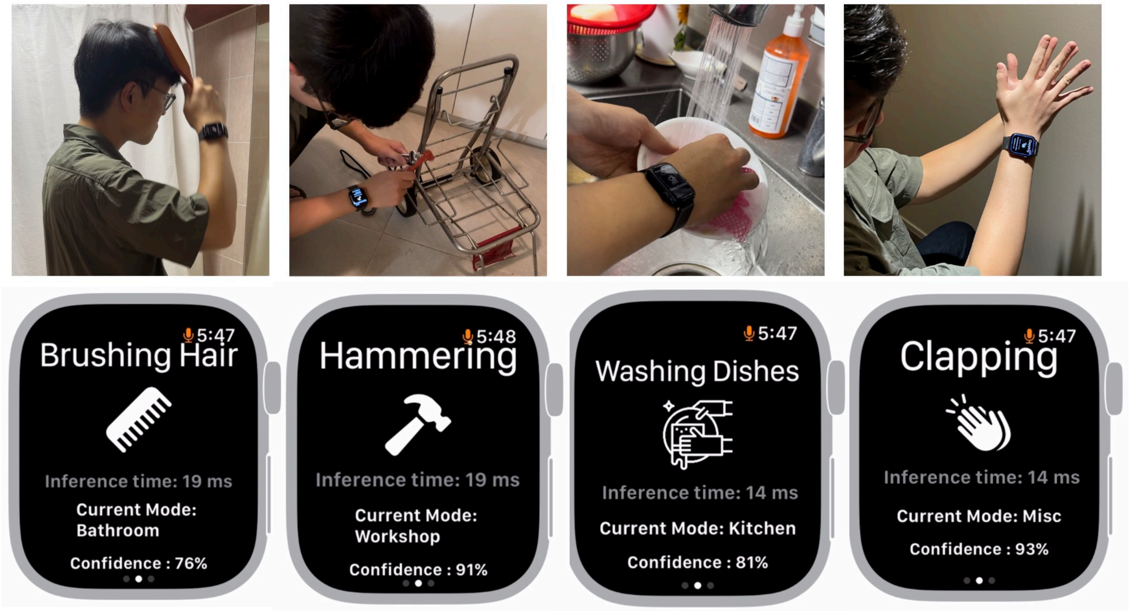
EdgeHAR: Real-time On-device Human Activity Recognition System for Smartwatches

ANONYMOUS AUTHOR(S)



Brushing Hair
5:47
Inference time: 19 ms
Current Mode: Bathroom
Confidence : 76%

Hammering
5:48
Inference time: 19 ms
Current Mode: Workshop
Confidence : 91%

Washing Dishes
5:47
Inference time: 14 ms
Current Mode: Kitchen
Confidence : 81%

Clapping
5:47
Inference time: 14 ms
Current Mode: Misc
Confidence : 93%

Fig. 1. Our EdgeHAR System Running on Apple Watch Series 7 with four different contexts.

Despite advances in practical and multimodal human activity recognition (HAR), a system that runs entirely on smartwatches in unconstrained environments remains elusive. We present EdgeHAR, 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, EdgeHAR achieves compounding performance gains. We introduce a novel architecture that unifies sensor data pre-processing and inference into an end-to-end trainable module, significantly accelerating performance by 25x while maintaining over 90% accuracy on 25+ activity classes. EdgeHAR outperforms state-of-the-art models in terms of accuracy while running on the smartwatch directly at 30 Hz for multimodal activity classification and 40 Hz for activity event detection. This research advances edge-based activity recognition, realizing smartwatches' potential as standalone, minimally-invasive activity tracking devices.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; • **Security and privacy** → *Privacy protections*; • **Applied computing** → Health informatics.

Additional Key Words and Phrases: Smartwatches, On-device processing, Real-time monitoring, Human activity recognition, privacy and security in mobile devices

ACM Reference Format:

Anonymous Author(s). 2018. EdgeHAR: Real-time On-device Human Activity Recognition System for Smartwatches. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Human Activity Recognition (HAR) has become a cornerstone of ubiquitous computing, with applications ranging from health monitoring to context-aware services. While significant strides have been made in developing accurate and robust HAR systems, 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 systems rely on external data processing, raising concerns about privacy, latency, and the need for constant connectivity.

EdgeHAR addresses these challenges by introducing a novel HAR system that operates entirely on smartwatches, leveraging both audio and inertial data. By eliminating the need for external data processing, EdgeHAR 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. EdgeHAR's key innovation lies in its end-to-end trainable preprocessing module that computes a Short-Time Fourier Transform and approximates a mel-filter bank as a 1D convolutional operation, thus enabling efficient GPU-based processing directly on the smartwatch.

Through careful optimization of each component in the pipeline, EdgeHAR achieves compounding performance gains. The system outperforms state-of-the-art models in terms of accuracy while running on the smartwatch at 30 Hz for multimodal activity classification and 40 Hz for activity event detection. EdgeHAR's novel architecture unifies sensor data pre-processing and inference into a single, trainable module, significantly accelerating performance by 25x while maintaining over 90% accuracy in more than 25 activity classes. These advancements demonstrate EdgeHAR's potential to revolutionize edge-based activity recognition, realizing the full potential of smartwatches as standalone, minimally-invasive activity tracking devices.

2 RELATED WORK

Human Activity Recognition (HAR) has seen significant advancements in recent years, particularly in the realm of wearable technology. A wide range of wearable devices have been explored for HAR, including wrist-worn sensors [1, 12, 14], smart rings [16], earbuds [18], and smartwatches [4, 5, 11, 13, 17]. These devices have proven effective in identifying various activities, from fitness exercises to daily tasks.

Smartwatches, with their array of sensors including IMUs and microphones, have emerged as particularly powerful data providers for consumer HAR systems. Recent works have demonstrated the potential of multimodal approaches in HAR using off-the-shelf commodity smartwatches. SAMoSA [13] and Bhattacharya et al. [4] showcased the benefits of combining audio and IMU data. SAMoSA achieved 92.2% accuracy across various contexts using 1kHz audio and 50Hz IMU data, proving that even lower-sampled audio can significantly enhance activity sensing while preserving privacy. They also introduced an IMU-based activity detector to optimize the use of power-hungry microphones. Bhattacharya et al. explored various multimodal sensor fusion techniques, and showcased performance in both controlled and in-the-wild scenarios. Despite these advancements, most systems treat wearables merely as data providers, offloading processing to smartphones or desktops [4, 13, 16, 17]. 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 [10] exercise monitoring system using natural magnetism in exercise equipment, and Zhang et al.'s [18] cough detection system that uses IMU sensor values to activate cough detection. Kunwar et al. [11] 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.

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. EdgeHAR 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 SYSTEM ARCHITECTURE

Our system architecture is designed to balance computational load and power consumption for real-time activity recognition on smartwatches. The system consists of two main components: an IMU Activity Detector and a Multimodal Activity Classifier.

To ensure our system runs efficiently on smartwatches, we implement several optimization techniques on the trained models (described below). We convert all models to the Apple CoreML format[2] for optimized execution on the Apple Watch hardware. We apply 16-bit float quantization, which reduces model size and improves inference speed with negligible impact on accuracy. We carefully tune the window sizes and hop lengths for both the activity 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 Activity Detector

We employ a lightweight IMU-based activity detector to initiate the more resource-intensive multimodal classifier. Our detector uses a 1D depthwise Convolutional Neural Network (CNN) architecture [7], 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 activity detection. To ensure rapid detection of activity onset while maintaining robustness against false positives from small motions, we create a new 3-second window every 20 ms (the time between two IMU samples at 50 Hz). The CNN model processes each window, outputting a binary classification (activity detected or not). We then apply a 2-second moving average to the model outputs, which helps filter out spurious detections.

Our system follows a two-stage detection process. The IMU Activity Detector continuously monitors for potential activities. When activity is detected, we activate the Multimodal Activity Classifier. If the Classifier confirms an activity, we continue processing; otherwise, we return to the IMU Activity Detector stage. This approach optimizes power consumption by only activating the more resource-intensive classifier when necessary.

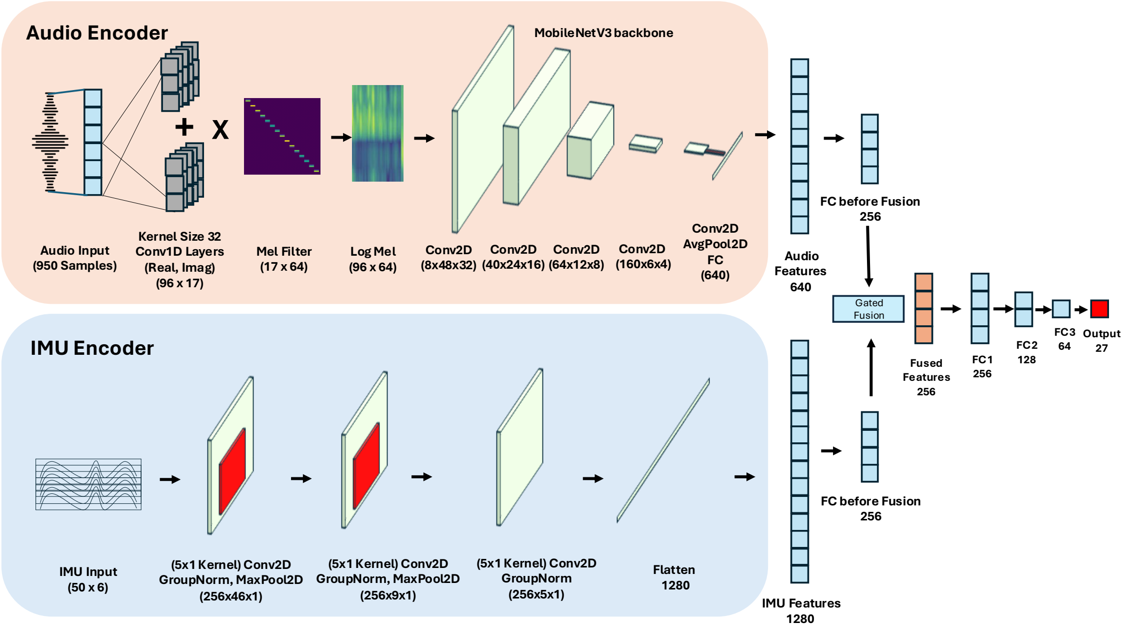


Fig. 2. Overall architecture of EdgeHAR's Multimodal Activity Classifier

3.2 Multi-modal Activity Classification

Our Multimodal Activity Classifier (Figure 2) processes both IMU and audio data to achieve high-accuracy activity recognition. As audio adds a significant compute burden, we use shorter window sizes to enable faster processing and reduce latency. Both IMU and audio data use 1-second windows. Like the activity detector, we use a 20 ms hop length for both IMU and audio data, allowing for fine-grained temporal resolution in our classifications.

Inspired by the nnAudio framework [6], we implement an end-to-end trainable audio preprocessing module directly within our neural network. Our architecture consists of three main components: a Short-Time Fourier Transform (STFT) implemented using a 1D convolutional layer, a mel-filter bank realized 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 the filters for HAR might be very different from speech recognition tasks which mel filters are tuned for. Lastly, the amplitude-to-DB conversion uses a logarithmic activation function to produce a spectogram which is passed to the Audio feature encoder.

For audio encoder, we utilize a MobileNetV3[9] backbone pretrained on the AudioSet dataset[8]. 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[15]. This model uses a 2D CNN structure optimized for efficient processing of multivariate time series data such as IMU signals. Key features of the ConvBoost model include depthwise separable convolutions for efficient feature extraction, shortcut connections inspired by ResNet architectures, and channel attention mechanisms to focus on the most relevant sensor axes.

To effectively combine information from both IMU and audio modalities, we implement a Gated Fusion mechanism[3] rather than simple feature concatenation. This approach allows the model to dynamically weigh the importance of each modality based on the input, potentially improving performance on activities where one modality may be more informative than the other.

4 DATASETS

To evaluate EdgeHAR we make use of the following existing smartwatch datasets to benchmark our results.

SAMoSA Dataset[13]: This was collected from 20 participants (mean age 23.3, all right-handed) across 60 diverse environments. Data was captured using a Fossil Gen 5 smartwatch running Google Android wearOS 2.23, which collected synchronized streams of 9-axis IMU data (accelerometer, gyroscope, and orientation) at 50 Hz and uncompressed audio at 16 kHz. The audio is later post-processed to 1kHz to make it privacy-preserving and hence removes any speech intelligibility from it. The dataset covers 26 activities across four contexts: kitchen, bathroom, workshop, and miscellaneous. Each participant performed each activity 3 times within each context, resulting in a total of 14.2 hours of data. This includes 5.9 hours of labeled activity data and 8.3 hours of in-transition "Other" data. The activities were performed in participants' homes with their own appliances and tools, incorporating associated contextual background noise profiles. The dataset presents evaluation protocols for both activity detection and classification.

Semi-Naturalistic Dataset[4]: This was collected from 15 participants (9 females and 6 males) with ages ranging from 23 to 64 (mean 43.6), representing diverse professions and socioeconomic backgrounds. Data was captured using a Fossil Gen 4 smartwatch running Android Wear OS 2.12, which collected synchronized streams of accelerometer, gyroscope, and microphone data. Inertial sensors were sampled at 50 Hz, while acoustic data was sampled at 22.05 kHz. The dataset covers 23 activities in total. The dataset presents evaluation protocols for only activity classification.

Data collection was conducted remotely via video calls due to social distancing regulations, with participants performing activities in their own homes. Each participant performed all 23 activities twice, once in each of two sessions, with a 15-minute break between sessions to introduce variability and test wearable placement sensitivity. Each activity lasted for a minimum of 30 seconds. To facilitate annotation, participants knocked on a surface to indicate the start and end of each activity. The entire data collection process was continuous within each session, capturing all activities and in-between movements. Data annotation was manually carried out by researchers using both the sensor data and the recorded video calls. The total duration of the dataset is not explicitly mentioned, but given the study design, it likely contains several hours of labeled activity data across various home environments.

5 RESULTS

We evaluate EdgeHAR'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. We found that converting 32-bit models to 16-bit didn't affect model predictions or accuracy. Performance evaluations were conducted on an M1 Max Macbook (2021) and Apple Watch Series 7.

Table 1. Event Detection Model Performance Comparison between SAMoSA and EdgeHAR.

Method	F1 Score (%)	Watch FPS	Onset Latency (sec)	Offset Latency (sec)
SAMoSA [IMU @ 50 Hz]	88.0	29.8	0.62	0.16
EdgeHAR [IMU @ 50Hz]	**93.5**	**40.6**	**0.27**	**0.07**

5.1 Activity Detector

We compared our Depthwise CNN1D Activity Detection Model with SAMoSA's random forest event detection model. The performance metrics are shown in Table 1. F1 Score represents the balanced accuracy of the model in detecting activity events. Watch FPS indicates the number of frames per second the model can process on the Apple Watch Series 7. Onset Latency measures the delay between the physical start of an activity and its detection by the model, while Offset Latency measures the delay in detecting the end of an activity.

EdgeHAR's activity detector achieves a higher F1 Score (93.5% vs 88.0%), faster processing speed on the watch (40.6 FPS vs 29.8 FPS), and lower onset (0.27s vs 0.62s) and offset (0.07s vs 0.16s) latencies compared to SAMoSA. This improvement is partly due to our model's ability to generate predictions every 20 ms, compared to SAMoSA's 200 ms interval. To address potential mispredictions due to data skewness, we implemented a moving average smoothing mechanism. This improved our F1 Score from 92.5% to 93.5% without significant computational overhead.

The SAMoSA model was transferred to the watch by implementing its random forest algorithm and feature computation in Swift. However, we found that computing features for SAMoSA's model took about 28.76ms leading to significant delays, while our CNN1D model only requires normalization of raw IMU data, which takes around 3ms.

5.2 Multimodal Activity Classification

Table 2. Multimodal Activity Classification across different approaches, datasets and Audio Sampling Rates

Model Name	Dataset	Sampling Rate (kHz)	Context-wise (%)	LOPO (%)	P-LOPO (%)	Inference Time (ms)	Watch FPS	Params (M)
SAMoSA	SAMoSA	1	92.2	-	-	38.98	-	21.41
Bhattacharya et al.	Semi-Naturalistic	22.05	-	89.7	**94.3**	109.8	-	80.689
EdgeHAR	SAMoSA	1	**92.34**	-	-	**0.45**	**31.5**	**2.360**
EdgeHAR	Semi-Naturalistic	22.05	-	**90.4**	93.8	9.42	10.98	6.061

For activity classification, we compared EdgeHAR with two main prior works: SAMoSA and Bhattacharya et al. The performance metrics are shown in Table 2.

For SAMoSA, we used their model with 1kHz audio and 50Hz motion data, as these were their primary results. Context-wise accuracy for SAMoSA refers to the average classification accuracy across four contexts (kitchen, bathroom, workshop, and miscellaneous). For the Semi-Naturalistic dataset from Bhattacharya et al., we used 22kHz audio to match their setup. LOPO (Leave-One-Participant-Out) evaluation tests the model's generalization to unseen participants. P-LOPO (Personalized-LOPO) includes some data from the test participant in training, simulating a partially personalized model. Inference Time represents the end-to-end processing time on the M1 Mac, including pre-processing, model inference, and any post-processing. Watch FPS indicates the real-time performance running on the Apple Watch Series 7. Params (M) shows the number of parameters in millions for each model.

EdgeHAR achieves comparable or better accuracy across datasets while significantly reducing model size and inference time. On the SAMoSA dataset, EdgeHAR slightly outperforms SAMoSA in context-wise accuracy (92.34% vs

92.2%) while reducing inference time by a factor of 86 (0.45ms vs 38.98ms) and model size by a factor of 9 (2.360M vs 21.41M parameters).

On the Semi-Naturalistic dataset, EdgeHAR achieves slightly better LOPO accuracy (90.4% vs 89.7%) and comparable P-LOPO accuracy (93.8% vs 94.3%) compared to Bhattacharya et al. EdgeHAR's model is significantly smaller (6.061M vs 80.689M parameters) and faster (9.42ms vs 109.8ms inference time) than Bhattacharya et al.'s approach.

For per activity confusion matrix, refer to the Appendix where we include results on SAMOSA for each activity context in Figure 4. Similarly, the activity confusion matrix on the Semi-Naturalistic dataset is depicted in Figure 5.

6 LIVE DEMO

To demonstrate the real-world applicability and performance of EdgeHAR, we developed a fully functional demo application that runs in real-time on the Apple Watch (refer to Video Figure). The application and all associated code will be open-sourced upon publication, providing researchers and developers with a valuable resource for further exploration and development in the field of on-device human activity recognition.

Figure 3 illustrates the workflow and user interface of our demo application. The system operates in two main modes: activity detection and activity classification. Initially, the system runs in detection mode, utilizing only the IMU sensor to conserve power. This is evident in the first frame, where the watch display shows "Detecting Activity" with an inference time of 9 ms, highlighting the efficiency of our IMU-based detector. When an activity is detected (second frame), the system activates the microphone, indicated by the question mark icon. This triggers the multimodal classifier, which combines IMU and audio data for more accurate activity recognition. Importantly, the audio is sampled at 1 kHz, preserving privacy in line with the approach used in SAMoSA. The third frame shows successful activity classification, identifying "Coughing" with high confidence (91%). Note the inference time of 19 ms for this multimodal classification, demonstrating the system's ability to perform complex analysis in near real-time on the watch. After the activity ends, the system returns to the IMU-only detection mode (fourth frame), again showing the 9 ms inference time for the detector.

Another key feature of our demo is the real-time visualization of sensor data directly on the watch. The bottom row of each frame shows live audio spectrograms and accelerometer graphs, providing immediate feedback and insight into the data being processed for debugging.

7 LIMITATIONS AND FUTURE WORK

While EdgeHAR demonstrates significant advancements in on-device human activity recognition, our approach has several limitations that present opportunities for future research. Primarily, we were unable to conduct a comprehensive power consumption analysis due to hardware and software constraints of the Apple Watch platform. The inability to adjust audio sampling rates and access app-specific battery consumption information through Apple's APIs limited our ability to directly measure the efficiency of our model and architecture in real-world scenarios. Although we optimized battery usage by restricting microphone activation, further investigation is needed to fully understand the system's energy impact in daily use. Future work should explore testing on platforms that allow for more granular control over sensor sampling rates and provide detailed power consumption metrics.

Another limitation is the scope of our device testing. While we focused on the Apple Watch due to its neural accelerator capabilities, we recognize the need to evaluate EdgeHAR's performance on a broader range of devices, particularly those with less powerful hardware. Expanding our testing to include Android smartwatches and other

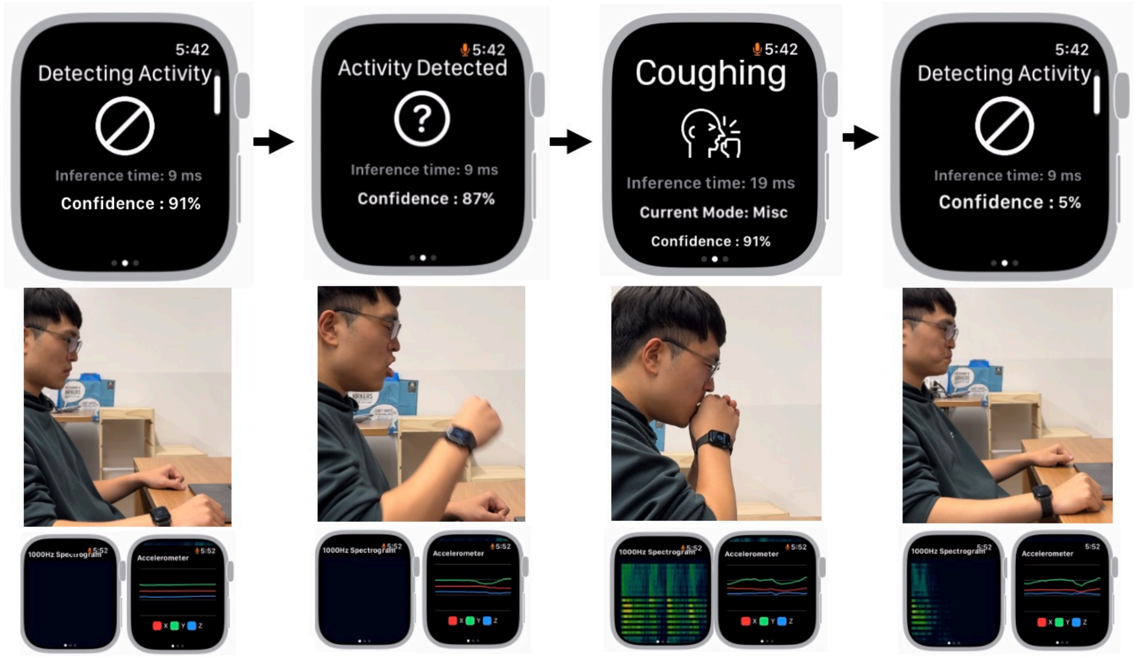


Fig. 3. Our Application Activity Detection Workflow and User Interface. Microphone is activated only after activity is detected. Also our Application gives real-time visualization of Audio Spectrogram and Accelerometer graph.

wearable platforms would provide valuable insights into the system's scalability and adaptability across different hardware configurations.

Additionally, our current evaluation lacks a longitudinal study to assess the model's performance over extended periods and in diverse real-world situations. Given we can now run on-device, future work should plan out passive data collection studies over the course of multiple days, if not weeks or months. A user study focused on long-term use would not only validate the system's sustained accuracy but also reveal interesting opportunities for new applications and user personalization without compromising generalizability.

8 CONCLUSION

EdgeHAR represents a significant leap forward in on-device human activity recognition for smartwatches. By successfully implementing a multimodal system that processes both IMU and audio data entirely on-device, 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. The open-sourcing of our implementation and demo application (upon publication) paves the way for further research and development in this field, potentially leading to new applications in health monitoring, context-aware computing, and personal analytics.

REFERENCES

[1] 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.

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

[3] 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

[4] 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

[5] 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

[6] 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

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

[8] 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

[9] 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.

[10] 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.

[11] 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.

[12] 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

[13] 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

[14] 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.

[15] 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

[16] 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

[17] 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

[18] 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

A CONFUSION MATRICES OF EDGEHAR ON SAMOSA AND SEMI-NATURALISTIC DATASET

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

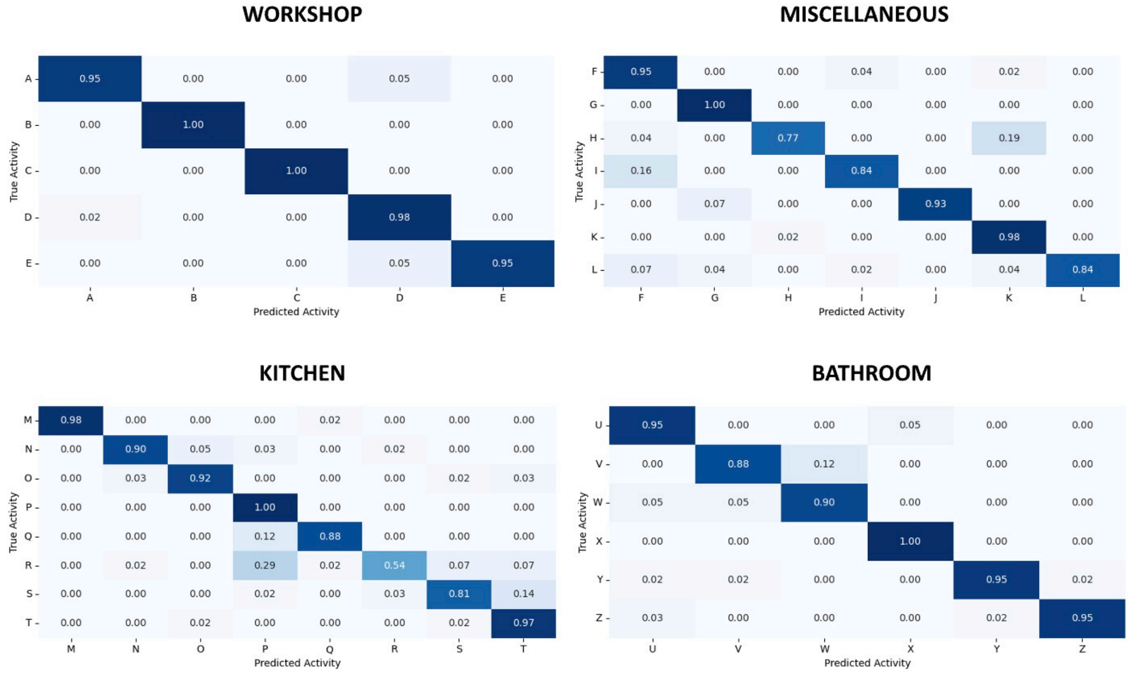
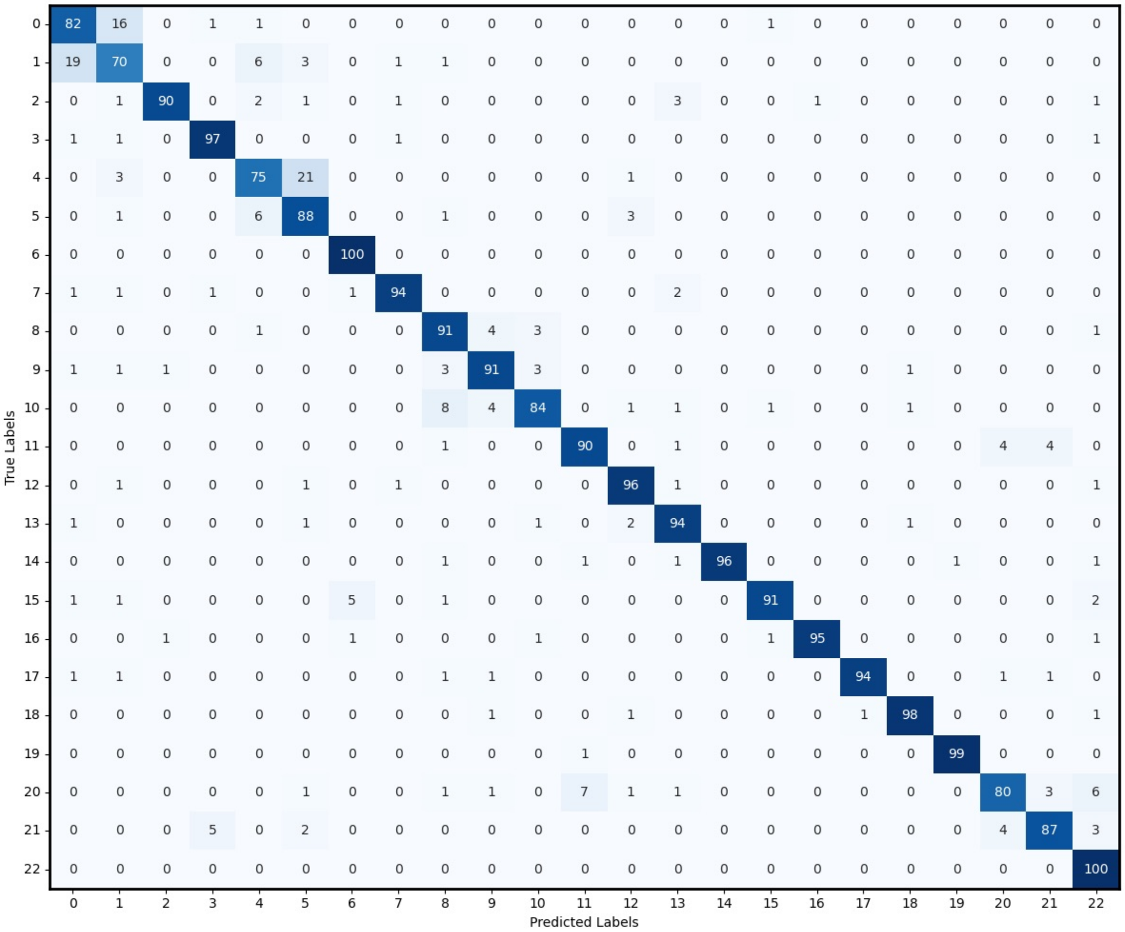


Fig. 4. Context-wise Confusion Matrix of our multimodal model from SAMoSA Dataset. Labels in alphabetic order as follows, A: Drill in use, B: Hammering, C: Sanding, D: Screwing, E: Vacuum in use, F: Alarm clock, G: Clapping, H: Coughing, I: Drinking, J: Knocking, K: Laughing, L: Scratching, M: Blender in use, N: Chopping, O: Grating, P: Microwave, Q: Pouring pitcher, R: Twisting jar, S: Washing Utensils, T: Wiping with rag, U: Brushing hair, V: Hair dryer in use, W: Shaver in use, X: Toilet flushing, Y: Toothbrushing, Z: Washing hands.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
0	82	16	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
1	19	70	0	0	6	3	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	1	90	0	2	1	0	1	0	0	0	0	0	3	0	0	1	0	0	0	0	0	1
3	1	1	0	97	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	0	3	0	0	75	21	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
5	0	1	0	0	6	88	0	0	1	0	0	0	3	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	1	0	1	0	0	1	94	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	91	4	3	0	0	0	0	0	0	0	0	0	0	0	1
9	1	1	1	0	0	0	0	0	3	91	3	0	0	0	0	0	0	0	1	0	0	0	0
10	0	0	0	0	0	0	0	0	8	4	84	0	1	1	0	1	0	0	1	0	0	0	0
11	0	0	0	0	0	0	0	0	1	0	0	90	0	1	0	0	0	0	0	0	4	4	0
12	0	1	0	0	0	1	0	1	0	0	0	0	96	1	0	0	0	0	0	0	0	0	1
13	1	0	0	0	0	1	0	0	0	0	1	0	2	94	0	0	0	0	1	0	0	0	0
14	0	0	0	0	0	0	0	0	1	0	0	1	0	1	96	0	0	0	0	1	0	0	1
15	1	1	0	0	0	0	5	0	1	0	0	0	0	0	0	91	0	0	0	0	0	0	2
16	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	95	0	0	0	0	0	1
17	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	94	0	0	1	1	0
18	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	98	0	0	0	1
19	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	99	0	0	0
20	0	0	0	0	0	1	0	0	1	1	0	7	1	1	0	0	0	0	0	0	80	3	6
21	0	0	0	5	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	87	3
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

Fig. 5. Confusion Matrix of EdgeHAR activity classification on Bhattacharya et al.[4] semi-naturalistic dataset. Labels are numbered as follows, 1: Writing, 2: Drawing, 3: Cutting paper, 4: Typing on keyboard, 5: Typing on phone, 6: Browsing on phone, 7: Clapping, 8: Shuffling cards, 9: Scratching, 10: Wiping table, 11: Brushing hair, 12: Washing hands, 13: Drinking, 14: Eating snacks, 15: Brushing teeth, 16: Chopping, 17: Grating, 18: Frying, 19: Sweeping, 20: Vacuuming, 21: Washing dishes, 22: Filling water, 23: Using microwave.