

μ Dpad: A Large-Scale Multimodal PPG and IMU Dataset for Wrist-Worn Microgesture Recognition

Lars Hauptmann
Department of Computer Science
ETH Zürich
Zürich, Switzerland
lars.hauptmann@inf.ethz.ch

Dominik Hollidt
Department of Computer Science
ETH Zürich
Zürich, Switzerland
dominik.hollidt@inf.ethz.ch

Xintong Liu
Department of Computer Science
ETH Zürich
Zürich, Switzerland
xintong.liu@inf.ethz.ch

Manuel Meier
Department of Computer Science
ETH Zürich
Zürich, Switzerland
manuel.meier@inf.ethz.ch

Christian Holz
Department of Computer Science
ETH Zürich
Zürich, Switzerland
chrisitan.holz@inf.ethz.ch

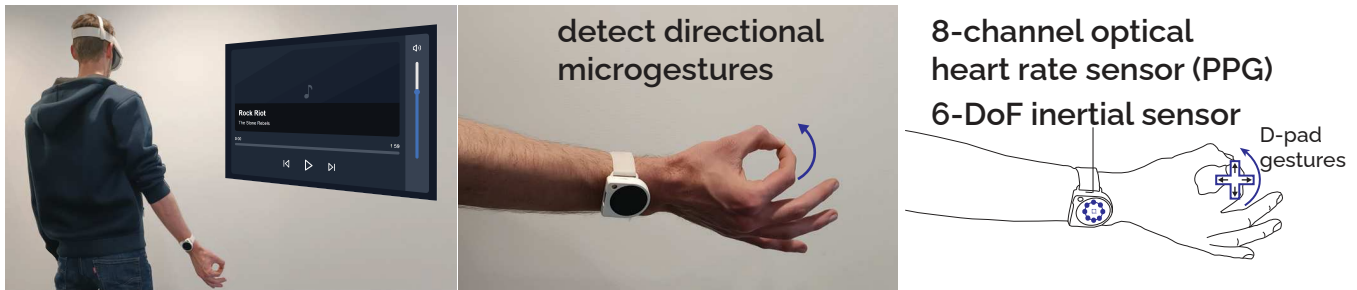


Figure 1: μ Dpad is a wrist-worn device that detects one-hand directional thumb microgestures for microinteraction-style input: thumb swipes in four directions (akin to a D-pad) as well as pinch close/open and wrist rotation for continuous control during pinching. μ Dpad’s sensing system integrates an inertial measurement unit (IMU) and an 8-channel spatial optical sensor, reflecting an emerging smartwatch sensing configuration that we use here to collect a multimodal dataset and establish baseline microgesture recognition performance.

Abstract

Gestures provide a natural interaction modality for mobile and wearable devices. However, most prior work focuses on macro gestures involving large wrist or arm movements, which are fatiguing and socially conspicuous. In contrast, microgestures, subtle, low-effort finger movements that can be performed one-handed and without visual attention, offer a promising alternative but remain significantly harder to sense and recognize. In this work, we present a large-scale multimodal dataset for wrist-based microgesture recognition, collected with an integrated wearable prototype that combines an IMU with an 8-channel spatial PPG sensor configuration similar to one emerging in commercial fitness watches. Because such devices typically do not expose raw optical signals, our goal is to provide the community with a resource for studying whether spatial PPG can capture the minute tissue deformations associated with microgestures. Our dataset comprises 25 hours of

synchronized IMU and PPG recordings from 66 participants, including 9 hours of aligned video annotations, and captures substantial variability in users, hand poses, and execution styles. To establish reference performance, we benchmark a range of recognition pipelines and present a simple multimodal baseline. This baseline already shows that adding spatial PPG improves microgesture detection over IMU-only sensing, reaching 83.7% accuracy with a 3.7% false-positive rate, while also underscoring that microgesture recognition remains substantially harder than recognizing coarser macro-gesture vocabularies. By releasing the dataset, we aim to enable more substantive algorithmic advances toward practical, low-effort, and socially acceptable microgesture interaction on future wrist-worn devices. Project page: siplab.org/projects/microDPad.

CCS Concepts

• **Human-centered computing** → **Gestural input**; • **Hardware** → **Sensor devices and platforms**.

Keywords

Microinteractions, microgestures, smartwatch input, wrist-worn sensing, IMU, photoplethysmography (PPG), online decoding.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

ICMI '26, Napoli, Italy

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2318-6/2026/10

<https://doi.org/10.1145/3776574.3831197>

ACM Reference Format:

Lars Hauptmann, Dominik Hollidt, Xintong Liu, Manuel Meier, and Christian Holz. 2026. μ Dpad: A Large-Scale Multimodal PPG and IMU Dataset for Wrist-Worn Microgesture Recognition. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26), October 05–09, 2026, Napoli, Italy*. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3776574.3831197>

1 Introduction

Hand gesture recognition is a key technology for enabling intuitive interaction in mobile and wearable computing [7, 10, 11]. Beyond explicit touch input, wearable sensors support interaction through hand and finger gestures [9, 29, 30]. Prior work demonstrated gesture input for domains such as robotic teleoperation [15, 16, 37], communication [28, 71], and user interface control [12]. However, much of this literature focuses on macro gestures or static hand poses. Macro gestures require larger hand or finger movements, which can be fatiguing and visually noticeable in social settings.

In contrast, subtle *microgestures*—small, low-effort gestures—offer an appealing alternative [2, 63]. These quick, eyes-free finger movements can enable socially acceptable control for tasks such as media playback, notification triage, and lightweight UI navigation, but reliably sensing them remains difficult. Compared with macro gestures, microgestures are comparatively underexplored: prior systems often rely on specialized hardware, target narrow gesture vocabularies, or are evaluated on limited datasets.

A wide range of sensing modalities has been explored for gesture recognition, including cameras [44, 68], EMG [13, 24, 58], radar [34], and sonar [30]. Many of these approaches have been successful for larger macro gestures, but reliable detection of subtle microgestures remains unresolved. While some methods achieve strong performance, they often depend on specialized hardware or remain difficult to deploy robustly in everyday use. For example, camera-based methods are sensitive to occlusion, viewpoint changes, and the requirement that the hand remain within the field of view [18, 59].

Our goal is therefore to study microgesture recognition using sensing modalities that are already integrated into commodity wearable devices. Inertial measurement units (IMUs) and photoplethysmography (PPG) sensors are widely available in modern wearables and offer complementary sensing capabilities. IMUs capture global motion dynamics and have been widely used for gesture recognition, particularly for larger macro gestures [10, 29, 40, 66]. Still, their effectiveness for fine-grained microgestures is limited, as subtle finger motions are easily confounded by placement variability, motion artifacts, and sensor noise.

PPG sensors, commonly used for heart-rate monitoring, measure changes in blood volume and tissue deformation at the skin [3, 52]. While primarily used for physiological sensing, their sensitivity to local pressure and deformation patterns makes them a promising signal source for capturing finger motion. Prior work has begun to explore PPG for gesture recognition [13, 56, 70], demonstrating its feasibility. Yet, these approaches typically rely on single-channel PPG signals and focus on larger, more distinct gestures. In contrast, modern wearable devices often provide multi-channel PPG measurements, which capture spatially distributed deformation patterns and may offer richer information for resolving subtle microgestures.

Together, IMU and PPG provide complementary information, combining global motion cues with local deformation signals for microgesture recognition. Yet subtle wrist-based microgestures on commodity sensors remain insufficiently understood: most prior work targets macro gestures, relies on specialized hardware, or uses small-scale evaluations, and no established benchmark exists for combining IMU and multi-channel PPG for this task.

In this paper, we address this gap primarily by contributing a benchmark dataset for wrist-based microgestures, alongside a systematic empirical study and a simple real-time multimodal baseline. We introduce a large-scale dataset comprising synchronized 6-axis IMU and spatially distributed 8-channel PPG recordings from 66 participants, totaling 25 hours of sensor data and 9 hours of aligned video annotations, collected with an integrated wearable prototype that reflects an emerging sensor configuration in commercial fitness watches. Because such devices typically do not expose raw optical signals, we created this corpus to enable the community to study how multi-channel spatial PPG can contribute to subtle microgesture recognition. Using this dataset, we benchmark established methods, including feature-based models (e.g., SVM and Random Forest) and deep architectures (e.g., VGG, ResNet, U-Net, and Transformers), to analyze the impact of sensor modality, model design, and user variability. We further present μ Dpad as a straightforward multimodal baseline that combines deep learning-based event detection with Hidden Markov Model (HMM) temporal decoding for continuous recognition of directional thumb swipes and pinch open/close events. This baseline achieves 78.6% accuracy with a false-positive rate of 5.3% in online gesture detection.

We further demonstrate its practical applicability through interactive prototypes supporting both discrete input and continuous control. Despite these results, performance remains below that of systems targeting macro gestures, highlighting the intrinsic difficulty of microgesture recognition. We therefore position the released dataset and benchmark as the paper's main contribution, intended to support more substantive future advances in multimodal wearable microgesture recognition.

In summary, we make the following contributions:

- (1) **A large-scale multimodal microgesture dataset for commodity wearables.** We present 25 hours of synchronized wrist-worn sensor data from 66 participants, including 6-axis IMU and 8-channel spatial PPG recordings of directional swipe, pinch, and rotation microgestures across varied arm orientations. For 9 hours, we also provide synchronized RGB hand video.
- (2) **A systematic benchmark of sensor modalities and recognition models for microgestures.** Using this dataset, we compare classical and deep models for fine-grained microgesture recognition, including SVMs, Random Forests, VGG, ResNet, U-Net, and Transformers. We show that multi-channel spatial PPG complements IMU, especially for users where IMU-only recognition is less reliable.
- (3) **μ Dpad, real-time multimodal baseline.** We introduce a hybrid pipeline that fuses 6D IMU and 8-channel PPG signals for continuous recognition of directional thumb swipes and pinch open/close events. It combines a learning-based event detection with Hidden Markov Model temporal decoding for online robustness. The baseline reaches 83.7% accuracy, and with personalization improves accuracy to 88%.

2 Related Work

Our work explores gesture recognition systems that leverage sensor-based devices, focusing on swiping gestures as a form of directional microgestures. We target mobile and wearable microinteractions where input should be quick, low-effort, and robust to everyday contexts, and where camera-based sensing may be unavailable or undesirable.

Gesture recognition is the process of identifying and interpreting physical gestures, which are meaningful body movements such as motions of the fingers, hands, arms, head, or full body. These gestures convey information or facilitate interaction with the environment. [10] The field spans a wide range of applications, such as medical rehabilitation [48], sign language recognition [28], and mobile and wearable interaction. Immersive systems (e.g., VR) also make use of gesture-based input [4, 41], but they are not the primary focus of this paper.

2.1 Hand Gesture Recognition

Hand gesture recognition has become a prominent research area, leveraging a wide range of sensing modalities. Cameras excel in hand tracking but suffer from occlusions and processing typically requires more compute [19, 64]. Acoustic signals have also been shown to help for gesture recognition [20, 21] but could suffer in real-world applications with background noise. Other research used signals from infrared (IR) sensors [17, 26], electromyography (EMG) [24, 57, 67], pressure sensors [23], radar [34], stretch sensors [49], magnetic sensors [8, 43], bio-capacitive sensing [53], photoplethysmography (PPG) [13, 50, 56]. However, many of the works on gesture detection with PPG sensors focus on macro gestures; we extend this line of research by proposing efficient and low-fatigue directional microgestures.

Among all sensing modalities, inertial measurement units (IMUs) are particularly noteworthy for their affordability, low power consumption, and extensive integration into consumer-grade wearable devices. As a result, IMUs are widely used for capturing motion-based hand gestures, particularly those related to arm and hand movements [29, 54, 65, 66].

While IMU-based approaches build a good baseline for accurate gesture recognition, many researchers use a hybrid approach where they combine the IMU sensors with another sensor such as EMG [57, 67], and FMG [47]. Wand et al. [56] used a combination of PPG, FMC, and accelerometer to recognize a set of 12 hand gestures, giving a detailed analysis of the contribution of each sensor to the model's prediction. However, their gesture set features primarily macro gestures such as wrist flexion or extension. In our work, we focus on dynamic directional microgestures.

While early gesture recognition techniques such as dynamic time warping (DTW) [35] and hidden Markov models (HMM) [39] were successful in recognizing relatively simple gestures that followed predictable, smooth trajectories, they are limited in their ability to capture the complexity of more intricate or expressive gestures. These methods excelled with minimal data but struggled when dealing with gestures involving non-linear movement or high variability. To overcome these challenges, newer methods have been introduced, often relying on large datasets of predefined gestures, enabling the use of more sophisticated models. Another class of

gesture recognition models are traditional classifiers like support vector machines (SVMs) and tree-based methods [17, 20]. Recent work has mostly been investigating deep learning architectures for gesture recognition, as they do not require manual feature engineering but learn relevant features from the data directly. These technologies include convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [19, 68].

2.2 Microgestures

Microgestures, subtle and precise hand or finger movements, enable efficient, non-intrusive interactions, making them ideal for applications requiring seamless input. However, their small scale presents challenges in accurate sensing and classification.

Extensive research has been done to explore the space of single-handed microgestures [5, 22, 31]. Many of these microgestures are thumb-to-volar (palm) interactions [46, 62, 69], which benefit from a large surface area on the palm. This enables large sets of gesture, which can be used for input modalities such as typing [32] or interactive touch surface [45, 60].

On the other hand, thumb-to-index microgestures focus on interactions between the thumb and the index finger. An example includes actions like rubbing the thumb against the radial side of the index finger [27, 32, 45, 60]. These gestures can be classified using static [27] or worn [55] cameras, sensors such as capacitive rings [9, 60], IMU-equipped rings [33], magnetic contacts embedded in gloves [45], or small wearable sensors which are placed at the fingertips [25]. To our knowledge, no existing solution recognizes these thumb-to-finger swipes.

3 Sensing Principle

Wrist-worn IMUs are effective for detecting taps and coarse hand motions [1, 10, 29, 40, 61, 66], but often insufficient to resolve subtle finger microgestures such as directional swipes. These gestures produce only weak wrist rotation and limited inertial signatures, making them difficult to distinguish using IMU signals alone.

We focus on IMU and PPG because both sensing modalities are already well established in commodity wearable hardware—including smartwatches and fitness trackers, and increasingly multi-channel PPG systems in devices such as the Apple Watch and Huawei Watch—making them attractive building blocks for deployable microgesture recognition. Our approach complements IMU sensing with *spatial* PPG measurements. While PPG is typically used for heart-rate monitoring, it is also sensitive to local tissue deformation and pressure changes at the skin [3, 52]. During fine finger motion, tendon and muscle activity in the forearm causes small, spatially varying deformations beneath the worn device. We hypothesize that these deformation patterns provide discriminative cues for microgesture recognition that are not captured by IMU alone.

3.1 Sensing Hardware

We developed a wrist-worn prototype that records and streams synchronized IMU and PPG signals (Figure 4). The device uses a QMI8658 6-axis IMU on a Waveshare ESP32-S3-LCD board to capture acceleration and angular velocity at 112 Hz. A second microcontroller (XIAO ESP32-S3) relays the data to a host computer

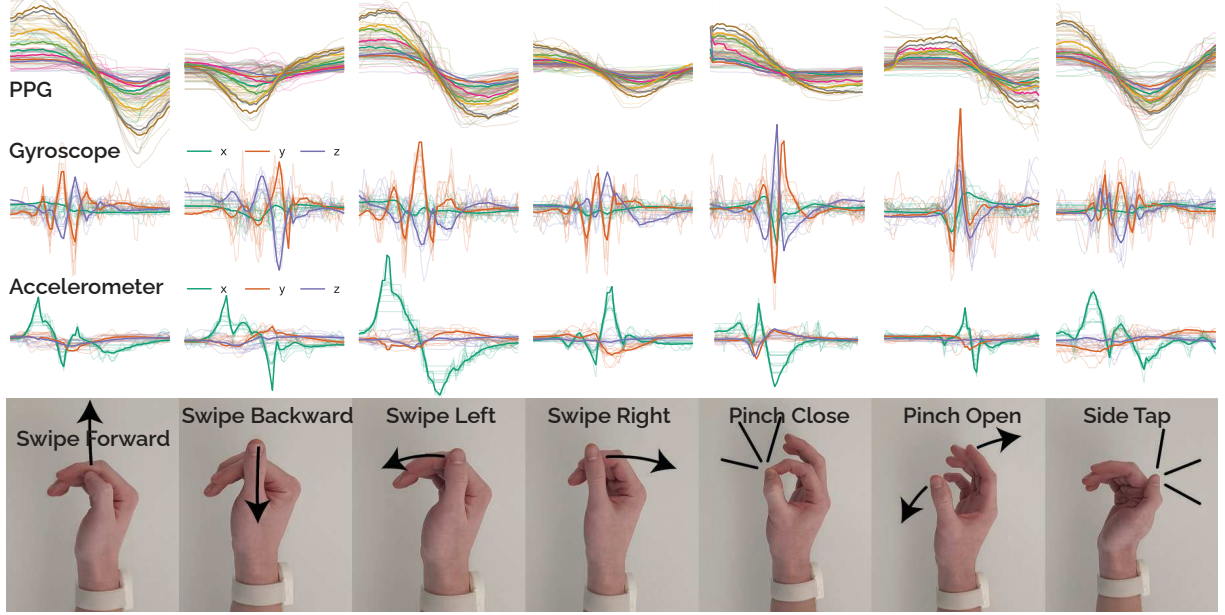


Figure 2: Overlaid signals from our gesture set. The opaque line is the barycenter of all corresponding signals from three different participants. The transparent lines are recordings with a high similarity to the barycenter.

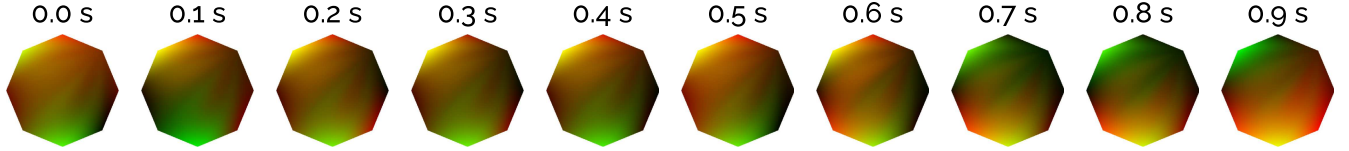


Figure 3: Spatial PPG sensing used by μ Dpad. Interpolating eight photodiode readings over a one-second gesture window reveals non-uniform reflectivity changes (a deformation “pattern”) that complements wrist IMU features for classifying directional thumb microgestures. Green and IR PPG channels are encoded into the green and red RGB values.

over serial. For optical sensing, the prototype integrates an Analog Devices MAX86177A PPG module, also sampled at 112 Hz.

We chose 112 Hz as it provides sufficient temporal resolution for our window-based classifier and online decoding, while keeping data rate and computation low for wearable deployment. The PPG sensor is mounted on the dorsal wrist and uses eight photodiodes arranged in a circular layout. We modified the MAX86177A photodiode layout to span a diameter of 2 cm, increasing spatial coverage under the device. All components are housed in a 3D-printed enclosure and attached with a flexible wrist strap.

4 Gesture Vocabulary

Following our motivation for low-effort, socially acceptable microinteractions, we designed a vocabulary of thumb-to-index-finger microgestures for wearable input (Fig. 2). It includes four directional thumb swipes (up, down, left, right), pinch close and pinch open events, a side tap, and wrist rotation during a held pinch.

This vocabulary supports both discrete and continuous control. The swipes enable navigation and menu interaction, pinch events provide command-like triggers, and wrist rotation during a held

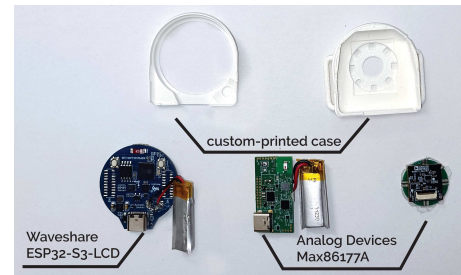


Figure 4: μ Dpad hardware prototype integrating a 6D IMU and an 8-channel spatial PPG sensor. The circular photodiode layout captures local deformation patterns under the device that help disambiguate directional thumb swipes.

pinch supports actions such as scrolling, scrubbing, or scaling. All gestures rely on small finger motions and self-contact, making them suitable for eyes-free use without requiring large arm movements.

At the same time, this gesture set is challenging for wrist-based sensing. Unlike macrogestures, microgestures produce only weak

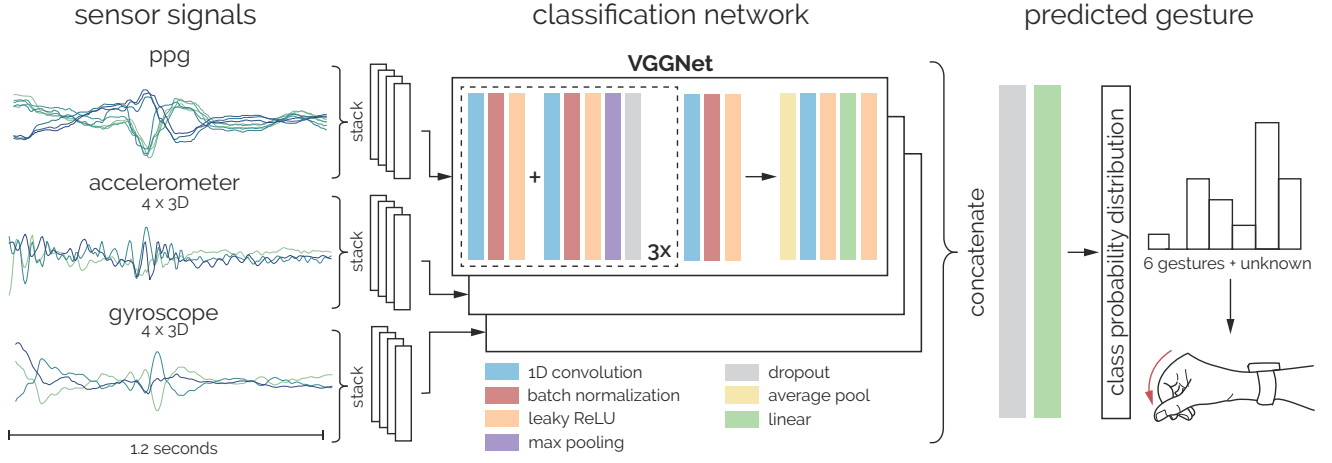


Figure 5: μ Dpad pipeline. A window-based multimodal classifier maps IMU+PPG windows to class probabilities, which are decoded over time with an HMM to enforce plausible transitions and improve robustness for online microinteraction input.

inertial signatures and subtle local tissue deformations, and some classes differ mainly in the direction of a small thumb movement. While prior work has explored thumb-to-finger microgestures using rings, gloves, cameras, or fingertip sensors, we are not aware of prior work targeting this class of directional microgestures using only wrist-worn commodity sensors such as IMU and PPG.

5 Dataset and Recording Procedure

To evaluate the proposed sensing approach, we collected a large-scale multimodal dataset from 68 participants (29 female, 39 male; mean age 24.7 ± 2 years). We excluded 2 participants due to hardware issues during recording and filtered out 6 outlier participants in our evaluation, resulting in a final cohort of 60 participants used in our evaluation. The dataset comprises 25 hours of time-synchronized 6-axis IMU and 8-channel spatial PPG signals, totaling 28,000 gesture samples and 16,000 negative samples. Participants performed randomized gesture sequences under controlled prompt timings. To ensure real-world robustness, data was collected across two distinct arm postures: with the arm bent at 90 degrees in front of the body, and with the arm fully extended and resting at the side. We additionally collected negative samples of non-gesture everyday actions, such as typing, handwriting, and object grasping. For 9 hours of the dataset, we recorded synchronized RGB video of the participants' hands. We utilized this video footage to manually verify and refine the prompt-based labels, resulting in high-fidelity annotations where only 0.8% of labels required adjustment. This synchronized video subset provides a foundation for future research into multimodal video-sensor gesture detection.

5.1 Safe and Responsible Innovation Statement

After consultation with our institution's ethics committee, the study was determined not to require formal IRB review. Participants provided written informed consent, participated voluntarily, could withdraw at any time, and consented to public release of

anonymized sensor data and hand-only videos. They were compensated USD 12.50 for a 30-minute session. Released data exclude faces and other personally identifiable information. The dataset is intended for research on wearable microgesture detection and not for identification, surveillance, or other privacy-invasive applications.

6 Benchmarking Microgesture Recognition

To rigorously evaluate microgesture recognition on our dataset, we frame the task as both an offline window-based classification problem (for benchmarking architectures) and a continuous online detection problem (for real-world applicability).

We implement several baselines from prior work, ranging from classical feature-based algorithms (SVM [29], Random Forests [17, 20]) to deep convolutional neural networks (ResNet [10], VGG [10], U-Net [14]) and modern Transformer architectures [42].

6.1 μ Dpad Architecture

μ Dpad utilizes an intermediate-fusion, VGG-style 1D CNN to process the multimodal streams. As shown in Figure 5, it processes raw IMU and PPG windows through dedicated feature extractors. Each extractor comprises six convolutional blocks (Conv1D, Batch Normalization, LeakyReLU) with interspersed dropout and max-pooling. To capture the complementary nature of the modalities, the extracted features are concatenated and passed through two fully connected layers to predict gesture probabilities. The resulting highly efficient model contains 670k parameters.

Preprocessing and Training. To ensure the model learns robust features directly from raw data, we do not apply hand-designed filtering. Signals are simply z-normalized per channel and zero-padded to a fixed window length of 150 samples. To build robustness against sensor displacement and varying execution speeds, we apply spatial augmentations (IMU scaling and 3D rotation) and temporal augmentations (Gaussian window shifting to mitigate label timing inaccuracies). The network is trained using cross-entropy

Algorithm 1 Online decoding pipeline combining window-based classification, HMM sequence decoding, and RSC activity gating.

Require: stream of IMU+PPG samples; packet size $N = 32$; window length $W = 150$; refractory $\tau = 300$ ms.

State: ring buffer of last W samples; previous RSC_{t-1} ; HMM state.

- 1: Append the new packet to the ring buffer and extract the last W samples.
- 2: Compute class probabilities $\mathbf{o}_t = f_\theta(\mathbf{x}_t)$ with the window-based classifier.
- 3: Update the HMM with observation $\mathbf{o}_t$ and decode the current state via Viterbi.
- 4: Compute an activity gate using the rate-of-change score: $RSC_t = \frac{RSC_{t-1}}{1.6} + \sum_{i \in \{x, y, z\}} |a'_{i,t}|$.
- 5: Emit a gesture event when (i) the HMM state is non-negative, (ii) $RSC_t > t_{\text{peak}} = 2$, and (iii) the refractory period τ has elapsed.

loss with label smoothing [36, 51] to reduce overconfidence and improve generalization. A comprehensive breakdown of network hyperparameters, augmentation ranges, and optimization details is provided in the Supplementary Material.

6.2 Continuous Online Detection Pipeline

For real-time applications, system robustness against false positives is more critical than sheer offline accuracy. Window-based CNN classifiers are not inherently shift-invariant [6] and are susceptible to spontaneous misclassifications during continuous non-gesture activities [29, 40]. To transition μDpad from isolated window classification to real-world streaming input, we implement a multi-stage decoding pipeline (Algorithm 1).

HMM Sequence Decoding. Our gesture set exhibits strict sequential dependencies (e.g., a *pinch close* is followed by a *pinch open*). We incorporate a Hidden Markov Model (HMM) as a post-processing step to enforce these plausible transitions. By treating the CNN’s output probabilities as observations, we decode the most likely gesture state sequence in real-time using the Viterbi algorithm.

Tracking Continuous Pinch States. By explicitly modeling the *pinch closed* state within the HMM, μDpad extends discrete gestures into continuously controllable interactions. Once a pinch is detected, μDpad activates a Madgwick filter [38] to track relative wrist rotation. This allows intuitive, one-handed control over tasks like scrolling or scrubbing, where angular velocity translates directly to interaction speed.

Activity Gating. To enable the reliable detection of gestures occurring in quick succession while suppressing noise, we employ a rate-of-change score (RSC) as an activity gate, coupled with a temporal refractory period.

7 Results

We evaluate all models using a rigorous Leave-One-Subject-Out (LOSO) cross-validation scheme to test generalization to unseen

Table 1: Model architecture comparison (mean $\pm$ std.). The VGG-style window-based model used by μDpad performs best on this dataset, outperforming alternative deep architectures and feature-based baselines.

Model	Accuracy	F1 Score
ResNet	82.1% $\pm$ 2.9	74.3% $\pm$ 2.0
U-Net	80.3% $\pm$ 3.4	71.6% $\pm$ 2.6
Transformer	62.0% $\pm$ 5.8	48.0% $\pm$ 2.0
Random Forest	61.9% $\pm$ 8.4	48.9% $\pm$ 2.5
SVM	64.7% $\pm$ 7.9	53.5% $\pm$ 2.9
μDpad (early fusion)	83.4% $\pm$ 2.9	76.6% $\pm$ 2.3
μDpad (ours)	83.7% $\pm$ 2.6	77.0% $\pm$ 2.3
μDpad (fine-tuned)	88.0% $\pm$ 5.5	82.6% $\pm$ 7.3

users. Ablation studies utilize an 8-fold cross-validation split (allocating 6/8 of participants for training, 1/8 for validation, and 1/8 for testing). We primarily report accuracy and macro F1-score.

7.1 Benchmarking Architectures and Modalities

Comparison of Machine Learning Classifiers. To establish a baseline for our multimodal dataset, we benchmarked several architectures. As shown in Table 1, μDpad (our VGG-style 1D CNN) achieved the highest accuracy at 83.7%, outperforming alternative deep networks and classical feature-based methods. An early fusion variant of μDpad performed comparably (83.4%), while ResNet and U-Net achieved 82.1% and 80.3%, respectively. The Transformer baseline struggled on this dataset (62.0%), likely due to the limited sequence length of the microgesture windows.

Traditional machine learning models performed poorly, with SVM and Random Forest yielding accuracies below 65%. The substantial gap in F1-scores between convolutional deep learning models (71.6%–77.0%) and traditional approaches (48.0%–53.5%) highlights the need for deep feature extraction for subtle microgestures. The higher standard deviations observed in traditional models indicate poor generalization across different users. Finally, because microgestures are highly individual, adding brief user-specific calibration (fine-tuning with 20 samples per class) effectively bridges this cross-user gap, pushing μDpad ’s accuracy to 88.0%.

Ablation of Sensing Modalities. To understand the value of multimodal fusion for microgestures, we evaluated the individual and combined contributions of the sensors (Table 2). Accelerometer data alone achieved 80.8% accuracy, significantly outperforming the gyroscope (73.1%). Combining both IMU sensors improved accuracy to 82.9%, confirming their complementary kinematic features.

Conversely, unimodal spatial PPG (green and IR channels) showed limited effectiveness (47.1%), indicating that optical sensing of sub-surface tendon shifts is insufficient for standalone gesture detection. However, fusing all modalities (IMU+PPG) yielded the best overall performance at 83.7% accuracy and 77.0% F1-score, while simultaneously decreasing cross-user variance.

While the absolute accuracy gain from adding PPG to the full IMU is modest (+0.8%), it is statistically reliable. A linear mixed-effects model on per-participant accuracy and F1 scores, with modality as a fixed effect and participant as a random intercept, showed a

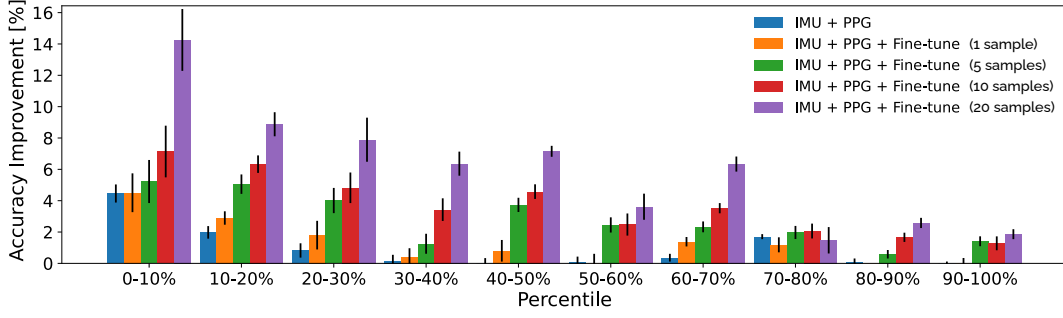


Figure 6: Relative accuracy gains in leave-one-subject-out evaluation. Participants are grouped by IMU-only accuracy percentiles (6 participants per bin). Spatial PPG yields the largest relative benefit for low-accuracy users, while fine-tuning improves performance across all bins.

Table 2: Sensor-modality ablation (mean $\pm$ std.) with PPG green+IR channels. IMU features dominate overall accuracy, while the complementary information from spatial PPG improves performance for some IMU-only failure cases.

Modalities	Accuracy	F1 Score
Acc	80.8% $\pm$ 2.5	72.8% $\pm$ 1.9
Acc + PPG	80.4% $\pm$ 2.7	72.3% $\pm$ 2.9
Gyro	73.1% $\pm$ 5.5	63.8% $\pm$ 3.6
Gyro + PPG	75.0% $\pm$ 3.9	64.3% $\pm$ 3.4
Acc + Gyro	82.9% $\pm$ 3.0	76.1% $\pm$ 2.4
PPG	47.1% $\pm$ 5.3	35.3% $\pm$ 2.0
Acc + Gyro + PPG	83.7% $\pm$ 2.6	77.0% $\pm$ 2.3

small but significant benefit of adding PPG over IMU+Gyro: +0.73% accuracy (95% CI [0.32, 1.13], $p = .0003$) and +0.76% F1 (95% CI [0.26, 1.27], $p = .003$), consistent with paired two-sided t-tests across participants ($p = .0007$ for accuracy, $p = .004$ for F1). As shown in the confusion matrices (Supplementary Figure 11), the spatial PPG subtly reduces misclassifications among mechanically similar gestures (e.g., directional swipes and pinch motions) and helps the model generalize to challenging users whose subtle thumb movements fail to register strongly on the IMU. As illustrated in Figure 6, fusing PPG improved accuracy by up to 4% for users in the bottom 10th percentile, suggesting that optical sensing offers a modest but consistent complementary gain, while IMU remains the dominant modality and PPG remains a promising target for future algorithmic work on the released dataset.

7.2 Continuous Online Detection and HMM Post-Processing

While offline classification evaluates perfectly pre-segmented windows, continuous real-time detection is inherently much harder. In a continuous stream, the system must navigate a massive class imbalance dominated by the "null class" (everyday background activities) while handling misaligned windows, transition frames, and partial gesture executions. Consequently, raw window-based classifiers typically suffer significant performance drops and high false-positive rates when deployed online.

Table 3: Online evaluation: window-based classifier vs. HMM decoding. The HMM improves accuracy/F1 by enforcing plausible state transitions.

Model	Accuracy	F1 Score
Window-based	75.0% $\pm$ 10.5	66.9% $\pm$ 6.8
HMM	78.6% $\pm$ 9.0	71.5% $\pm$ 7.7

To evaluate μ Dpad in this realistic continuous setting, we streamed unsegmented sequences from our dataset through our pipeline. A detection was classified as a true positive if it overlapped the ground-truth annotation by at least 70%, while windows with less than 20% overlap were assigned to the negative class (ambiguous 20–30% overlaps were excluded).

As shown in Table 3, the raw window-based classifier struggles with the continuous data stream, yielding a high false positive rate (FPR) of 8.3%. However, applying our computationally lightweight HMM post-processing effectively mitigates this. By enforcing plausible biomechanical state transitions and smoothing transient noise, the HMM improves overall accuracy to 78.6% and the F1-score to 71.5% (see Supp. Figure 13 for continuous confusion matrices).

Crucially, the HMM reduces the FPR to just 5.3%. Because microgestures rely on tiny thumb movements that are easily conflated with normal hand twitches during daily tasks, maintaining a low FPR in an unsegmented stream is the most critical metric for usability. Achieving 78.6% accuracy with only a 5.3% FPR demonstrates that μ Dpad serves as a practical baseline for reliable, always-available interactive systems.

8 Demonstration Scenarios

To demonstrate our pipeline’s practical utility, we developed prototypes leveraging μ Dpad’s always-active, contact-free input. Requiring no wake-up gesture, it seamlessly supports microinteractions when visual attention is limited or hands are occupied (e.g., cooking, walking, Fig. 7). These prototypes illustrate how μ Dpad functions as a one-handed D-pad, while the pinch-hold state enables continuous parameter control:

Eyes-Free Music Control. Controlling audio on the go typically requires retrieving a phone or interacting with small, memorized

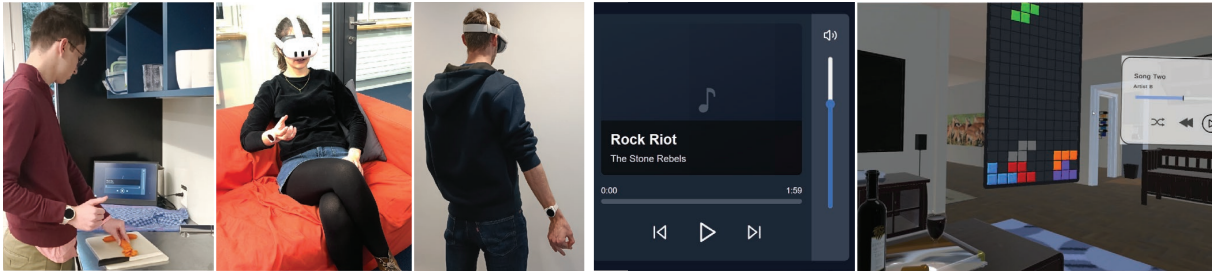


Figure 7: Example microinteraction scenarios enabled by μ Dpad. Left: always-available, low-effort input in everyday contexts without touching a screen. Right: prototype mappings for music control and lightweight game input using directional swipes, taps, and pinch-hold combined with wrist rotation.

headphone buttons. μ Dpad enables a continuous, eyes-free alternative. Swiping left or right navigates tracks, a side tap plays/pauses, and initiating a pinch-hold while rotating the forearm provides smooth, continuous adjustment of playback volume.

Lightweight Gaming. μ Dpad can also support low-effort, always-available game control. In a prototyped Tetris-like game, directional swipes map intuitively to lateral block movements, while pinching and rotating the forearm controls block rotation.

Because always-on wearable input remains susceptible to occasional false positives, we recommend that application designers pair μ Dpad’s robust HMM decoding with error-tolerant interaction designs, such as mapping microgestures to easily reversible actions or utilizing context-dependent activation.

9 Discussion

Our systematic benchmark highlights the intrinsic difficulty of wrist-based microgesture recognition. While models targeting large, macro arm movements often achieve 90% accuracy, thumb–index microinteractions plateau around 84%. This gap stems from minimal inertial energy and high anatomical variability across users. Our results suggest that multimodal fusion is a promising path forward: while the IMU captures global kinematics, spatial PPG captures local subsurface tendon deformations, improving robustness particularly for users where inertial signals are weak or ambiguous. Furthermore, our continuous evaluation shows that for real-world deployment, explicitly modeling biomechanical constraints (via HMM decoding) is essential to suppress transient noise and achieve a usable false-positive rate (5.3%) in unsegmented data streams.

While μ Dpad provides a useful baseline, several open challenges remain. Our prototype utilizes an 8-channel PPG layout to maximize spatial coverage; however, optical PPG signals remain sensitive to variations in sensor contact, motion artifacts, and skin properties. Stronger multimodal baselines building on our VGG/ResNet architectures, combined with improved sensing strategies and models that learn invariances to these perturbations, could yield better performance.

To catalyze progress on these challenges, we contribute our full 25-hour, 66-participant dataset to the research community. Crucially, this release includes 9 hours of synchronized RGB hand video, along with sensor data from an integrated IMU + spatial-PPG wearable configuration that is becoming available in consumer

devices but is rarely accessible as raw data. As traditional supervised learning on sensor data alone shows diminishing returns with increasing dataset size (Supplementary Figure 10), we argue that advances in microgesture recognition will stem from better algorithms and cross-modal learning. By providing aligned video and multi-channel sensor data, our dataset establishes a foundation for exploring video-to-sensor knowledge distillation, contrastive multimodal pretraining, and self-supervised approaches for wearable interaction.

10 Conclusion

We presented a large-scale dataset for wearable microgesture recognition using multimodal signals from commodity sensors, comprising synchronized IMU and PPG data from 66 participants, totaling 25 hours and 44,000 annotated gestures. The dataset was captured with an integrated wearable prototype using an IMU + spatial PPG sensing configuration that is emerging in consumer fitness watches, but whose raw optical signals are unavailable for research. To support future multimodal research, we further complement the dataset with 9 hours of aligned RGB video for cross-modal learning.

Using this dataset, we benchmark a range of commonly used architectures for gesture recognition and introduce μ Dpad as a simple multimodal baseline for recognizing one-handed directional thumb microgestures for mobile and wearable interaction. μ Dpad detects thumb swipes in four directions (akin to a D-pad), as well as pinch close/open events, and supports continuous parameter control via wrist rotation during pinch holds.

Across 60 participants, our multimodal IMU+PPG baseline achieves 83.7% accuracy. Furthermore, our HMM-based post-processing improves robustness in continuous settings by enforcing simple, interaction-relevant transition constraints without requiring additional training data.

We hope that the release of this dataset, benchmark, and baseline will enable future algorithmic work to better exploit spatial PPG for microgesture recognition and to close the gap between wearable microgesture recognition and the performance levels achieved for macro gestures.

Acknowledgments

We thank Silvan Roth for his support in manufacturing and assembling the circular sensor design.

References

- [1] Ahmad Akl, Chen Feng, and Shahrokh Valaee. 2011. A Novel Accelerometer-Based Gesture Recognition System. *IEEE Transactions on Signal Processing* 59, 12 (Dec. 2011), 6197–6205. doi:10.1109/TSP.2011.2165707
- [2] D Ashbrook. 2007. *Supporting mobile microinteractions*. Ph.D. Dissertation. Ph.D. Dissertation. PhD thesis, Georgia Institute of Technology.
- [3] Valentin Bieri, Paul Strel, Berken Utku Demirel, and Christian Holz. 2023. BeliefPPG: Uncertainty-aware heart rate estimation from PPG signals via belief propagation. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (Proceedings of Machine Learning Research, Vol. 216)*, Robin J. Evans and Ilya Shpitser (Eds.). PMLR, Pittsburgh, PA, USA, 173–183. <https://proceedings.mlr.press/v216/bieri23a.html>
- [4] Doug Bowman, Chadwick Wingrave, Joshua Campbell, and Vinh Ly. 2001. Using pinch gloves (tm) for both natural and abstract interaction techniques in virtual environments. (2001).
- [5] Adrien Chaffangeon Caillet, Alix Goguy, and Laurence Nigay. 2023. μGlyph: A Microgesture Notation. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–28. doi:10.1145/3544548.3580693
- [6] Anadi Chaman and Ivan Dokmanic. 2021. Truly shift-invariant convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3773–3783.
- [7] Victor Chang, Rahman Olamide Eniola, Lewis Golightly, and Qianwen Ariel Xu. 2023. An exploration into human-computer interaction: Hand gesture recognition management in a challenging environment. *SN Computer Science* 4, 5 (2023), 441.
- [8] Ke-Yu Chen, Shwetak N. Patel, and Sean Keller. 2016. Finexus: Tracking Precise Motions of Multiple Fingertips Using Magnetic Sensing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1504–1514. doi:10.1145/2858036.2858125
- [9] Taizhou Chen, Tianpei Li, Xingyu Yang, and Kening Zhu. 2023. Efring: Enabling thumb-to-index-finger microgesture interaction through electric field sensing using single smart ring. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–31.
- [10] Wenqiang Chen, Ziqi Wang, Pengrui Quan, Zhencan Peng, Shupe Lin, Mani Srivastava, Wojciech Matusik, and John Stankovic. 2023. Robust Finger Interactions with COTS Smartwatches via Unsupervised Siamese Adaptation. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, San Francisco CA USA, 1–14. doi:10.1145/3586183.3606794
- [11] Ming Jin Cheok, Zaid Omar, and Mohamed Hisham Jaward. 2019. A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics* 10, 1 (2019), 131–153.
- [12] SN David Chua, KY Richard Chin, SF Lim, and Pushpdant Jain. 2022. Hand gesture control for human-computer interaction with deep learning. *Journal of Electrical Engineering & Technology* 17, 3 (2022), 1961–1970.
- [13] Ethan Eddy, Evan Campbell, Ulysse Côté-Allard, Scott Bateman, and Erik Scheme. 2024. Discrete Gesture Recognition Using Multimodal PPG, IMU, and Single-Channel EMG Recorded at the Wrist. *IEEE Sensors Letters* 8, 9 (Sept. 2024), 1–4. doi:10.1109/LSSENS.2024.3447240
- [14] Meghaa Elangovan and Mithun Kumar Kar. 2025. Hand Gesture Segmentation using UNet-VAE Deep Learning Architecture. In *2025 5th International Conference on Intelligent Technologies (CONIT)*. IEEE, 1–7.
- [15] O. Fukuda, T. Tsuji, M. Kaneko, and A. Otsuka. 2003. A human-assisting manipulator teleoperated by EMG signals and arm motions. *IEEE Transactions on Robotics and Automation* 19, 2 (2003), 210–222. doi:10.1109/TRA.2003.808873
- [16] Qing Gao, Zhaojie Ju, Yongquan Chen, Qiwen Wang, and Chuliang Chi. 2023. An Efficient RGB-D Hand Gesture Detection Framework for Dexterous Robot Hand-Arm Teleoperation System. *IEEE Transactions on Human-Machine Systems* 53, 1 (2023), 13–23. doi:10.1109/THMS.2022.3206663
- [17] Marcus Georgi, Christoph Amma, and Tanja Schultz. 2025. Recognizing Hand and Finger Gestures with IMU Based Motion and EMG Based Muscle Activity Sensing. In *International Conference on Bio-Inspired Systems and Signal Processing*. 99–108.
- [18] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. 2020. MEGATrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (ToG)* 39, 4 (2020), 87–1.
- [19] Fang Hu, Peng He, Songlin Xu, Yin Li, and Cheng Zhang. 2020. FingerTrak: Continuous 3D Hand Pose Tracking by Deep Learning Hand Silhouettes Captured by Miniature Thermal Cameras on Wrist. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2 (June 2020), 71:1–71:24. doi:10.1145/3397306
- [20] Yasha Iravantchi, Mayank Goel, and Chris Harrison. 2019. BeamBand: Hand Gesture Sensing with Ultrasonic Beamforming. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3290605.3300245
- [21] Yasha Iravantchi, Yang Zhang, Evi Bernitsas, Mayank Goel, and Chris Harrison. 2019. Interferi: Gesture Sensing Using On-Body Acoustic Interferometry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300506
- [22] Nikhita Joshi, Parastoo Abtahi, Raj Sodhi, Nitzan Bartov, Jackson Rushing, Christopher Collins, Daniel Vogel, and Michael Glueck. 2023. Transferable Microgestures Across Hand Posture and Location Constraints: Leveraging the Middle, Ring, and Pinky Fingers. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3586183.3606713
- [23] Pyeong-Gook Jung, Gukchan Lim, Seonghyok Kim, and Kyoungchul Kong. 2015. A Wearable Gesture Recognition Device for Detecting Muscular Activities Based on Air-Pressure Sensors. *IEEE Transactions on Industrial Informatics* 11, 2 (April 2015), 485–494. doi:10.1109/TII.2015.2405413
- [24] Patrick Kaifosh and Thomas R Reardon. 2025. A generic non-invasive neuromotor interface for human-computer interaction. *Nature* (2025), 1–10.
- [25] Keiko Katsuragawa, Ju Wang, Ziyang Shan, Ningshan Ouyang, Omid Abari, and Daniel Vogel. 2019. Tip-Tap: Battery-free Discrete 2D Fingertip Input. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. ACM, New Orleans LA USA, 1045–1057. doi:10.1145/3332165.3347907
- [26] Wolf Kienle and Ken Hinckley. 2014. LightRing: Always-Available 2D Input on Any Surface. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*. Association for Computing Machinery, New York, NY, USA, 157–160. doi:10.1145/2642918.2647376
- [27] Kenrick Kin, Chengde Wan, Ken Koh, Andrei Marin, Necati Cihan Camgöz, Yubo Zhang, Yujun Cai, Fedor Kovalev, Moshe Ben-Zacharia, Shannon Hoople, Marcos Nunes-Ueno, Mariel Sanchez-Rodriguez, Ayush Bhargava, Robert Wang, Eric Sauter, and Shugao Ma. 2024. STMG: A Machine Learning Microgesture Recognition System for Supporting Thumb-Based VR/AR Input. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–15. doi:10.1145/3613904.3642702
- [28] Karly Kudrinko, Emile Flavin, Xiaodan Zhu, and Qingguo Li. 2021. Wearable Sensor-Based Sign Language Recognition: A Comprehensive Review. *IEEE Reviews in Biomedical Engineering* 14 (2021), 82–97. doi:10.1109/RBME.2020.3019769
- [29] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. ViBand: High-Fidelity Bio-Acoustic Sensing Using Commodity Smartwatch Accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. Association for Computing Machinery, New York, NY, USA, 321–333. doi:10.1145/2984511.2984582
- [30] Chi-Jung Lee, Ruidong Zhang, Devansh Agarwal, Tianhong Catherine Yu, Vipin Gunda, Oliver Lopez, James Kim, Sicheng Yin, Boao Dong, Ke Li, et al. 2024. Echowrist: Continuous hand pose tracking and hand-object interaction recognition using low-power active acoustic sensing on a wristband. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.
- [31] Guangchuan Li, David Rempel, Yue Liu, Weitao Song, and Carisa Harris Adamson. 2021. Design of 3D Microgestures for Commands in Virtual Reality or Augmented Reality. *Applied Sciences* 11, 14 (Jan. 2021), 6375. doi:10.3390/app11146375
- [32] Chen Liang, Chi Hsia, Chun Yu, Yukang Yan, Yuntao Wang, and Yuanchun Shi. 2023. DRG-KeyBoard: Enabling Subtle Gesture Typing on the Fingertip with Dual IMU Rings. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4 (Jan. 2023), 170:1–170:30. doi:10.1145/3569463
- [33] Chen Liang, Chun Yu, Yue Qin, Yuntao Wang, and Yuanchun Shi. 2021. DualRing: Enabling Subtle and Expressive Hand Interaction with Dual IMU Rings. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3 (Sept. 2021), 115:1–115:27. doi:10.1145/3478114
- [34] Jaime Lien, Nicholas Gillian, M. Emre Karagozler, Patrick Amihoud, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: Ubiquitous Gesture Sensing with Millimeter Wave Radar. *ACM Trans. Graph.* 35, 4 (July 2016), 142:1–142:19. doi:10.1145/2897824.2925953
- [35] Jiayang Liu, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. 2009. uWave: Accelerometer-based Personalized Gesture Recognition and Its Applications. *Pervasive and Mobile Computing* 5, 6 (Dec. 2009), 657–675. doi:10.1016/j.pmcj.2009.07.007
- [36] Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. 2020. Does Label Smoothing Mitigate Label Noise? arXiv:2003.02819 [cs] doi:10.48550/arXiv.2003.02819
- [37] Jing Luo, Zhidong Lin, Yanan Li, and Chenguang Yang. 2020. A Teleoperation Framework for Mobile Robots Based on Shared Control. *IEEE Robotics and Automation Letters* 5, 2 (2020), 377–384. doi:10.1109/LRA.2019.2959442
- [38] Sebastian O. H. Madgwick, Andrew J. L. Harrison, and Ravi Vaidyanathan. 2011. Estimation of IMU and MARG Orientation Using a Gradient Descent Algorithm. In *2011 IEEE International Conference on Rehabilitation Robotics*. 1–7. doi:10.1109/ICORR.2011.5975346
- [39] Stephen J. McKenna and Kenny Morrison. 2004. A Comparison of Skin History and Trajectory-Based Representation Schemes for the Recognition of User-Specified Gestures. *Pattern Recognition* 37, 5 (May 2004), 999–1009. doi:10.1016/j.patcog.2003.09.007

- [40] Manuel Meier, Paul Strel, Andreas Fender, and Christian Holz. 2021. TapID: Rapid touch interaction in virtual reality using wearable sensing. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 519–528.
- [41] George B. Mo, John J Dudley, and Per Ola Kristensson. 2021. Gesture Knitter: A Hand Gesture Design Tool for Head-Mounted Mixed Reality Applications. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–13. doi:10.1145/3411764.3445766
- [42] Mansoorh Montazerin, Elahe Rahimian, Farnoosh Naderkhani, S. Farokh Atashzar, Svetlana Yanushkevich, and Arash Mohammadi. 2023. Transformer-Based Hand Gesture Recognition from Instantaneous to Fused Neural Decomposition of High-Density EMG Signals. *Scientific Reports* 13, 1 (July 2023), 11000. doi:10.1038/s41598-023-36490-w
- [43] Farshid Salemi Parizi, Eric Whitmire, and Shwetak Patel. 2020. AuraRing: Precise Electromagnetic Finger Tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4 (Sept. 2020), 150:1–150:28. doi:10.1145/3369831
- [44] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. 2024. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9826–9836.
- [45] Anna Peshock, Julia Duvall, and Lucy E. Dunne. 2014. Argot: A Wearable One-Handed Keyboard Glove. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers: Adjunct Program (ISWC '14 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 87–92. doi:10.1145/2641248.2641266
- [46] Manuel Prätorius, Dimitar Valkov, Ulrich Burghbacher, and Klaus Hinrichs. 2014. DigiTap: An Eyes-Free VR/AR Symbolic Input Device. In *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology (VRST '14)*. Association for Computing Machinery, New York, NY, USA, 9–18. doi:10.1145/2671015.2671029
- [47] Mahdi Rasouli, Karthik Chellamuthu, John-John Cabibihan, and Sunil L. Kukreja. 2016. Towards Enhanced Control of Upper Prosthetic Limbs: A Force-Myographic Approach. In *2016 6th IEEE International Conference on Biomedical Robotics and Biomechanics (BioRob)*. 232–236. doi:10.1109/BIOROB.2016.7523629
- [48] Francesca Stival, Stefano Michieletto, Matteo Cognolato, Enrico Pagello, Henning Müller, and Manfredo Atzori. 2019. A Quantitative Taxonomy of Human Hand Grasps. *Journal of NeuroEngineering and Rehabilitation* 16, 1 (Feb. 2019), 28. doi:10.1186/s12984-019-0488-x
- [49] Paul Strohmeier, Roel Vertegaal, and Audrey Girouard. 2012. With a Flick of the Wrist: Stretch Sensors as Lightweight Input for Mobile Devices. In *Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction (TEI '12)*. Association for Computing Machinery, New York, NY, USA, 307–308. doi:10.1145/2148131.2148195
- [50] Karthik Subramanian, Celal Savur, and Ferat Sahin. 2020. Using Photoplethysmography for Simple Hand Gesture Recognition. In *2020 IEEE 15th International Conference of System of Systems Engineering (SoSE)*. 307–312. doi:10.1109/SoSE50414.2020.9130489
- [51] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. arXiv:1512.00567 [cs] doi:10.48550/arXiv.1512.00567
- [52] Andriy Temko. 2017. Accurate heart rate monitoring during physical exercises using PPG. *IEEE Transactions on Biomedical Engineering* 64, 9 (2017), 2016–2024.
- [53] Hoang Truong, Shuo Zhang, Ufuk Muncuk, Phuc Nguyen, Nam Bui, Anh Nguyen, Qin Lv, Kaushik Chowdhury, Thang Dinh, and Tam Vu. 2018. CapBand: Battery-free Successive Capacitance Sensing Wristband for Hand Gesture Recognition. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems (SenSys '18)*. Association for Computing Machinery, New York, NY, USA, 54–67. doi:10.1145/3274783.3274854
- [54] Tran Huy Vu, Archan Misra, Quentin Roy, Kenny Choo Tsu Wei, and Youngki Lee. 2018. Smartwatch-based early gesture detection 8 trajectory tracking for interactive gesture-driven applications. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–27.
- [55] Anandghan Waghmare, Roger Boldu, Eric Whitmire, and Wolf Kienzle. 2023. OptiRing: Low-Resolution Optical Sensing for Subtle Thumb-to-Index Micro-Interactions. In *Proceedings of the 2023 ACM Symposium on Spatial User Interaction (SUI '23)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3607822.3614538
- [56] Hong Wang, Peiqi Kang, Qinghua Gao, Shuo Jiang, and Peter B. Shull. 2022. A Novel PPG-FMG-ACC Wristband for Hand Gesture Recognition. *IEEE Journal of Biomedical and Health Informatics* 26, 10 (Oct. 2022), 5097–5108. doi:10.1109/JBHI.2022.3194017
- [57] Le Wang, Jianting Fu, Hui Chen, and Bin Zheng. 2023. Hand Gesture Recognition Using Smooth Wavelet Packet Transformation and Hybrid CNN Based on Surface EMG and Accelerometer Signal. *Biomedical Signal Processing and Control* 86 (Sept. 2023), 105141. doi:10.1016/j.bspc.2023.105141
- [58] Nana Wang, Jianwei Niu, Xuefeng Liu, Dongqin Yu, Guogang Zhu, Xinghao Wu, Mingliang Xu, and Hao Su. 2024. BeyondVision: an EMG-driven micro hand gesture recognition based on dynamic segmentation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. 6044–6052.
- [59] Robert Wang, Sylvain Paris, and Jovan Popović. 2011. 6D hands: markerless hand-tracking for computer aided design. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. 549–558.
- [60] Martin Weigel, Aditya Shekhar Nittala, Alex Olwal, and Jürgen Steimle. 2017. SkinMarks: Enabling Interactions on Body Landmarks Using Conformal Skin Electronics. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3095–3105. doi:10.1145/3025453.3025704
- [61] Hongyi Wen, Julian Ramos Rojas, and Anind K. Dey. 2016. Serendipity: Finger Gesture Recognition Using an Off-the-Shelf Smartwatch. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3847–3851. doi:10.1145/2858036.2858466
- [62] Eric Whitmire, Mohit Jain, Divye Jain, Greg Nelson, Ravi Karkar, Shwetak Patel, and Mayank Goel. 2017. DigiTouch: Reconfigurable Thumb-to-Finger Input and Text Entry on Head-mounted Displays. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3 (Sept. 2017), 113:1–113:21. doi:10.1145/3130978
- [63] Katrin Wolf, Anja Naumann, Michael Rohs, and Jörg Müller. 2011. A taxonomy of microinteractions: Defining microgestures based on ergonomic and scenario-dependent requirements. In *IFIP conference on human-computer interaction*. Springer, 559–575.
- [64] Erwin Wu, Ye Yuan, Hui-Shyong Yeo, Aaron Quigley, Hideki Koike, and Kris M. Kitani. 2020. Back-Hand-Pose: 3D Hand Pose Estimation for a Wrist-worn Camera via Dorsum Deformation Network. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20)*. Association for Computing Machinery, New York, NY, USA, 1147–1160. doi:10.1145/3379337.3415897
- [65] Chao Xu, Parth H. Pathak, and Prasant Mohapatra. 2015. Finger-Writing with Smartwatch: A Case for Finger and Hand Gesture Recognition Using Smartwatch. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications (HotMobile '15)*. Association for Computing Machinery, New York, NY, USA, 9–14. doi:10.1145/2699343.2699350
- [66] Xuhai Xu, Jun Gong, Carolina Brum, Lilian Liang, Bongsoo Suh, Kumar Gupta, Yash Agarwal, Laurence Lindsey, Runchang Kang, Behrooz Shahsavari, Tu Nguyen, Heriberto Nieto, Scott E. Hudson, Charlie Maalouf, Seyed Mousavi, and Gierad Laput. 2022. Enabling Hand Gesture Customization on Wrist-Worn Devices. In *CHI Conference on Human Factors in Computing Systems*. 1–19. arXiv:2203.15239 [cs] doi:10.1145/3491102.3501904
- [67] Momona Yamagami, Claire L. Mitchell, Alexandra A. Portnova-Fahreeva, Junhan Kong, Jennifer Mankoff, and Jacob O. Wobbrock. 2024. Customized Mid-Air Gestures for Accessibility: A SB Recognizer for Multi-Dimensional Biosignal Gestures. arXiv:2409.08402 [cs] doi:10.48550/arXiv.2409.08402
- [68] Hui-Shyong Yeo, Erwin Wu, Juyoung Lee, Aaron Quigley, and Hideki Koike. 2019. Opisthenar: Hand Poses and Finger Tapping Recognition by Observing Back of Hand Using Embedded Wrist Camera. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST '19)*. Association for Computing Machinery, New York, NY, USA, 963–971. doi:10.1145/3332165.3347867
- [69] Cheng Zhang, Anandghan Waghmare, Pranav Kundra, Yiming Pu, Scott Gilliland, Thomas Ploetz, Thad E. Starner, Omer T. Inan, and Gregory D. Abowd. 2017. FingerSound: Recognizing Unistroke Thumb Gestures Using a Ring. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3 (Sept. 2017), 120:1–120:19. doi:10.1145/3130985
- [70] Tianming Zhao, Jian Liu, Yan Wang, Hongbo Liu, and Yingying Chen. 2018. PPG-based Finger-level Gesture Recognition Leveraging Wearables. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*. 1457–1465. doi:10.1109/INFOCOM.2018.8486006
- [71] Tianming Zhao, Jian Liu, Yan Wang, Hongbo Liu, and Yingying Chen. 2019. Towards low-cost sign language gesture recognition leveraging wearables. *IEEE Transactions on Mobile Computing* 20, 4 (2019), 1685–1701.

A Supplementary Material

We provide a video of μ Dpad in action. For that, we refer to the submitted video.

B Detailed Collection Procedure

B.1 Participant Demographics and Sample Count

The 68 participants (66 right-handed, 2 left-handed) had a mean wrist circumference of 16 ± 1.4 cm, a mean hand width of 8.2 ± 0.75 cm, and a mean hand length of 17.4 ± 1.3 cm. The dataset contains $\approx 28,000$ gesture samples across eight classes: four swipe directions, pinch close/open, and side tap each have $\approx 3,300 - 3,400$ samples, while held pinch/rotation has $\approx 4,400$. An additional 16,000 negative samples bring the total to 44,000. Per-participant counts were uniform by design. Recordings included forward- and downward-facing wrist positions, and negative samples were not tightly constrained.

B.2 Recording Protocol

The study comprised one initial trial session for familiarization, followed by four recorded sessions. To account for day-to-day wear variability, participants completely removed and reattached the sensor prototype between all sessions. During the recorded sessions, participants followed randomly generated prompts displayed on a screen. Each session contained 10 sequences of 12 gestures (120 gestures total). Each gesture prompt was displayed for 1.5 s, followed by a 0.5 s break. This prompt timing served as the label anchor; any remaining onset or offset jitter was addressed via temporal augmentation during training. The randomization ensured each gesture appeared exactly 15 times per session with equal transition probabilities. If an experimenter observed an execution error, the sequence was immediately deleted and repeated.

B.3 Data Processing

For data recording, the incoming streams of the 8-channel spatial PPG signal and 6-channel IMU signal were saved into text files and later aligned using the accelerometer recordings of both devices. For alignment, we computed the normalized cross-correlation between the accelerometer signals and shifted the IMU signal to maximize the metric. For live processing, we directly aligned both signals using a fixed delay time. The accelerometer signal was recorded with a maximum range of $\pm 2g$ and the gyroscope with a maximum range of ± 1024 degrees/second, capturing fine changes in the signal while accounting for large movement values.

B.4 Signal Characteristics and Variability

Qualitative analysis of the data demonstrates high intra-participant consistency for individual gestures. Figure 2 overlays signals from different gestures and shows the minor variations from their respective barycenters. We controlled only for exceptionally loose wristband attachment; qualitative inspection indicated comparable signal quality for moderate and tight fits. Tighter wristbands generally improved PPG signal quality, while looser fits introduced more motion artifacts in the PPG. For instance, gyroscope signals reliably disambiguate opposing directional swipes through mirrored

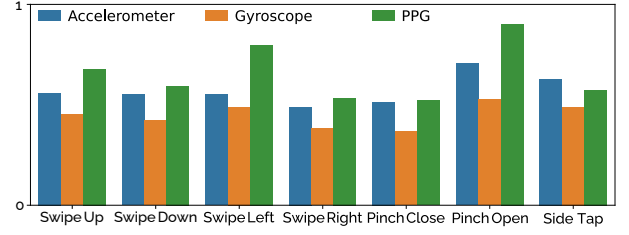


Figure 8: Signal magnitude by gesture and modality, illustrating complementarity. Directional swipes produce discriminative patterns in both IMU and spatial PPG, motivating multimodal fusion.

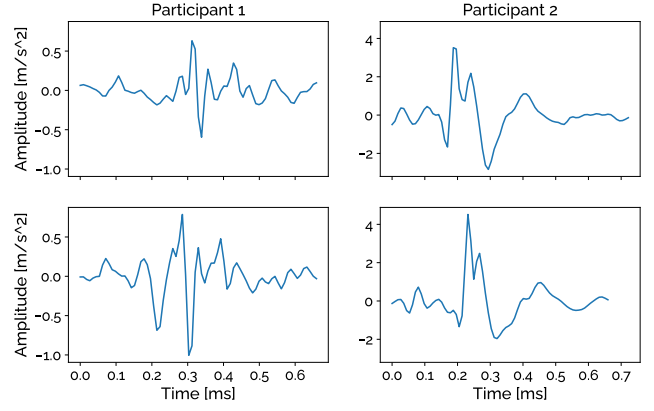


Figure 9: Example accelerometer magnitude during pinch gestures for two participants. Within-user trials are relatively consistent, while between-user differences are larger, motivating cross-user robustness techniques and optional personalization.

z-axis signals. However, significant inter-participant variability exists in both the inertial and spatial PPG patterns (Figure 9). This variance underscores the necessity of a large participant pool for robust model benchmarking and highlights the requirement for user-specific adaptation in wearable gesture recognition.

C Training Details

C.1 Implementation and Training Details

This section details the hyperparameter configurations, data augmentation strategies, and loss functions utilized to train the μ Dpad architecture and the baseline models discussed in the main text.

Data and Temporal Augmentation. To prevent overfitting and enhance robustness against variations in sensor placement and execution style, we applied several augmentations during training:

- **Scaling:** Each IMU signal is scaled by a factor uniformly sampled between 0.6 and 1.6.
- **Rotation:** To simulate sensor displacement around the wrist, we apply a random rotation along an arbitrary 3D axis within the range of -15° to $+15^\circ$.

- **Temporal Shifting:** We introduce random shifts to each window in time. We sample a new window center from a Gaussian distribution centered on the annotated label, and the signal window is subsequently aligned to this new center. This mitigates label timing inaccuracies (e.g., when a user starts a gesture slightly before or after the prompted time window).

Hyperparameters and Optimization. All neural network models were trained for 400 epochs using a batch size of 512. We utilized the Adam optimizer with an initial learning rate of 10^{-3} and a weight decay of 10^{-4} . A learning rate scheduler was applied to reduce the learning rate based on the validation loss plateau. To avoid overfitting, we employed early stopping based on the validation loss and applied a dropout rate of 50% within the fully connected layers.

For the SVM and Random Forest baselines, we extracted standard TSFEL statistical, temporal, and spectral features, e.g., moments, energy/RMS, entropy, spectral descriptors, and wavelet features, per window and modality, then concatenated them for multimodal classification.

Label Smoothing. We train our models using a cross-entropy loss and apply label smoothing. Label smoothing is a regularization technique that assigns the ground truth labels a smoothed probability distribution instead of a hard one-hot encoded representation [51]. By redistributing a fraction of the probability mass across all classes, the model’s overconfidence in its predictions is reduced. This reduces overfitting and allows for better generalization to noisy sensor data [36].

Mathematically, given a dataset with C classes and a smoothing parameter α (set to 0.1 in our experiments), the smoothed label vector y_{smooth} is calculated as:

$$y_{\text{smooth}} = (1 - \alpha) \cdot y_{\text{one-hot}} + \frac{\alpha}{C} \quad (1)$$

where $y_{\text{one-hot}}$ is the original one-hot encoded vector.

D Additional Results

D.0.1 Dataset Scaling. As shown in Figure 10, model accuracy increases with the number of participants. Accuracy rises rapidly from 57.5% with 5 participants to approximately 75% with 18 participants. However, the improvement rate eventually plateaus at 82.8% with 54 participants. This trend indicates suggests that our dataset is sufficiently large in both total size and participant diversity.

D.1 Impact of Data Augmentation

Many augmentation techniques are known for signal data that can aid generalization and classification accuracy. We implemented many and evaluated their benefits, see Table. 4. Random rotation (up to 15°) of the IMU signal showed the strongest improvement with +1.0% accuracy and scaling the signal yielded modest gains of +0.4% accuracy. Interestingly, many other augmentation techniques, such as zooming into the window and adding random noise slightly degraded performance, decreasing accuracy by 0.8% and 0.7% respectively. We included these results in the training approach and only included scaling and rotation as data augmentations in our final model.

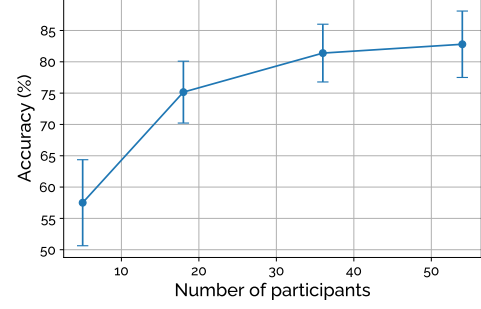


Figure 10: Accuracy of μ Dpad with increasing number of participants in the training split.

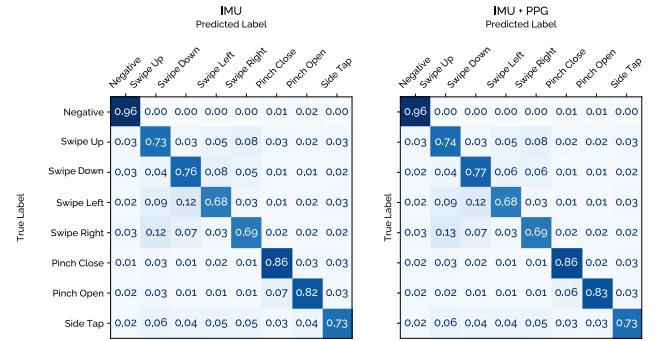


Figure 11: Leave-one-subject-out confusion matrices for IMU-only vs. IMU+PPG input. Spatial PPG reduces confusion among directional swipes.

Table 4: Data augmentation ablation. Values are absolute performance changes relative to training without augmentation; rotation and scaling are most beneficial for cross-user robustness.

Augmentation	Accuracy	F1 Score	Recall	Precision
Jitter	0.1%	0.0%	-0.0%	0.1%
Mixup	-0.1%	0.3%	0.2%	0.4%
Noise	-0.7%	-0.6%	-0.5%	-0.8%
Rotation	1.0%	0.7%	0.7%	0.6%
Scale	0.4%	0.5%	0.4%	0.5%
Wrap	0.2%	-0.0%	-0.1%	0.0%
Zoom	-0.8%	-1.2%	-1.0%	-1.2%

D.2 Per-User Results

In Figure 12, we show the per-participant accuracy when training the base model with and without PPG, as well as fine-tuning the base model using 5 samples of each gesture from the participant’s data. The average accuracy for the IMU+PPG base model is 83.9%, with a maximum accuracy of 94.3% and a minimum accuracy of 71.7%.

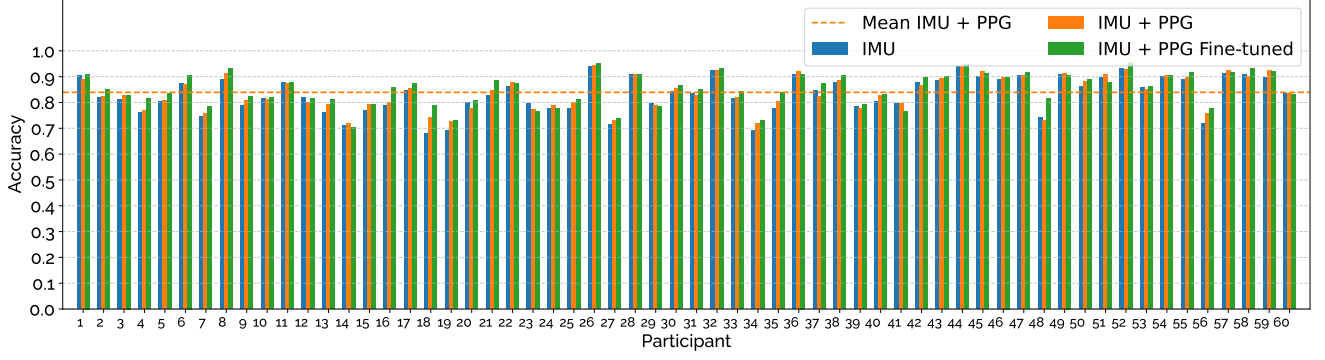


Figure 12: Classification accuracy of our method per participant comparing the classifier with only IMU input; IMU and PPG input, as well as finetuning with five gestures per class.

Table 5: Effect of lightweight personalization. Leave-one-subject-out performance when adding k labeled samples per class from the target user to fine-tune the model; increasing k steadily improves accuracy and F1.

fine-tuning samples	Accuracy	F1 Score
0	83.9% $\pm$ 6.4	78.0% $\pm$ 7.3
1	84.0% $\pm$ 6.5	78.3% $\pm$ 7.6
5	85.4% $\pm$ 6.4	79.9% $\pm$ 7.4
10	86.2% $\pm$ 6.2	80.8% $\pm$ 7.6
20	88.0% $\pm$ 5.5	82.6% $\pm$ 7.3

E Online Implementation and HMM Configuration

This section provides the specific parameters and transition rules used for the continuous online decoding pipeline of μ Dpad (as outlined in Algorithm 1 of the main text).

Stream Processing. In online inference mode, the wristband streams data in packets of 32 samples (equivalent to 280 ms). Upon the arrival of each packet, the previous 150 samples (1.3 s) are extracted from a rolling ring buffer and passed to the CNN classifier. The resulting class probabilities are then used as the observation to update the HMM state.

HMM Matrices. To improve robustness, the HMM explicitly models the biases of the gesture set. We define a discrete HMM with a transition matrix A and an emission matrix B .

- **Emission Matrix (B):** We treat the classifier output at time step t as an observation $\mathbf{o}_t$. The emission matrix assigns a 0.95 probability to correct (diagonal) classifications, distributing the remaining 0.05 probability uniformly across all off-diagonal entries.
- **Transition Matrix (A):** The transition matrix constrains state changes to reflect physical dependencies. For example, we explicitly model the *pinch closed* state, which is only reachable upon detecting a *pinch close*. Subsequently, the *pinch open* gesture is heavily constrained so that it is primarily accessible from *pinch close* or *pinch closed*. Furthermore,

to account for variations in gesture speed, we lower the transition probability from *side tap* to itself, as side taps are typically executed much faster than swipes. The full transition matrix is provided in Table 6.

Activity Gating and Refractory Period. To prevent excessive or stuttering detections from a single physical movement, the system utilizes a rate-of-change score (RSC). The RSC accepts a gesture only when the score surpasses a predefined threshold of $t_{\text{peak}} = 2$. Finally, we impose a strict refractory period of 300 ms after a successful detection. This effectively limits the maximum theoretical detection rate to one gesture every 300 ms, ensuring clean separation between consecutive inputs.

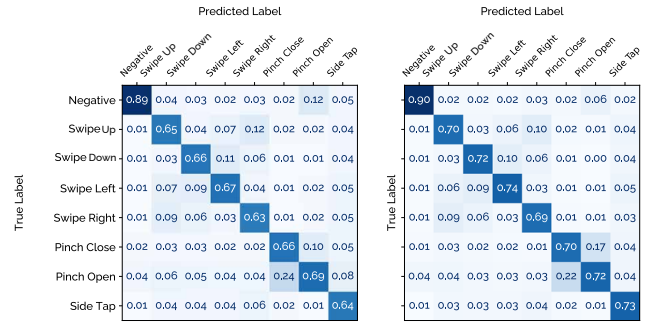


Figure 13: Online evaluation confusion matrices computed over 8 splits of continuous signal. Left: window-based model predictions. Right: HMM-decoded predictions, showing improved pinch-state consistency and fewer confusions across gestures.

E.1 HMM Filter

We describe the transition matrix for our HMM-based online filtering approach in Table 6. We model each gesture as a state and set the probability to move from one state to another state. *Pinch Closed* models the state when index finger and thumb are held together, i.e., between *Pinch Close* and *Pinch Open*.

To \ From	Negative	Swipe Fwd	Swipe Bwd	Swipe Right	Swipe Left	Pinch Close	Pinch Closed	Pinch Open	Side Tap
Negative	66.1%	5.6%	5.6%	5.6%	5.6%	0.1%	0.1%	5.1%	7.2%
Swipe Fwd	5.6%	66.1%	5.6%	5.6%	5.6%	0.1%	0.1%	5.1%	7.2%
Swipe Bwd	5.6%	5.6%	66.1%	5.6%	5.6%	0.1%	0.1%	5.1%	7.2%
Swipe Right	5.6%	5.6%	5.6%	66.1%	5.6%	0.1%	0.1%	5.1%	7.2%
Swipe Left	5.6%	5.6%	5.6%	5.6%	66.1%	0.1%	0.1%	5.1%	7.2%
Pinch Close	5.6%	5.6%	5.6%	5.6%	5.6%	59.6%	0.1%	5.1%	7.2%
Pinch Closed	0.1%	0.1%	0.1%	0.1%	0.1%	19.9%	59.6%	5.1%	0.1%
Pinch Open	0.1%	0.1%	0.1%	0.1%	0.1%	19.9%	39.7%	59.6%	0.1%
Side Tap	5.6%	5.6%	5.6%	5.6%	5.6%	0.1%	0.1%	5.1%	56.6%

Table 6: Transition probabilities for HMM filtering. Each cell $p_{i,j}$ represents the probability (%) of transitioning from state j (column) to state i (row).