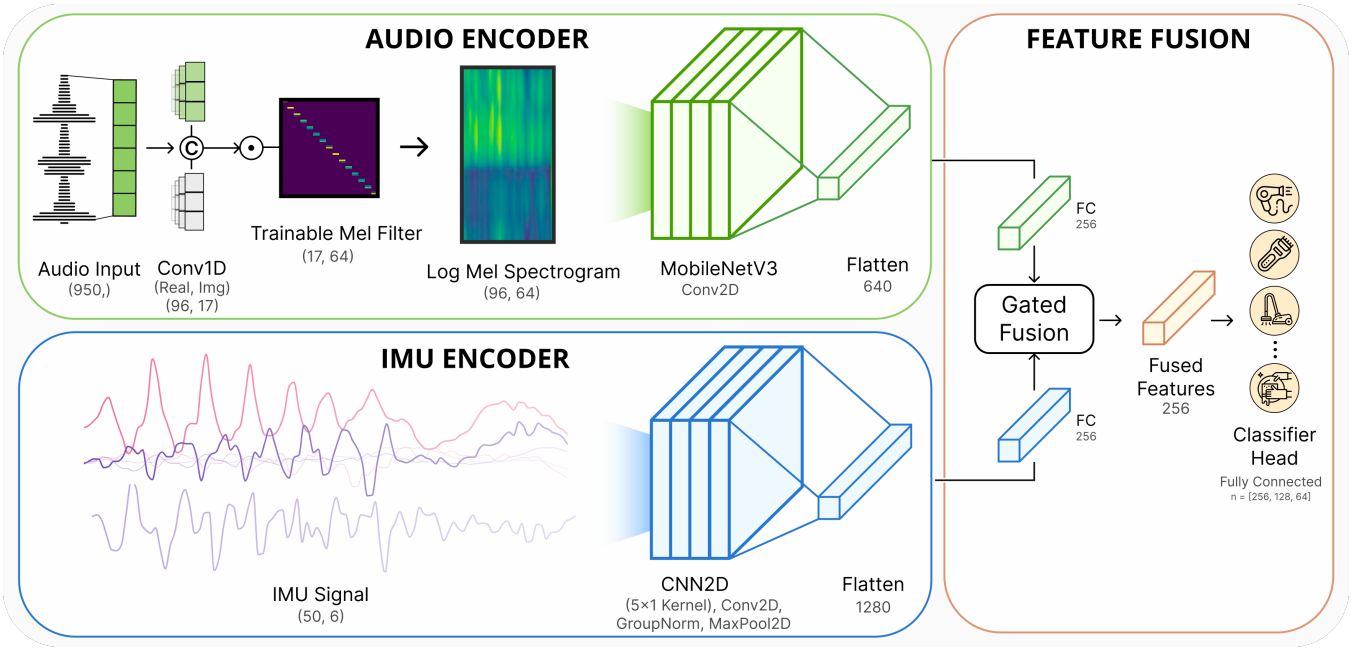


**Figure 3: Overview of WatchHAR's Multimodal Activity Classifier. The model processes 1-second windows of raw audio and 6-axis IMU data to predict activities. Audio preprocessing is integrated directly into the neural network as trainable layers.**

To effectively combine information from IMU and audio modalities, we implement a Gated Fusion mechanism [7] rather than simple feature concatenation. Both IMU and audio embeddings are first projected into a shared 256-dimensional space using separate linear layers. Each projected embedding is then passed through its own gating network – a linear layer followed by a sigmoid activation – that produces a vector of weights between 0 and 1. These weights are applied to the projected features through element-wise multiplication, effectively scaling each feature dimension by its learned importance. The gated IMU and audio features are then summed together and passed through an additional linear layer to produce the final 256-dimensional fused representation. This fused vector is then passed through a final classifier head to predict class probabilities. All components are fully differentiable and jointly trained with the rest of the model.

The gated fusion mechanism is trained end-to-end with binary cross-entropy loss, learning to weight modalities based on their relevance to each activity. This approach is particularly effective in the P-LOPO setting (Section 6) where some personalization data is available for user adaptation.

## 4 Datasets

To train and evaluate WatchHAR, we utilize the following publicly available smartwatch datasets, which were collected in previous studies.

**SAMoSA Dataset [19]:** The SAMoSA dataset was collected from 20 participants (mean age 23.3, all right-handed) across 60 diverse environments. Data was recorded using a Fossil Gen 5 smartwatch, capturing synchronized 9-axis IMU data (accelerometer, gyroscope, and orientation) at 50 Hz and uncompressed audio at 16 kHz, later post-processed to 1 kHz for privacy preservation. The dataset includes 26 activities performed in four contexts: kitchen, bathroom, workshop, and miscellaneous. Each participant performed every activity three times per context, resulting in 14.2 hours of data in total – 5.9 hours of labeled activity data and 8.3 hours of transition ("Other") data. All activities were performed in participants' homes using their own appliances and tools, naturally incorporating ambient background noise. This dataset is used to train and evaluate both activity detection and classification.

**Semi-Naturalistic and In-the-Wild Dataset [8]:** We used two complementary datasets from Bhattacharya et al.'s work. First, the **Semi-Naturalistic dataset** was collected from 15 participants (9 female, 6 male, mean age 43.6), representing diverse professional and socioeconomic backgrounds. Data was captured using a Fossil Gen 4 smartwatch, recording accelerometer and gyroscope data at 50 Hz, and audio data at 22.05 kHz. Participants performed 23 activities twice across two sessions, with each activity lasting a minimum of 30 seconds. Data collection was conducted remotely via video calls, and participants knocked on a surface to mark the start and end of activities. Continuous recordings captured all activities, including in-between movements, and manual annotation was performed using sensor data and video footage.

In addition, an **In-the-Wild dataset** was gathered from five additional participants (4 males, 1 female, mean age 27). These participants wore the same smartwatch alongside a smartphone mounted on their chest, which captured 25-second egocentric video clips every minute using a dedicated mobile application. Data collection for the in-the-wild study was performed without any predefined activity scripts, allowing participants to engage in their daily routines

**Table 1: Event Detection Model Performance Comparison between SAMoSA and WatchHAR on Apple Watch Series 7**

| Method | F1 score (%) | Processing Time (ms) | Onset Latency (sec) | Offset Latency (sec) |
|---|---|---|---|---|
| **SAMoSA [IMU @ 50 Hz]** | 88.0 | 55.4 | 0.62 | 0.16 |
| **WatchHAR [IMU @ 50 Hz]** | **93.5** | **9.3** | **0.27** | **0.07** |

naturally over two separate sessions each, totaling 10 in-the-wild sessions.

## 5 Evaluation

We evaluate WatchHAR's performance against prior smartwatch-based approaches, focusing on aspects critical for real-time applications: processing time, model size, and accuracy across different settings. All models were implemented using PyTorch version 2.1.2 and converted to CoreML format using coremltools version 7.1 with float16 quantization for running on Apple Devices. Converting 32-bit models to 16-bit had no impact on accuracy while halving the model size. On-device performance evaluations were conducted using an Apple Watch Series 7 (45mm) GPU via CoreML library.

### 5.1 IMU-Only Event Detector

We compare our Depthwise CNN1D Event Detection Model with SAMoSA's Random Forest event detection model, as shown in Table 1. WatchHAR outperforms SAMoSA across all metrics, achieving higher F1 scores (93.5% vs. 88.0%). As noted in Section 3.1, to address potential mispredictions due to data skewness, we applied a 2-second moving average to the output probabilities, smoothing predictions and reducing spurious outputs. This improved our F1 score from 92.5% to 93.5% with negligible computational overhead.

Compared to SAMoSA, WatchHAR demonstrates significantly faster processing times (9.3 ms vs. 55.4 ms)[2]. Processing time includes data preprocessing and model inference for a single IMU window. WatchHAR's processing efficiency stems from streamlined GPU feature computation. SAMoSA calculates eight statistical features – mean, standard deviation, max, min, median, variance, skewness, and kurtosis – for each of the nine IMU values, requiring 28.76 ms on the Apple Watch Series 7 (45mm). In contrast, our 1D CNN only normalizes raw IMU data, completing in just 3 ms. We also measured onset latency (i.e., the delay between the physical start of an event and its detection by the model) and offset latency (i.e., the delay in detecting the end of an event). WatchHAR achieves lower onset (0.27s vs 0.62s) and offset (0.07s vs 0.16s) latencies compared to SAMoSA.

### 5.2 Multimodal Activity Classification

We compare our Multimodal Activity Classifier with two prior works, SAMoSA [19] and Bhattacharya et al. [8] on their respective datasets, as summarized in Table 2. All models were evaluated using a leave-one-participant-out (LOPO) cross-validation scheme or Personalized-LOPO (P-LOPO), which incorporates a subset of test participants' data during training to simulate partial personalization. In our P-LOPO evaluation, we use personalized data from public datasets, though future work could explore on-device fine-tuning to preserve user privacy. For the In-the-Wild dataset, we report weighted-F1 scores evaluated in a P-LOPO protocol, aligning with the evaluation protocol from Bhattacharya et al. [8]. We also compare processing times, measured on an Apple Watch Series 7 (45mm), which reflects the total time to process a single window of audio and IMU data, from log-mel generation to prediction.

On the SAMoSA dataset, we compare with the authors' pretrained model configured for 1 kHz audio and 50 Hz motion data, matching their primary evaluation setup. We compute the context-wise accuracy defined as the average classification accuracy across four contexts: kitchen, bathroom, workshop, and miscellaneous. WatchHAR achieves slightly higher (92.34% vs 92.2%) context-wise accuracy with 47× lower computational cost (0.036 GFLOPs vs 1.71 GFLOPs) and 5× faster processing time (11.8 ms vs 56.4 ms).

On the Semi-Naturalistic dataset, we reproduced the best performing architecture from their paper. CNN14 [16] for audio and Attend&Discriminate [1] for IMU with concatenation late fusion method, since no pretrained model was provided. WatchHAR achieves slightly better LOPO accuracy (90.4% vs 89.7%) and comparable P-LOPO accuracy (93.8% vs 94.3%), while requiring 11× fewer FLOPs (0.917 GFLOPs vs 4.24 GFLOPs) and achieving 6× faster inference time (71.1 ms vs 438.3 ms). Note that our FLOPs increased from the SAMoSA dataset (0.036 GFLOPs) to the Semi-Naturalistic dataset (0.917 GFLOPs) due to adjusting our CNN2D IMU encoder to handle the longer 10-second input windows compared to 1-second windows used for SAMoSA dataset.

On the In-the-Wild dataset, we followed Bhattacharya et al.'s evaluation protocol: WatchHAR was first pretrained using the Semi-Naturalistic dataset and then fine-tuned with within-session data, leading to a personalized leave-one-participant-out (P-LOPO) evaluation. Similar to Bhattacharya et al., we augmented the training data with Semi-Naturalistic samples to address class imbalance and missing labels. WatchHAR achieves higher weighted F1 scores compared to Bhattacharya et al. (56.7% vs 55.8%). This drop in accuracy – relative to those achieved on the Semi-Naturalistic dataset (over 90%) – are expected, as the In-the-Wild dataset contains noisier ground truth labels due to limitations in the video-based annotation system used during data collection. Despite these challenges, our system achieves a higher F1 score while requiring 11× lower computational cost. WatchHAR, without any per-user fine-tuning and using only the model trained on the Semi-Naturalistic dataset, obtains a weighted F1 score of 28.5 compared to 26.8 from Bhattacharya et al. [8].

We refer readers to the Appendix for per-activity confusion matrices on the SAMoSA, Semi-Naturalistic, and In-the-Wild dataset across each activity context. All our models, code, and evaluation

[2]SAMoSA originally reported 4.8 ms on MacBook Air (M1). We measured both methods on Apple Watch Series 7 (45mm) for fair comparison. WatchHAR's event detector inference time is 2.1 ms on M1 hardware

**Table 2: Multimodal Activity Classification across different approaches and datasets. For SAMoSA & Semi-Naturalistic, their evaluation metrics are LOPO or P-LOPO. Their values are classification accuracies (%). Note, for In-the-Wild dataset, their evaluation metrics are weighted F1 scores.**

| Model Name | Dataset | Sampling Rate (kHz) | LOPO | P-LOPO | Processing Time (ms) | FLOPs (G) |
|---|---|---|---|---|---|---|
| **SAMoSA** | SAMoSA | 1 | 92.2 | N/A | 56.4 | 1.71 |
| **WatchHAR** | SAMoSA | 1 | **92.34** | N/A | **11.8** | **0.036** |
| **Bhattacharya et al.** | Semi-Naturalistic | 22.05 | 89.7 | **94.3** | 438.3 | 4.24 |
| **WatchHAR** | Semi-Naturalistic | 22.05 | **90.4** | 93.8 | **71.1** | **0.917** |
| **Bhattacharya et al.** | In-the-Wild | 22.05 | N/A | 55.8 | 438.3 | 4.24 |
| **WatchHAR** | In-the-Wild | 22.05 | N/A | **56.7** | **71.1** | **0.917** |

scripts are publicly available at https://github.com/SPICExLAB/WatchHAR to foster community use and adoption.

## 6 Ablation Study

We analyze how different architectural choices affect the accuracy and computational complexity through an ablation on the *Semi-Naturalistic* dataset. We explore performance-efficiency tradeoffs across different pretrained audio backbone models, IMU encoder architectures, and multimodal feature fusion techniques. Model variants are trained and evaluated using the protocols outlined in Sections 3.2 and 5.2, respectively. Additionally, we report model size and computational cost (FLOPs) to assess their suitability for on-device applications.

We evaluated four publicly available pretrained audio encoders — CNN14, ResNet-22, MobileNetV1, and our choice of MobileNetV3 — all pretrained on AudioSet [16]. Of note, we excluded the pretrained VGGish model used in SAMoSA as it is hard-coded for 96×64 spectrogram inputs, making it incompatible with the 690×64 dimensions of the *Semi-Naturalistic* dataset. As Table 3 shows, ResNet-22 yields the highest LOPO score (79.8%), but MobileNetV3 attains the best P-LOPO (86.7%) while being ~30× smaller and ~13× lighter in FLOPs.

We also evaluated four IMU backbone architectures: CNN1D from SAMoSA [19], DeepConvLSTM [21], Attend&Discriminate [1], and our choice of CNN2D [21]. Unlike the audio encoders, all IMU encoders were initialized from scratch. As shown in Table 4, Attend&Discriminate achieved the highest P-LOPO accuracy (90.8%), consistent with prior results from Bhattacharya et al. [8], while our CNN2D achieved the best LOPO performance (85.5%) with the lowest computational cost (0.13 GFLOPs).

Finally, we reproduced the three late-fusion schemes proposed by Bhattacharya et al.[8]: (1) simple concatenation of modality embeddings, (2) cross-modal self-attention, and (3) softmax averaging of independent classifications. We compare these schemes with our gated-fusion method, which adds small gated block that learns per-sample modality weights. As shown in Table 5, softmax averaging achieved the highest LOPO accuracy (90.5%) but lower P-LOPO performance (92.8%). Our gated-fusion approach nearly matched the LOPO score while improving P-LOPO accuracy to 93.8%, with minimal computational overhead (<0.5M parameters, negligible FLOPs increase). These results demonstrate that our fusion method particularly excels when user-specific data is available for fine-tuning.

| Model | LOPO | P-LOPO | Params (M) | FLOPs (G) |
|---|---|---|---|---|
| CNN14 | 74.5 | 82.4 | 80.80 | 14.54 |
| ResNet-22 | **79.8** | 86.5 | 63.73 | 10.36 |
| MobileNetV1 | 77.8 | 85.5 | 5.36 | 1.25 |
| MobileNetV3 | 77.8 | **86.7** | **2.19** | **0.79** |

**Table 3: Audio Model Ablation Study Results.**

| Model | LOPO | P-LOPO | Params (M) | FLOPs (G) |
|---|---|---|---|---|
| DeepConvLSTM | 72.0 | 77.1 | **0.72** | 0.19 |
| CNN1D | 75.9 | 80.4 | 246.80 | 0.56 |
| Attend&Discriminate | 84.0 | **90.8** | 0.97 | 0.35 |
| CNN2D | **85.5** | 89.0 | 4.57 | **0.13** |

**Table 4: IMU Model Ablation Study Results.**

| Fusion Method | LOPO | P-LOPO | Params (M) | FLOPs (G) |
|---|---|---|---|---|
| Concatenation | 89.4 | 93.2 | 6.76 | 0.917 |
| Self-Attention | 89.8 | 93.3 | 8.53 | 0.918 |
| Softmax Averaging | **90.5** | 92.8 | **6.76** | **0.917** |
| Gated Fusion | 90.4 | **93.8** | 7.18 | 0.917 |

**Table 5: Multimodal Fusion Ablation Study results.**

## 7 Application Scenarios

WatchHAR broadens the scope of activity recognition by running entirely on-device, enabling real-time, privacy-preserving, and on-the-go applications across diverse application domains (Figure 4). For instance, it can be used for behavioral intervention. WatchHAR can instantly recognize gestures such as nail biting, delivering a gentle vibration, logging the event, or prompting a brief intervention - supporting real-time behavior change without external devices [22].

It can also be used for worker safety in factory settings. Manual tasks in industrial settings – such as drilling, lifting, or assembly –

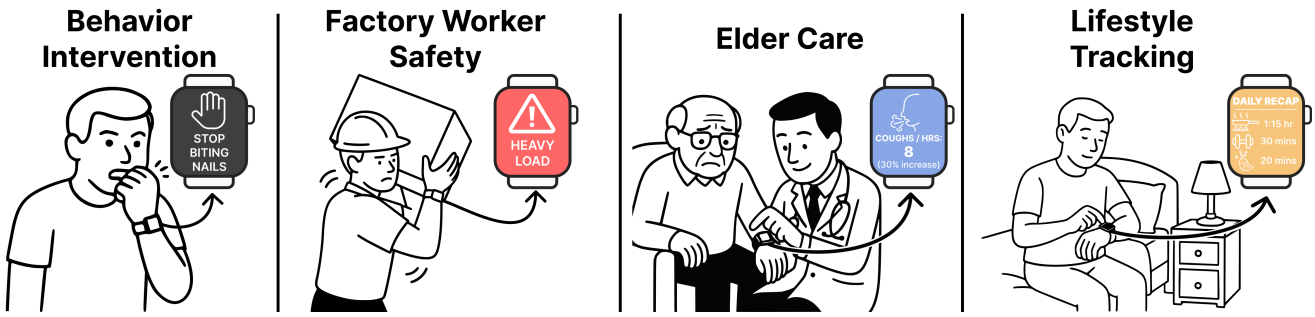


**Figure 4: Illustrations of four representative application scenarios enabled by WatchHAR. *Left→right:* Behavior Intervention—recognition of nail-biting triggers haptic nudges; Worker Safety—real-time detection of over-the-shoulder lifting issues on-the-spot warnings; Elder Care—the watch tracks cough frequency and summarizes trends for clinicians or family; Lifestyle Tracking—an on-device diary reports time spent on daily activities, giving users actionable insights without exporting raw data.**

produce distinctive motion and audio patterns. Early systems recognized these patterns using body-worn sensors [24], but required specialized hardware and offline processing. WatchHAR brings these capabilities to scale using only off-the-shelf smartwatches, enabling real-time procedure tracking, automated task logging, and alerts for high-risk movements like over-the-shoulder lifts.

WatchHAR can also be used for health sensing and elder care. By monitoring daily motion profiles [25] for actions such as walking, dressing, or coughing entirely on the device, it can passively track digital health biomarkers over time. Similarly, it can be used to turn raw sensor data into a daily activity journal [10] - capturing time spent in daily activities such as cooking, cleaning, exercising, and many more - supporting actionable lifestyle insights [23].

## 8 Limitations and future work

While WatchHAR demonstrates significant advancements in on-device human activity recognition, our approach has several limitations that present opportunities for future research.

First, watchOS offers limited access to microphone sampling rate adjustments and fails to provide fine-grained battery usage metrics for third-party apps. In practice, this means we cannot precisely adjust the sampling frequency based on energy availability. Addressing these constraints may require kernel-level modifications or leveraging the latest hardware releases, which might include more efficient microphone interfaces or battery optimization APIs.

In addition, although the current WatchHAR model is effective, it may not be fully optimized for the computational constraints of smartwatch hardware. Users with older devices may experience slower inference and increased battery drain due to limited processing power, while newer models often include hardware accelerators – such as Neural Processing Units (NPUs) in recent Apple devices – that can significantly improve energy efficiency when properly leveraged. In practical scenarios, energy efficiency could be improved by enabling the model to reduce sampling rates, generate fewer predictions, or skip sensor readings during periods of low activity, such as sitting or resting.

Finally – and most importantly – our current evaluation lacks a longitudinal study to assess the model's performance over extended periods and in diverse real-world situations. Similar to prior works, our accuracies see a severe performance degradation going from semi-controlled datasets to in-the-wild datasets. Since the model can now run on-device, future work should plan passive data collection studies over multiple days—or even weeks or months. A user study focused on long-term use would not only validate the system's sustained accuracy but also reveal new opportunities for dataset collection, applications, and personalization.

## 9 Safe and Responsible Innovation Statement

WatchHAR performs all inference on-device, so raw audio and IMU stay on the watch, eliminating server-side leakage risks. We train and evaluate solely on previously published, consented datasets (SAMoSA, Semi-Naturalistic, In-the-Wild) and will release code, weights, and evaluation scripts to foster reproducibility. Because dataset demographics are limited, we report per-dataset results to reveal potential bias and encourage future work on broader populations. Public releases must preserve our privacy-by-design constraint: no data storage or export for secondary purposes. We see no foreseeable harms but will address any reports of misuse promptly.

## 10 Conclusion

WatchHAR represents a significant leap forward in on-device human activity recognition on smartwatches for activities of daily living. By successfully implementing a multimodal system that processes both IMU and audio data entirely on-device with a novel end-to-end differentiable preprocessing plus inference pipeline, we have addressed key challenges in privacy, latency, and power efficiency that have long hindered the widespread adoption of continuous activity tracking. Our system's ability to maintain high accuracy across a diverse range of activities while operating in real-time on commodity smartwatch hardware demonstrates the viability of edge-based HAR solutions. Our implementation and demo application are openly available at https://github.com/SPICExLAB/WatchHAR, paving the way for further research and development in this field, potentially leading to new applications in health monitoring, context-aware computing, and personal analytics.

## Acknowledgments

We thank Chenfeng (Jesse) Gao for helping film the video and providing feedback on our figures. Vasco Xu's and Henry Hoffmann's work on this project is supported by NSF (CCF-2119184 CNS-2313190 CCF-1822949 CNS-1956180).

## References

[1] Alireza Abedin, Mahsa Ehsanpour, Qinfeng Shi, Hamid Rezatofighi, and Damith C. Ranasinghe. 2021. Attend and Discriminate: Beyond the State-of-the-Art for Human Activity Recognition Using Wearable Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 1 (mar 2021), 22 pages. https://doi.org/10.1145/3448083

[2] Karan Ahuja. 2024. Practical and Rich User Digitization. *arXiv preprint arXiv:2403.00153* (2024).

[3] Karan Ahuja, Eric Whitmire, Joseph Greer, and Wolf Kienzle. 2022. ActivityPoser: Activity driven Full-Body Pose Estimation from Sparse IMU Configurations. In *Proceedings of the 2022 ACM Symposium on Spatial User Interaction* (Online, CA, USA) *(SUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 19, 2 pages. https://doi.org/10.1145/3565970.3567687

[4] Sayma Akther, Nazir Saleheen, Mithun Saha, Vivek Shetty, and Santosh Kumar. 2021. mteeth: Identifying brushing teeth surfaces using wrist-worn inertial sensors. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 2 (2021), 1–25.

[5] Apple Inc. 2024. CoreML. https://developer.apple.com/documentation/coreml/. [Software library].

[6] Riku Arakawa, Hiromu Yakura, Vimal Mollyn, Suzanne Nie, Emma Russell, Dustin P. DeMeo, Haarika A. Reddy, Alexander K. Maytin, Bryan T. Carroll, Jill Fain Lehman, and Mayank Goel. 2023. PrISM-Tracker: A Framework for Multimodal Procedure Tracking Using Wearable Sensors and State Transition Information with User-Driven Handling of Errors and Uncertainty. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 156 (jan 2023), 27 pages. https://doi.org/10.1145/3569504

[7] John Arevalo, Thamar Solorio, Manuel Montes y Gómez, and Fabio A. González. 2017. Gated Multimodal Units for Information Fusion. arXiv:1702.01992 [stat.ML] https://arxiv.org/abs/1702.01992

[8] Sarnab Bhattacharya, Rebecca Adaimi, and Edison Thomaz. 2022. Leveraging Sound and Wrist Motion to Detect Activities of Daily Living with Commodity Smartwatches. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 42 (jul 2022), 28 pages. https://doi.org/10.1145/3534582

[9] Gino Brunner, Darya Melnyk, Birkir Sigfússon, and Roger Wattenhofer. 2019. Swimming style recognition and lap counting using a smartwatch and deep learning. In *Proceedings of the 2019 ACM International Symposium on Wearable Computers* (London, United Kingdom) *(ISWC '19)*. Association for Computing Machinery, New York, NY, USA, 23–31. https://doi.org/10.1145/3341163.3347719

[10] Lora E. Burke, Jing Wang, and Mary Ann Sevick. 2011. Self-Monitoring in Weight Loss: A Systematic Review of the Literature. *Journal of the American Dietetic Association* 111, 1 (2011), 92–102. https://doi.org/10.1016/j.jada.2010.10.008

[11] K. W. Cheuk, H. Anderson, K. Agres, and D. Herremans. 2020. nnAudio: An on-the-Fly GPU Audio to Spectrogram Conversion Toolbox Using 1D Convolutional Neural Networks. *IEEE Access* 8 (2020), 161981–162003. https://doi.org/10.1109/ACCESS.2020.3019084

[12] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1251–1258.

[13] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 776–780. https://doi.org/10.1109/ICASSP.2017.7952261

[14] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1314–1324.

[15] Jiha Kim, Younho Nam, Jungeun Lee, Young-Joo Suh, and Inseok Hwang. 2023. ProxiFit: Proximity Magnetic Sensing Using a Single Commodity Mobile toward Holistic Weight Exercise Monitoring. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–32.

[16] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. arXiv:1912.10211 [cs.SD] https://arxiv.org/abs/1912.10211

[17] Utkarsh Kunwar, Sheetal Borar, Moritz Berghofer, Julia Kylmälä, Ilhan Aslan, Luis A Leiva, and Antti Oulasvirta. 2022. Robust and deployable gesture recognition for smartwatches. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 277–291.

[18] Hong Li, Shishir Chawla, Richard Li, Sumeet Jain, Gregory D. Abowd, Thad Starner, Cheng Zhang, and Thomas Plötz. 2018. Wristwash: towards automatic handwashing assessment using a wrist-worn device. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers* (Singapore, Singapore) *(ISWC '18)*. Association for Computing Machinery, New York, NY, USA, 132–139. https://doi.org/10.1145/3267242.3267247

[19] Vimal Mollyn, Karan Ahuja, Dhruv Verma, Chris Harrison, and Mayank Goel. 2022. SAMoSA: Sensing Activities with Motion and Subsampled Audio. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 132 (sep 2022), 19 pages. https://doi.org/10.1145/3550284

[20] Dan Morris, T Scott Saponas, Andrew Guillory, and Ilya Kelner. 2014. RecoFit: using a wearable sensor to find, recognize, and count repetitive exercises. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3225–3234.

[21] Shuai Shao, Yu Guan, Bing Zhai, Paolo Missier, and Thomas Plötz. 2023. ConvBoost: Boosting ConvNets for Sensor-based Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 2 (2023), 75. https://doi.org/10.1145/3596234

[22] Muhammad Shoaib, Stephan Bosch, Hans Scholten, Paul JM Havinga, and Ozlem Durmaz Incel. 2015. Towards detection of bad habits by fusing smartphone and smartwatch sensors. In *2015 IEEE international conference on pervasive computing and communication workshops (PerCom Workshops)*. IEEE, 591–596.

[23] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. 2017. Recognizing Detailed Human Context in the Wild from Smartphones and Smartwatches. *IEEE Pervasive Computing* 16, 4 (2017), 62–74. https://doi.org/10.1109/MPRV.2017.3971131

[24] Jamie A Ward, Paul Lukowicz, Gerhard Troster, and Thad E Starner. 2006. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE transactions on pattern analysis and machine intelligence* 28, 10 (2006), 1553–1567.

[25] Vasco Xu, Chenfeng Gao, Henry Hoffmann, and Karan Ahuja. 2024. Mobileposer: Real-time full-body pose estimation and 3d human translation from imus in mobile consumer devices. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–11.

[26] Weitao Xu, Huanqi Yang, Jiongzhang Chen, Chengwen Luo, Jia Zhang, Yuliang Zhao, and Wen Jung Li. 2024. WashRing: An Energy-Efficient and Highly Accurate Handwashing Monitoring System via Smart Ring. *IEEE Transactions on Mobile Computing* 23, 1 (2024), 971–984. https://doi.org/10.1109/TMC.2022.3227299

[27] Cheng Zhang, AbdelKareem Bedri, Gabriel Reyes, Bailey Bercik, Omer T. Inan, Thad E. Starner, and Gregory D. Abowd. 2016. TapSkin: Recognizing On-Skin Input for Smartwatches. In *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces* (Niagara Falls, Ontario, Canada) *(ISS '16)*. Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/2992154.2992187

[28] Shibo Zhang, Ebrahim Nemati, Minh Dinh, Nathan Folkman, Tousif Ahmed, Mahbubur Rahman, Jilong Kuang, Nabil Alshurafa, and Alex Gao. 2022. Coughtrigger: Earbuds IMU Based Cough Detection Activator Using An Energy-Efficient Sensitivity-Prioritized Time Series Classifier. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. https://doi.org/10.1109/ICASSP43922.2022.9746334