BioQ: Towards Context-Aware Multi-Device Collaboration with Bio-cues

Adiba Orzikulova[†*], Diana A. Vasile[‡], Chi Ian Tang[‡], Fahim Kawsar[‡§], Sung-Ju Lee[†], Chulhong Min[‡]
[†]KAIST, Korea [‡]Nokia Bell Labs, Cambridge, UK [§]University of Glasgow, UK
{adiorz,profsj}@kaist.ac.kr
{diana-alexandra.vasile,ian.tang,fahim.kawsar,chulhong.min}@nokia-bell-labs.com

Abstract

The rapid growth of wearable devices has opened exciting opportunities for *context-aware multi-device collaboration*, where multiple devices can provide enhanced user experience tailored to user needs and conditions. However, it also presents a unique challenge of reliably determining whether a set of wearables is being used by the same individual. In real-world scenarios, device sharing, exchanging, or unintended use can cause privacy risks and degraded functionality. Existing solutions primarily rely on accelerometer data to match movement patterns across devices, but they perform poorly during stationary or varied non-repetitive activities. In this paper, we introduce BioQ, a method that unobtrusively detects wearable co-location by generating and matching *bio-cues*. These bio-cues are generated from on-body wearable sensor data and embedded into a common latent space. Furthermore, when devices share the same sensor types, BioQ can effectively integrate multiple sensor sources to improve cue generation and matching. Experimental results show that BioQ outperforms baselines in bio-cue generation and matching and is resource-effective in model training, inference, and energy use. Our code is available at https://github.com/Nokia-Bell-Labs/contextual-biological-cues.

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Computing methodologies** → **Machine learning**.

Keywords

Biological cues, multi-device collaboration, contrastive learning

1 Introduction

Recent advances in wearable technology have driven the rise of multi-device environments, where users simultaneously wear various smart devices [7, 15]. These wearables, ranging from smartwatches to smart earables, rings, and patches, offer exciting opportunities to enhance user experience through seamless and intelligent inter-device collaboration, adapting dynamically to user contexts and real-time conditions. Such *context-aware multi-device collaboration* goes beyond traditional sensor fusion [13, 19] by enabling personalized interactions and responses across devices in real-time. This approach improves context monitoring [1, 30], notification management [50], privacy [39], and energy efficiency [38].

However, realizing context-aware multi-device collaboration introduces a unique challenge: determining on the fly if a set of wearable devices is being used by the same individual. Although wearables are typically personal devices, real-world scenarios often involve sharing, exchanging, or unintended use of devices by others, leading to privacy risks and inconsistent device behavior. For example, teenagers may share earbuds to listen to music [21], family members may unknowingly use each other's wearables, and in social or public settings, devices can be easily passed between users [46]. Such scenarios can disrupt the user experience from multi-device collaboration: sensitive information may be unintentionally shared, the audio quality of stereo music may degrade when earbuds are shared, and health applications can produce inaccurate predictions by combining health data from different users.

Existing solutions to identify whether devices are worn by the same individual have largely relied on accelerometer [14, 32, 45, 53] to compare movement patterns across wearables. These approaches applied the coherence function [5, 6] to match periodic on-body device movement patterns such as walking [14, 32]. While effective in distinguishing movement patterns of different users, these methods have limited applicability, especially during varied daily activities that do not involve repetitive motion. Also, relying on periodic movements of the body is often error-prone, even when devices are worn by the same user, in scenarios such as stationary tasks or activities that involve isolated body movements, like typing or nodding. This limitation highlights the need for more adaptable methods for wearable device matching.

To address these challenges, we introduce BioQ (illustrated in Fig. 1), a method that unobtrusively verifies if multiple devices are worn by the same user by generating and matching *bio-cues*, sensor data embeddings in a shared latent space. These bio-cues are universal for an individual at a given time, regardless of device placement (provided the same sensor set is used), but distinct for different users or times. This facilitates an efficient and effective approach without requiring all devices to be placed at the same body location [32] or customized for each device pair [14]. A key reason this is possible is that vital signs such as heart rate and respiration, remain universal to an individual, irrespective of device placement. Modern wearables continuously monitor these vital signs, providing a natural source of bio-cues that fluctuate over time but remain distinctive to the same person [4, 8]. Leveraging these characteristics, BioQ generates bio-cues from physiological and motion sensing data by adopting contrastive representation learning [9]. Contrastive learning is a machine learning technique

[*] Work was done during the author's internship at Nokia Bell Labs.

SenSys '25, Irvine, CA, USA

ACM ISBN 979-8-4007-1479-5/2025/05
https://doi.org/10.1145/3715014.3722079

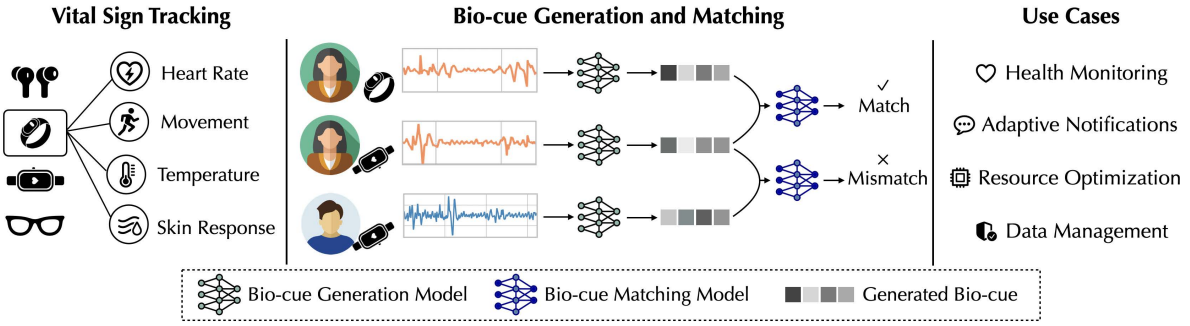


Figure 1: BioQ – Bio-cue generation and matching with use cases in context-aware multi-device collaboration.

that extracts meaningful representations by contrasting positive (similar) and negative (dissimilar) pairs of instances. In BioQ, synchronized sensor data from multiple wearables on the same user are treated as *positive* pairs, encouraging similarity. In contrast, data from different users or times serve as *negative* pairs, promoting distinction. Finally, a co-location decision is made by matching the generated bio-cues from wearable devices. While BioQ can effectively handle multiple sensor modalities (given that the devices share overlapping sensors), exploring cross-modal synergy holds great promise in future research, as we discuss in §6.

Extensive experimental results on real-world datasets show that BioQ outperforms baseline methods, achieving an average Fisher's discriminant ratio (FDR), a measure of distribution separability for device co-location detection on the same body, improvement of 1.586 in bio-cue generation, along with a 0.188 F1 score increase and a 0.151 reduction in equal error rate in bio-cue matching. In addition, BioQ exhibits low system overhead on user devices, with average inference latencies of 5.05 ms and 3.30 ms for bio-cue generation and matching, respectively, on Raspberry Pi devices. In summary, our work makes the following contributions:

- We introduce the concept of *bio-cues*, sensor embeddings generated through contrastive representation learning, which exhibit high similarity when devices are worn on the same body at a given time and clear differences when worn by different users or at different time intervals.
- We develop BioQ, a system that unobtrusively detects co-located on-body wearable devices using bio-cues. It operates without user intervention and is adaptable to leverage various physiological sensors beyond motion data.
- We show the BioQ's effectiveness through extensive experiments on real-world datasets, showcasing its superior performance over existing methods in bio-cue generation and matching.

2 Motivation

2.1 Design Rationale

Why vital signs? Mobile and wearable devices come with a wide range of sensors, including Inertial Measurement Unit (IMU), photoplethysmography (PPG), and electrodermal activity (EDA) sensors, capable of monitoring human physiology and behavior in real-time [16]. Key metrics such as heart rate, oxygen saturation, respiration rate, and blood pressure, which these sensors capture, remain consistent in the body for a particular person at a particular time, but change dynamically depending on contextual factors and activity [4, 8]. This offers a unique opportunity to generate bio-cues without complex user interactions or specialized hardware.

Why temporal dynamics? Vital signs do not remain static for the same user throughout the day due to activity or stress. To accurately determine if devices are worn by the same person, modeling the user's evolving physiological and physical conditions is essential. BioQ models temporal dynamics of vital signs to generate reliable and contextually relevant bio-cues.

Multimodal bio-cue generation: Since many wearable devices share multiple sensing modalities, BioQ can generate bio-cues by combining multiple vital signs from multimodal sensing sources (e.g., IMU + PPG) instead of relying on a single modality (e.g., PPG) in such cases. Our experimental results in §4.2.3 also show that utilizing various combinations of sensors enhances bio-cue performance. Another direction to expand BioQ's capability is to handle a potential cross-modal setting, where distinct sensor data, such as PPG from one device and motion from another, can be harmonized. We leave this as future work and will discuss its feasibility in §6.

2.2 Use Cases

With the unprecedented rise of wearable devices—earbuds, watches, rings, wristbands, headbands, patches—it is evident that we will be surrounded by more wearable devices. In such environments, we envision that these devices will increasingly collaborate with each other to enhance the multi-wearable user experience. For instance, this collaboration can recognize more diverse and complex contexts [10, 52], achieve better recognition accuracy [37, 38], and improve resource efficiency [18, 25, 31]. We believe BioQ's capability to verify device co-location unobtrusively will be a core functionality and offer new opportunities, as we showcase below.

Robust health monitoring. Many users rely on multiple wearables—a smartwatch, a fitness tracker, or a heart rate monitor—to track their health and activity throughout the day. For example, a user might wear a smartwatch to monitor daily steps and heart rate while occasionally using a chest strap or a ring sensor for more precise tracking during workouts or sleep. BioQ can enable reliable multi-device sensor fusion by ensuring that all data streams originate from the same person. This capability helps reduce errors that could arise if a device was mistakenly worn by someone else. Furthermore, when inconsistencies between devices are detected, such

as deviations from the user's baseline health metrics, the system can alert the user, pause logging, or trigger predefined responses.

Adaptive notification management. People rely on receiving notifications across multiple devices, including smartphones, smartwatches, and earables. BioQ can help applications determine when the same individual wears multiple devices, allowing more intelligent notification management. For instance, smart earbuds that automatically read out incoming messages can use BioQ to verify that they are still worn by their original user. If the earbuds are shared with someone else—for example, during a commute—this verification can enable applications to immediately mute confidential notifications, preventing unintended exposure. Similarly, in professional settings, such as shared workspaces, BioQ can support applications that temporarily pause or redirect sensitive alerts when a device is borrowed by someone else, ensuring privacy and contextual awareness.

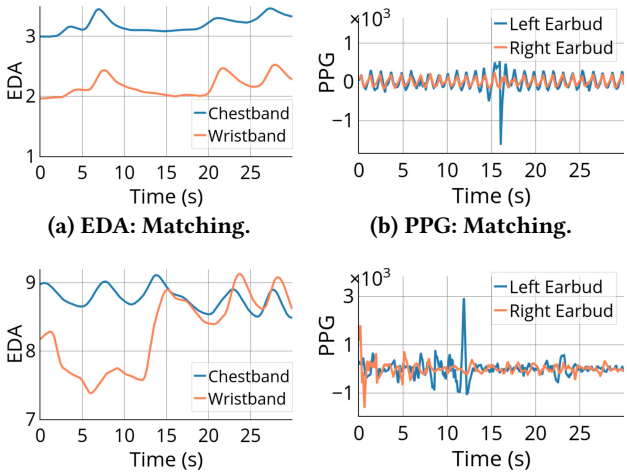
Privacy and data management. As wearable devices become more capable, they collect and store increasingly sensitive data, from daily step counts to detailed health recordings. BioQ can enable applications that require strong identity verification to ensure data integrity and privacy. For example, when a health insurance platform integrates data from multiple personal devices for activity-based premium adjustments, BioQ can help prevent incorrect data attribution by detecting mismatches in bio-cues that indicate device swaps or unintended sharing. When such discrepancies are identified, applications can choose to halt new data aggregation or trigger re-verification mechanisms, ensuring that health sensing insights remain accurate and trustworthy.

Energy efficiency and resource optimization. Many users wear multiple wearables with overlapping sensors (e.g., multiple accelerometers, microphones, and heart rate monitors). BioQ can enable more effective coordination among these devices by confirming that they are worn by the same individual. With this capability, applications can make smarter decisions, such as adjusting sensor sampling rates or offloading tasks to the most power-efficient device. For instance, a smartwatch could delegate GPS tracking to a phone if bio-cues verify they are co-located on the same body, reducing overall power draw and balancing battery usage across connected devices. This capability can support more sustainable operations and extend the longevity of wearable systems.

2.3 Challenges in Generating Bio-cues

To present challenges in a bio-cue generation, we conducted experiments on physiological sensor data using two conventional methods: direct comparison of clean, filtered data (§2.3.1) and analysis of engineered sensor features (§2.3.2).

2.3.1 Bio-cues from Filtered Sensor Data. A simple and intuitive approach to generate bio-cues is using clean, filtered data from physiological sensors. To assess its feasibility, we analyzed EDA and PPG data from WESAD [44] and FatigueSet [24] datasets, respectively. Details about datasets, pre-processing, and filtering methods used to clean the data are provided in §4.1. We use Spearman's rank correlation [54] to measure how similar or different the clean sensor data are from devices worn by the same user. The coefficient, ρ, ranges from -1 to 1, where 1 indicates perfect positive correlation, -1 indicates perfect negative correlation, and values near 0 suggest no correlation.



(a) EDA: Matching. (b) PPG: Matching.

(c) EDA: Mismatching. (d) PPG: Mismatching.

Figure 2: EDA on chestband and wristband in (a) and (c), and PPG on left and right earbuds in (b) and (d).

Our analysis, shown in Fig. 2, reveals intriguing patterns. For example, a high correlation of ρ=0.86 between EDA data from a chestband and wristband (Fig. 2a) and a correlation of ρ=0.68 between PPG data from left and right earbuds (Fig. 2b) on the same user can be observed. However, these correlations are not always consistent; in Fig. 2 (c) and (d), raw sensor data from the same user at the same time exhibit discrepancies, with low correlations of ρ= -0.02 and ρ=-0.08. These findings suggest that relying solely on sensor values, even after filtering, presents challenges for generating reliable and consistent bio-cues.

2.3.2 Bio-cues from Hand-crafted Features. The inherent issue with raw or even filtered data is its vulnerability to device placement, which can significantly degrade the quality of generated cues. As an improved strategy, hand-crafted, sensor-specific features can be considered. These features are designed to represent vital signs, which should remain consistent across different device placements for the same individual during similar time frames. We conducted experiments using EDA data from chestbands and wristbands in the WESAD dataset to explore this. Specifically, we computed hand-crafted features after separating the filtered EDA signal into Skin Conductance Level (SCL) and Skin Conductance Response (SCR) components, a common feature engineering approach [44]. Following prior work [44], we extracted the mean and standard deviation (std) from SCR and computed the mean, std, number of peaks, and mean peak amplitude from SCL using a 30-second window with no overlap. We calculated the correlation between extracted features, similar to the experiments in §2.3.1.

Table 1 presents the results for different subjects across three emotional contexts: neutral, stress, and amusement. The **highest** and second-highest correlation coefficient (ρ) values are **bolded** and underlined, respectively. The correlation values show high variability in feature effectiveness across individuals and emotional

Table 1: Correlation coefficient (ρ) of bio-cues from hand-crafted EDA features on subjects' chestband and wristband across three emotional states: neutral, stress, and amusement. SCR: Skin Conductance Response, SCL: Skin Conductance Level.

Subject ID	S-02			S-03			S-04		
Feature / State	Neutral	Stress	Amusement	Neutral	Stress	Amusement	Neutral	Stress	Amusement
SCR Mean	-0.038	0.639	-0.189	0.066	0.321	-0.063	0.051	0.481	0.035
SCR Std	0.398	0.598	**0.490**	0.612	**0.896**	**0.664**	**0.438**	0.683	**0.490**
SCL Mean	**0.898**	**0.908**	-0.343	0.803	0.481	0.552	-0.825	-0.066	0.245
SCL Std	0.801	0.362	-0.147	**0.860**	0.657	0.378	0.270	0.200	-0.007
SCR Num Peaks	0.197	0.801	0.319	0.407	0.761	0.419	0.116	0.597	0.373
SCR Mean Peak Amp.	-0.019	0.624	0.000	0.650	0.661	0.545	0.225	**0.712**	0.245

contexts. No singular feature or combination consistently correlated across all users and emotional states, highlighting the challenge of generating reliable bio-cues even with hand-crafted features.

2.3.3 Bio-cues from Motion Data. A range of studies [14, 32] leverage human gait patterns to detect device co-location on the same body by analyzing the coherence [5, 6] between accelerometer signals collected during walking. However, relying solely on accelerometer data can be noisy, especially when devices are positioned at different body locations (e.g., wrist vs. head) [32]. Additionally, these methods are limited to walking activities and fail when the user is stationary or engaged in non-walking activities, which is more common in real life. Thus, there is a need for a more robust approach that can integrate various sensing modalities and support diverse activities beyond motion.

3 Method

We explain the overall operational flow of the system and outline the key designs of our bio-cue modeling framework (§3.1), with formal definitions and implementation details given in (§3.2) and (§3.3), respectively. Before detailing the algorithm, we present an overview of the operational flow of `BioQ`, illustrating how the system functions in a real-world setting. As depicted in Fig. 3, the process is divided into three stages:

- **Pre-deployment** ①: Global bio-cue generation and matching models are trained before deployment. The bio-cue generation model learns to extract accurate bio-cues by analyzing vital sign data from various devices and users, building a strong foundation model. The matching model is then trained to predict whether two bio-cues originated simultaneously from the same user.
- **Upon deployment** ②: When a new user is added, both bio-cue generation and matching models can be retrained. Positive and negative non-aligned pairs are gathered from the new user, while negative pairs from existing users in the global dataset further enrich the candidate pool.
- **At runtime** ③: Finally, `BioQ` generates and matches bio-cues as needed, either upon request or as required, to support context-aware multi-device collaboration (§2.2) in real-time.

At runtime, when a matching request is made, bio-cues are generated on the device where the sensor data is collected, and delivered to a device in which the matching operation is executed, e.g., a smartphone or one of the matching devices depending on the service scenarios. However, when the processing capability of the wearables is not sufficient, bio-cue generation can be offloaded to a more powerful and trusted device, such as the user's smartphone. Failure scenarios—such as mismatched bio-cues—could further be mitigated through periodic recalibration to maintain robust performance in dynamic settings.

3.1 Key Designs of `BioQ`

Following the discussion in §2.3, we highlight the key designs in our bio-cue generation and matching framework aimed at addressing different challenges.

Key objectives. With bio-cue modeling, our key objective is to develop a model that identifies if pairs of data samples are matching (i.e., from the same user at the same time) or non-matching (i.e., from different users or different times). To achieve this, we need an effective feature extractor (bio-cue generator) that outputs distinct features from different data samples and a matching algorithm (bio-cue matching) to perform the actual matching. It is important to note that the `BioQ`'s matching is a binary classification, i.e., it verifies whether devices are co-located on the same user or not. User identification can be achieved by collaborating with capable devices, e.g., FaceID on a smartphone or passcode on a smartwatch.

Contrastive learning for bio-cue generation. Among representation learning techniques from sensor data, including autoencoders [28, 33], Siamese networks [11], and contrastive learning [9, 20, 22, 40, 49], we choose contrastive representation learning as it naturally aligns with our objective: extracting similar representations for *positive pairs* (time-aligned samples from devices worn by the same user), and dissimilar ones for *negative pairs* (samples from devices on different users or different time intervals).

Unlike conventional contrastive representation learning (e.g., SimCLR [9], MoCo [20]) that relies on manually designed data augmentations of an anchor sample (e.g., random cropping or color jitter for images), `BioQ` leverages *multi-wearable settings* to generate positive and negative samples. Since each device provides a genuine real-time view of the user's physiological or physical status for positive samples, `BioQ` uses *naturally aligned* sensor signals from different on-body devices as positive pairs. It does not need synthetic transformations and makes the positive pairs more organically representative of the target task. We also define negative pairs more strategically by incorporating signals not only from different users but also from different time windows of the same user, ensuring that `BioQ` learns to separate subtle but important fluctuations—such as normal shifts in heart rate or respiration—occurring over the day. Furthermore, we implement an *informative negative sampling* strategy, discarding those negatives that are trivially close or far in embedding space and instead focusing on "moderately close" negatives that provide richer training signals. These design choices move beyond merely adopting an existing contrastive framework:

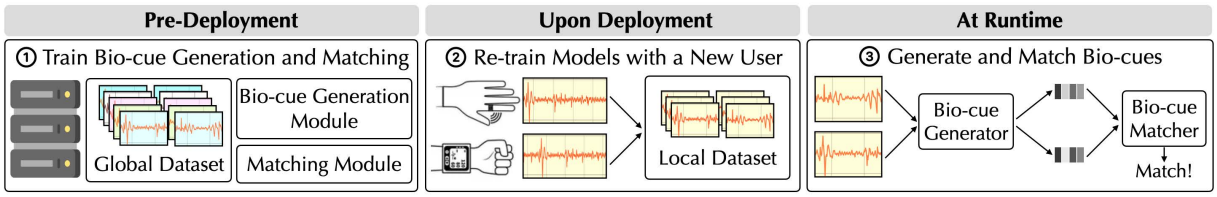


Figure 3: Operational flow of BioQ.

BioQ adapts contrastive principles to the unique demands of multi-device physiological data, enabling robust co-location detection even under varied user conditions and activities.

Naturally-occurring positive pairs. Defining positive and negative samples is central to any contrastive learning framework since it directly encodes the learning objective. Different from conventional contrastive representation learning frameworks where positive pairs are formed of augmented data pairs [9], which require carefully selected augmentation functions, we leverage the unique observation that multiple devices capture the physiological and physical status of the user from different viewpoints and form positive pairs that are effective for our target task. This allows models to generate effective bio-cues without substantial training data.

Informative negative pairs. Negative samples are often easier to define as virtually any samples that are not used in forming positive pairs with the anchor can be used, e.g., by varying the user set, device set, and time window. However, selecting *all possible* negative pairs for learning is often not feasible because the majority do not contribute significantly to the learning process, in which the model can distinguish the easy negatives. Therefore, a sampling strategy prioritizing quality over quantity is needed, ensuring that negative samples contribute effectively to the contrastive learning process. To achieve this, we introduce an informative negative sampling strategy that selects negatives based on their relative distance from the anchor sample so that the samples in the 'Goldilocks zone' are selected, forming a balanced set of positive and negative samples that contributes to effective training. This approach aligns with established principles in negative mining for optimizing contrastive learning outcomes [43, 55, 56].

Cue matching. Although contrastive learning can train our models to extract effective features as bio-cues, the final matching task is not fully encoded in the learning objective: contrastive learning is analogous to the clustering task [48], where there might exist multiple clusters for one particular class (a user in our case). This is one of the reasons why contrastive learning is a powerful self-supervised learning framework, but further modeling would be beneficial for the final task of user matching. Therefore, instead of using a simple similarity measure, we propose including a dedicated cue-matching model trained on pairs of aligned bio-cues from pairs of data samples to boost the performance in classifying if pairs are matching or non-matching, achieving our key objectives.

Both bio-cue generation and matching are device-agnostic and do not require any customization for each unique pair of devices as long as they contain the same set of sensors. This design facilitates an efficient and effective approach without the need for all devices to be located at a similar position on the body [14, 32] or requiring customization for each device pair [14].

3.2 Formal Definitions

In our proposal, we present the exact definitions of different components. Complete pseudocode is provided in Algorithm 1.

3.2.1 Problem Statement. We define the problem of bio-cue generation as follows. Given training data, let $\mathbb{U}$ represent a set of users, $\mathbb{D}$ a set of devices, $\mathbb{C}$ a set of sensor channels or vital signs, $\mathbb{T}$ a set of time windows, and $\mathbb{X} = \{x_t^{u,d,c} | u \in \mathbb{U}, d \in \mathbb{D}, c \in \mathbb{C}, t \in \mathbb{T}\}$ the dataset. Each sample $x_t^{u,d,c}$ in $\mathbb{X}$ is a sensor time series with length l_d and s_d channels, specific to device d, i.e. $x_t^{u,d,c} \in \mathbb{R}^{l_d \times s_d}$. For each user $u \in \mathbb{U}$, we assume that time-aligned data from multiple devices are available for the same sensor channels. Specifically,

$$x_t^{u,d,c} \in \mathbb{X} \wedge x_t^{u,d',c} \in \mathbb{X} \wedge d \neq d'.$$

We aim to define a bio-cue generation model, f, that projects raw sensor streams from the same user close together in the embedding space while positioning data from different users or times farther apart. We formalize this with the following constraint:

$$\delta(f(x_t^{u,d,c}), f(x_t^{u,d',c})) + m < \delta(f(x_t^{u,d,c}), f(x_{t'}^{u',d',c})), \quad (1)$$

where $d \neq d'$ and ($u \neq u'$ or $t \neq t'$). Here, $\delta(a, b) = |a - b|$ denotes the L_1 norm, used as the distance metric between two bio-cues, and m is the margin between positive (similar) and negative (distinct) pairs. Additionally, to detect whether the devices are worn on the same body at a given time, we train a bio-cue matching model g:

$$g(f(x_t^{u,d,c}), f(x_{t'}^{u',d',c})) = \begin{cases} 1 & \text{if } u = u' \wedge t = t' \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

3.2.2 Bio-cue Generation. Figure 4 provides an overview of the bio-cue generation process using contrastive learning. Formally, we define the set of positive pairs $\mathbb{P}$ as $\mathbb{P} = \{(x_t^{u,d,c}, x_t^{u,d',c}) | d \neq d'\}$ (i.e., samples collected simultaneously across different devices worn on the same body). For negative pairs, we select samples that either originate from different users or differ in time if from the same user. In contrast, negative pairs are selected based on two key criteria: A sample qualifies as a negative candidate if it (a) either originates from a different user ($u \neq u'$), or (b) if from the same user ($u = u'$), corresponds to a different timestamp ($t \neq t'$). While the device origin must differ from that of the anchor sample ($d \neq d'$), the channel c remains consistent, ensuring the comparison involves the same sensor type. To select informative negative candidates $\mathbb{N}_{\text{inform}}$ for each anchor sample $x_t^{u,d,c}$ that are effective for learning, we define the following constraint:

$$\mathbb{N}_{\text{inform}} = \left\{ x_{t'}^{u',d',c} \;\middle|\; \delta^{\min} \leq \delta(x^{u,d,c}, x_{t'}^{u',d',c}) < \delta^{\max} \right\},$$

where δ represents the $L1$ distance between the anchor and candidate negative samples. To ensure a balanced selection, $\delta^{\min}$ and

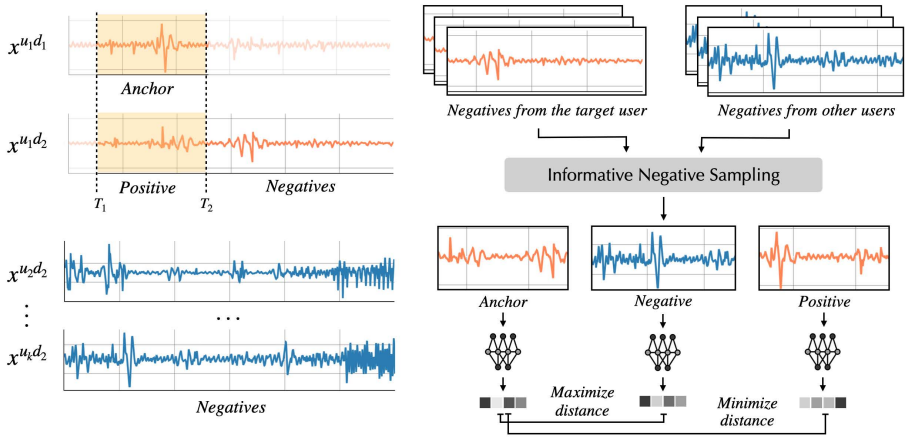


Figure 4: Bio-cue generation: positive and negative pairs establishment (left), informative negative sampling (right).

$\delta^{\max}$ exclude a percentage of the closest and farthest negatives, retaining only the middle portion of the distance distribution for sampling. We empirically define this percentage as 10% for both δ^{min} and δ^{max}. To further promote diversity in negative sampling, we introduce a probability parameter that increases the likelihood of selecting negatives from different users. This parameter is set to 0.8, ensuring that negatives are drawn from different users in 80% of cases. Finally, the bio-cue generation process is guided by a contrastive loss function to encourage separation between positive and negative pairs as follows:

$$\mathcal{L}_{con} = \frac{1}{N}\sum_{i=1}^{N} \max\Big(0, \delta(f(x_{t_i}^{u_i,d_i,c}), f(x_{t_i}^{u_i,d_i',c})) - \delta(f(x_{t_i}^{u_i,d_i,c}), f(x_{t_i'}^{u_i',d_i',c})) + m\Big), \quad (3)$$

where $\delta(f(x_t^{u,d,c}), f(x_t^{u,d',c}))$ is the distance between the anchor and positive cues, $\delta(f(x_t^{u,d,c}), f(x_{t'}^{u',d',c}))$ is the distance between the anchor and negative cues, and m is the margin, a hyperparameter that defines the minimum desired separation between positive and negative pairs. By minimizing this loss, the model learns to cluster similar bio-cues while ensuring sufficient separation from dissimilar ones, thereby enhancing the robustness of the learned representations.

3.2.3 Bio-cue Matching. To train the bio-cue matching model, we leverage the bio-cue generation model, f, that extracts bio-cues for anchor, positive, and negative samples, following the definition given above. The cue-matching model is trained to classify bio-cue pairs as matching or non-matching. To achieve this, we construct matching pairs $[f(x_t^{u,d,c}), f(x_t^{u,d',c}) \mid d \neq d']$ along with non-matching pairs $[f(x_t^{u,d,c}), f(x_{t'}^{u',d',c})] \mid d \neq d'$ and $(u \neq u'$ or $t \neq t')$, which are then fed to the matching model as concatenated vectors. Given a batch of such pairs, we define

$$\mathbf{X}_{\text{match}} = \{[f(x_i), f(x_j)] \mid (x_i, x_j) \in \mathcal{M} \cup \mathcal{N}\},$$

where $\mathcal{M}$ is the set of matching pairs, and $\mathcal{N}$ is the set of non-matching pairs. Their corresponding labels are given by $\mathbf{Y}_{\text{match}} = \{1, 0\}$, where 1 indicates a matching pair and 0 indicates a non-matching pair. The training objective for the matching model is to minimize the binary cross-entropy loss:

$$\mathcal{L}_{\text{match}} = -\frac{1}{N}\sum_{i=1}^{N}\Big(Y_{\text{match}}^{(i)} \log(p^{(i)}) + (1 - Y_{\text{match}}^{(i)}) \log(1 - p^{(i)})\Big), \quad (4)$$

where $p^{(i)}$ denotes the model's predicted probability that the i-th pair matches, and $Y_{\text{match}}^{(i)}$ is the true label for the pair. By minimizing this loss, the model learns to distinguish between similar and dissimilar bio-cues, refining its ability to classify whether two given samples are from the same contexts.

3.3 Implementation

We train the bio-cue generation model using an encoder with three linear layers and a projection head with two linear layers. Each linear layer, except the last, is followed by a ReLU activation. We also incorporate a 10% dropout to enhance generalizability. The matching model consists of four linear layers, each followed by ReLU activations except the last one. Similar to the bio-cue generation model, the matching model also uses a dropout layer. The bio-cue generation and matching models are trained for 400 and 200 epochs, respectively. Training hyperparameters are consistent across bio-cue generation and matching models and are set to a batch size of 64, a learning rate of 0.001, a weight decay of 0.001, and trained using the Adam optimizer. The bio-cue embedding dimension is set to 16. All experiments are conducted in Python 3.9 using the PyTorch [41] framework.

4 Evaluation

We compare the performance of `BioQ` and baselines in bio-cue generation (§4.2-§4.4) and matching (§4.3-§4.4) across various settings. Additionally, we analyze the system costs of `BioQ` in training and inference latency, CPU usage, and energy consumption (§4.6).

4.1 Experimental Setup

We conducted experiments using two publicly available multimodal, multi-device datasets (§4.1.1), comparing the performance of `BioQ` against four baselines (§4.1.2). A within-subject train-test split was applied, with 70% of the data used for training and the remaining 30% for testing. In both datasets, we only consider valid users with at least 15 samples per label (i.e., the original category for

Algorithm 1 BioQ Algorithm

Input: Dataset $\mathbb{X} = \{x_t^{u,d,c}\}$ with samples indexed by user u, device d, channel c, and time t, training epochs E, M for bio-cue generation and matching.
Bio-cue Generation:
  Initialize bio-cue generation model f
  for epoch $e = 1, \ldots, E$ **do**
    $B \leftarrow$ split $\mathbb{X}$ into batches of size $|B|$
    for batch $b \in B$ **do**
      $b_{anc}, b_{pos} \leftarrow$ Get anchor and positive pairs
      $b_{neg} \leftarrow$ Sample informative negatives
      $\mathcal{L}_{con} \leftarrow$ Compute contrastive loss
      $f \leftarrow$ Update using $\texttt{SGD}(f, \mathcal{L}_{con})$
    end for
  end for
Bio-cue Matching:
  Initialize bio-cue matching model g
  for epoch $m = 1, \ldots, M$ **do**
    $B \leftarrow$ split $\mathbb{X}$ into batches of size $|B|$
    for batch $b \in B$ **do**
      $b_{anc}, b_{pos}, b_{neg} \leftarrow$ Sample data for matching
      $b^e_{anc}, b^e_{pos}, b^e_{neg} \leftarrow$ Generate cues using f
      $Y_{match} \leftarrow$ Assign matching labels
      $\mathcal{L}_{match} \leftarrow$ Compute binary cross-entropy
      $g \leftarrow$ Update using $\texttt{SGD}(g, \mathcal{L}_{match})$
    end for
  end for

datasets, such as stress, amusement for WESAD and activity, fatigue for FatigueSet). We ran each experiment with five different seeds, reporting the mean and standard deviation in the results (§4).

4.1.1 Datasets. **FatigueSet** [24] is a dataset for exploring the impact of physical activity on mental fatigue. Data was gathered from 12 users wearing earbuds, a headband, a chestband, and a wristband, capturing physiological and motion data. We focus on accelerometer (ACC), gyroscope (GYR), and photoplethysmography (PPG) data, which are shared by at least two devices. We employed established noise-filtering techniques to mitigate the inherent noise in the physiological and motion-sensing data. For ACC, a low-pass filter with a 15 Hz cutoff frequency is applied, as 99% of human body motion is contained below 15 Hz [26], and magnitude from three axes was used. Similarly, for GYR, we applied a low-pass filter with a 20 Hz cutoff frequency to reduce high-frequency noise. The PPG is processed using a 3rd order Butterworth filter with a 0.5 Hz low-cut and 8 Hz high-cut. Noise filtering is implemented leveraging an open-source Python package library, Neurokit2 [34, 35].

WESAD [44] dataset is created for detecting stress and effects using wearable sensors. Collected from 15 participants wearing a chestband and a wristband, it includes physiological and motion data such as Electrodermal Activity (EDA), Skin Temperature (TEM), Blood Volume Pressure (BVP), and Accelerometer (ACC). Our analysis targets the shared sensors on the wrist and chest: ACC, EDA, and TEM. Following the authors of the WESAD dataset [44], we opted not to apply filtering to the temperature data. For EDA, a low-pass filter with a 5 Hz cutoff frequency and a 4th-order Butterworth filter was applied. For ACC, we applied the same preprocessing technique we used in the FatigueSet dataset.

Table 2: List of sensor-specific features.

Sensor	Features
PPG	min, max, mean, std, range, beats per minute, interbeat interval, std of normal-to-normal intervals, std of successive differences, area of Poincare plot, proportion of NN50 and NN20, breathing rate, heart rate median absolute deviation, root mean square of SDD std 1 and 2 from Poincare plot analysis, their ratio
EDA	min, max, mean, std, range, slope, SCL (skin conductance level) mean, SCL std, SCR (skin conductance response) mean, SCR std, SCR num peaks, SCR num peak amplitude
TEM	min, max, median, std, range, slope
ACC, GYR	min, max, mean, std, median

Following prior work [14, 32], we use an 8-second window for ACC data. Similarly, we apply an 8-second window for GYR data and a 20-second window for physiological sensing data (such as EDA or PPG, and multimodal combinations involving these sensors), as a window size of 20~60 seconds is effective for feature extraction [29].

4.1.2 Baselines. We consider the following as baselines:

- **Raw** treats clean, noise-filtered sensor data as bio-cues. We compute the L1 distance between bio-cues to assess the similarity for synchronized samples. For multimodal sensor streams, we average distance values from all sensor types. Note that the values are normalized to a 0~1 range for consistency. This represents a naïve baseline and acts as a lower bound for comparison.
- **Feature** uses the most representative hand-crafted features for each sensor type [34, 44, 51] as bio-cues. A full list of features used for each sensor type is provided in Table 2. Similar to Raw, the Feature method uses L1 distance and normalizes values before averaging distances across multimodal data streams.
- **Lester et al. [32]** applies the coherence [5, 6] function to noise-filtered accelerometer data to determine whether devices are worn on the same body. In this approach, coherence values serve as a measure of bio-cue similarity specifically for accelerometer data. We evaluated it solely for the accelerometer data. As this method is exclusively designed for accelerometer signals, we implemented and evaluated it solely for this data type.
- **Cornelius et al. [14]** use a feature-based method, extracting hand-crafted features from accelerometer data and computing coherence between the features. The method was developed for matching, so we use averaged coherence similarity values to derive a single metric to compare bio-cue generation performance.

Of the four baselines, Cornelius et al. [14] use Support Vector Machines (SVM) [3] trained on coherence features to determine if devices are worn on the same body. They also extend Lester et al. [32] by incorporating SVMs to dynamically estimate a threshold. Based on these approaches, we implemented matching baselines. We trained SVMs on the calculated distance values for both the Raw and Feature methods.

4.1.3 Evaluation Metrics. To assess bio-cue generation performance, we evaluate how consistently cues align within the same user and how distinctly they separate across different users or time intervals.

For this purpose, we use the Fisher Discriminant Ratio (FDR), a metric that quantifies the separability of two distributions by comparing their means and variances [17]. A higher FDR indicates better distinguishability, with more significant mean differences and smaller variances leading to improved bio-cue separability. Generally, FDR values above 1 indicate strong separation, while values over 0.5 suggest moderate separation. We use the macro F1 score and Equal Error Rate (EER) for bio-cue matching performance. The F1 score balances precision and recall, while EER indicates the threshold where false acceptance and rejection rates are equal. Together, they provide a comprehensive evaluation of matching performance.

4.2 Bio-cue Generation

An ideal system should generate similar bio-cues only when sensor data originates from the same user simultaneously. In all other cases, such as data from different users or from the same user at different times, the generated bio-cues should be distinct and separable (§3.2.1). To evaluate this, we conduct evaluations under two scenarios: (i) distinguishing between data from the same user at the same time versus different users (§4.2.1); and (ii) between the same user at the same time versus at different times (§4.2.2).

Overall, when comparing to state-of-the-art methods in bio-cue generation using accelerometer data (unimodal), the FDR of BioQ is on average 3.60× (in geometric mean, minimum 2.01×, maximum 13.12×, geometric std 2.39×) of the best-performing baseline across different settings. In multimodal settings (using all available data modalities), an average FDR of 9.30× of other baselines can be observed (minimum 8.42×, maximum 9.71×, geometric std 1.07×).

4.2.1 Bio-cue Generation Across Users. Table 3 presents the performance of bio-cue generation across different users, using the FDR (mean±std) metric. FDR quantifies the separability of bio-cue distributions between the same user and different users, referred to as 'across users' for brevity. The best results are highlighted in bold. Since the baselines by Lester et al.[32] and Cornelius et al.[14] are compatible only with ACC data, we limit comparisons for these models to ACC data in both datasets. BioQ consistently outperforms baselines on both datasets, achieving mean FDR scores of 0.445 and 3.185, highlighting the effectiveness of its contrastive-based bio-cue generation. By contrast, Lester et al. [32] and Cornelius et al. [14], which only support ACC data, perform variably but remain lower overall, with Raw showing the weakest performance, indicating that noise-filtered data alone lacks sufficient bio-cue quality. Moreover, BioQ maintains top performance across all other sensor types and in the multimodal setting, underscoring its robust and reliable bio-cue generation across diverse sensor sources.

4.2.2 Bio-cue Generation for the Same User Across Time. We evaluate the performance of BioQ and baselines in bio-cue generation for the same user across time, a critical requirement for data integrity and reliable health service provision. Table 4 shows the results that follow a similar trend to those in Table 3. In both the WESAD and FatigueSet datasets, BioQ achieves higher FDR values across all settings, demonstrating reliable bio-cue distribution separation between data collected simultaneously and data collected at different times. For instance, in the WESAD dataset, BioQ achieves an FDR of 0.458 in EDA and 0.965 in the combined All modality. The

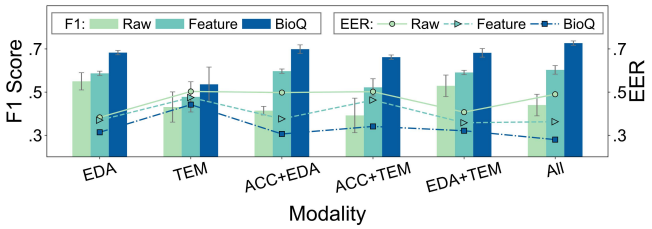


Figure 5: Bio-cue matching results in WESAD.

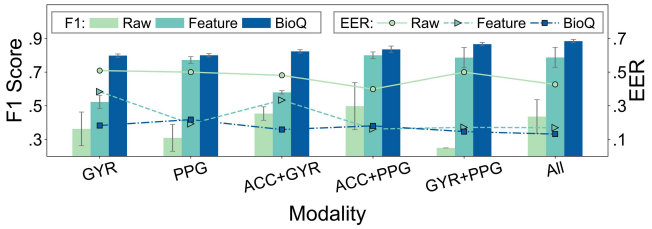


Figure 6: Bio-cue matching results in FatigueSet.

relatively high FDR values in EDA and the multimodal configuration suggest that these signals capture distinctive temporal features that enhance separation over different time intervals. Similarly, in the FatigueSet dataset, BioQ outperforms the baselines across all modalities, achieving FDR values of 1.906 in PPG and 4.093 in the All modality. The high FDR in the combined All modality setting underscores the advantage of integrating multiple modalities for enhanced temporal bio-cue generation, highlighting BioQ's effectiveness in maintaining user-specific data integrity over time.

4.2.3 Bio-cue Generation with Various Sensor Combinations. We also evaluate scenarios where devices utilize various combinations of available sensors to generate bio-cues. Tables 5 and 6 show results for bio-cue generation across users. Since Lester et al. [32] and Cornelius et al. [14] are limited to accelerometer data, we consider Raw and Feature baselines in this analysis. The results consistently show that BioQ outperforms all baselines across every sensor combination.

Our findings indicate that multimodal (i.e., ACC+GYR+PPG in FatigueSet and ACC+EDA+TEM in WESAD) approaches generally outperform unimodal ones. However, in unimodal settings, no single modality consistently excels across all scenarios. This variability is likely due to specific conditions under which each dataset was collected, such as mental/physical fatigue levels, or physiological states. For example, in the WESAD dataset—where participants remained mostly stationary while their affective states were measured—EDA showed the highest performance in both across-user and across-time evaluations. In contrast, in the FatigueSet dataset, which involves both physical and mental fatigue, ACC was most effective for distinguishing across users, while PPG was better in capturing temporal variations within individuals.

4.3 Bio-cue Matching Performance

Next, we present the bio-cue matching results. Fig. 5 and 6 display the results for all sensor types, including various sensor combinations. In the figures, the left y-axis represents the macro F1 score, shown as bar graphs, while the right y-axis represents the Equal

Table 3: Bio-cue generation results (FDR) across users. 'All' includes all available sensor types. Relative performance is calculated as a ratio between BioQ and the best-performing baseline.

Dataset	WESAD				FatigueSet			
Method / Modal	ACC	EDA	TEM	All	ACC	GYR	PPG	All
Raw	0.000±0.00	0.184±0.09	0.015±0.02	0.027±0.03	0.001±0.00	0.023±0.02	0.001±0.00	0.001±0.00
Feature	0.014±0.01	0.201±0.04	0.013±0.01	0.117±0.03	0.243±0.03	0.152±0.02	0.300±0.04	0.495±0.07
Lester et al. [32]	0.189±0.00	-	-	-	0.077±0.01	-	-	-
Cornelius et al. [14]	0.196±0.01	-	-	-	0.052±0.01	-	-	-
BioQ	**0.445**±0.02	**0.652**±0.10	**0.048**±0.03	**1.133**±0.07	**3.185**±0.22	**2.699**±0.21	**2.242**±0.20	**4.793**±0.14
Relative Performance	2.27×	3.25×	3.16×	9.71×	13.12×	17.80×	7.48×	9.67×

Table 4: Bio-cue generation results (FDR) for the same user across time. 'All' includes all available sensor types. Relative performance is calculated as a ratio between BioQ and the best-performing baseline.

Dataset	WESAD				FatigueSet			
Method / Modal	ACC	EDA	TEM	All	ACC	GYR	PPG	All
Raw	0.000±0.00	0.178±0.09	0.016±0.02	0.028±0.02	0.001±0.00	0.025±0.02	0.001±0.00	0.002±0.00
Feature	0.016±0.02	0.198±0.04	0.015±0.01	0.115±0.02	0.229±0.02	0.146±0.02	0.235±0.06	0.432±0.05
Lester et al. [32]	0.157±0.02	-	-	-	0.075±0.01	-	-	-
Cornelius et al. [14]	0.165±0.01	-	-	-	0.051±0.01	-	-	-
BioQ	**0.332**±0.04	**0.458**±0.04	**0.039**±0.02	**0.965**±0.05	**0.641**±0.02	**1.069**±0.07	**1.906**±0.23	**4.093**±0.38
Relative Performance	2.01×	2.31×	2.42×	8.42×	2.80×	7.30×	8.11×	9.48×

Table 5: Bio-cue generation results (FDR) across users with various sensor combinations.

Dataset	Modalities	Raw	Feature	BioQ
WESAD	ACC + EDA	0.088±0.04	0.175±0.03	**0.850**±0.11
	ACC + TEM	0.011±0.01	0.001±0.00	**0.423**±0.08
	EDA + TEM	0.036±0.03	0.124±0.05	**0.713**±0.05
FatigueSet	ACC + GYR	0.003±0.00	0.263±0.03	**4.297**±0.27
	ACC + PPG	0.001±0.00	0.457±0.06	**3.142**±0.52
	GYR + PPG	0.003±0.00	0.435±0.06	**4.400**±0.28

Table 6: Bio-cue generation results (FDR) for the same user across time with various sensor combinations.

Dataset	Modalities	Raw	Feature	BioQ
WESAD	ACC + EDA	0.090±0.04	0.182±0.02	**0.666**±0.07
	ACC + TEM	0.010±0.01	0.001±0.00	**0.382**±0.06
	EDA + TEM	0.034±0.02	0.116±0.04	**0.666**±0.10
FatigueSet	ACC + GYR	0.003±0.00	0.252±0.02	**1.178**±0.06
	ACC + PPG	0.002±0.00	0.382±0.06	**2.673**±0.28
	GYR + PPG	0.003±0.00	0.371±0.05	**3.608**±0.36

Error Rate (EER), depicted as line graphs. The matching results highlight the performance of BioQ against Raw and Feature baselines across different sensor types. Overall, average improvements of .097 and .132 in F1 scores are observed across all different choices of sensor types and their combinations in WESAD and FatigueSet, respectively. Similarly, average reductions of .068 and .079 in EER are observed in WESAD and FatigueSet, respectively.

Table 7 presents the matching performance for accelerometer data, including the baselines by Lester et al.[32] and Cornelius[14]. BioQ significantly outperforms all baseline methods, with improved F1 scores of .661 and .759 and reduced EER of .323 and .203 in WESAD and FatigueSet, respectively. Interestingly, BioQ also outperforms Lester et al.[32] and Cornelius et al.[14], which are specially designed for device co-location detection using accelerometer data, with F1 score higher by 0.069 and 0.180 in WESAD and FatigueSet. This is because their methods are limited to walking activity, making them less effective in diverse settings that include static and other non-walking activities. In contrast, BioQ is designed to handle both static and dynamic activities, enabling it to consistently surpass these specialized approaches.

Table 7: Bio-cue matching results with ACC. (Higher F1 and lower EER are better)

Dataset	WESAD		FatigueSet	
Metric	F1 ↑	EER ↓	F1 ↑	EER ↓
Raw	.464±.05	.498±.00	.510±.04	.471±.02
Feature	.508±.03	.450±.02	.554±.01	.363±.01
Lester et al. [32]	.585±.01	.399±.01	.590±.00	.412±.00
Cornelius et al. [14]	.592±.01	.397±.01	.574±.00	.427±.00
BioQ	**.661**±.01	**.323**±.01	**.759**±.00	**.203**±.00
Difference	+0.069	-0.074	+0.170	-0.160

4.4 Device-specific Analysis

The previous evaluation focused on scenarios where all devices participate in bio-cue generation. In real-world use cases, a wide range of device combinations is possible, therefore we also conduct a device-specific analysis to assess the impact of each device on performance. We focus our analysis on FatigueSet as WESAD only has data from two devices (wristband and chestband). The bio-cue generation results are presented in Tables 8 and 9, along with matching results provided in Tables 10 and 11. The results show that BioQ outperforms all baselines across various device combinations. In Tables 8 and 10 we observe that Lester et al. [32] and Cornelius et al. [14] perform quite well when the devices are located close to each other (left and right earbuds). However, they fail to perform effectively for devices that are not located in a similar position on the body. On the other hand, BioQ works consistently better than baselines, showing its generalizability across various body positions. In terms of the impact of device placement on BioQ's performance, BioQ performs better with devices located close to each other or on the same side of the body compared to other device combinations. For instance, an FDR of 6.922 is achieved with left and right earbuds, and similarly strong performance is observed with earbuds or a headband (FDR over 4). In contrast, incorporating

Table 8: Device-specific analysis of bio-cue generation performance (FDR) from ACC across users.

left earbud	right earbud	wristband	headband	Raw	Feature	Lester et al. [32]	Cornelius et al. [14]	BioQ
✓	✓	×	×	0.000±0.00	0.381±0.01	0.915±0.04	0.778±0.05	**6.922**±0.95
✓	×	✓	×	0.002±0.00	0.156±0.02	0.008±0.00	0.002±0.00	**2.701**±0.13
✓	×	×	✓	0.002±0.00	0.340±0.02	0.024±0.00	0.017±0.01	**4.329**±0.32
×	✓	✓	×	0.001±0.00	0.154±0.02	0.007±0.00	0.003±0.00	**2.527**±0.28
×	✓	×	✓	0.003±0.00	0.331±0.03	0.019±0.00	0.014±0.00	**4.057**±0.28
×	×	✓	✓	0.000±0.00	0.179±0.02	0.122±0.02	0.090±0.02	**2.438**±0.19
✓	✓	✓	×	0.001±0.00	0.234±0.01	0.115±0.01	0.080±0.01	**3.332**±0.18
✓	✓	×	✓	0.005±0.00	0.346±0.03	0.148±0.00	0.113±0.01	**4.838**±0.29
✓	×	✓	✓	0.001±0.00	0.215±0.03	0.042±0.00	0.026±0.00	**2.636**±0.16
×	✓	✓	✓	0.006±0.01	0.210±0.03	0.038±0.00	0.024±0.00	**2.672**±0.22
✓	✓	✓	✓	0.001±0.00	0.243±0.03	0.077±0.01	0.052±0.01	**3.185**±0.22

Table 9: Device-specific analysis of bio-cue generation performance (FDR) from GYR.

left earbud	right earbud	headband	Across users			Same user across time		
			Raw	Feature	BioQ	Raw	Feature	BioQ
✓	✓	×	0.001±0.00	0.171±0.03	**5.743**±0.61	0.001±0.00	0.168±0.03	**2.293**±0.20
✓	×	✓	0.001±0.00	0.174±0.01	**2.055**±0.22	0.001±0.00	0.175±0.02	**0.728**±0.06
×	✓	✓	0.002±0.00	0.135±0.02	**2.030**±0.30	0.002±0.00	0.136±0.02	**0.770**±0.08
✓	✓	✓	0.023±0.02	0.152±0.02	**2.699**±0.21	0.025±0.02	0.146±0.02	**1.069**±0.07

Table 10: Matching performance (F1) from ACC-based bio-cues with various device combinations.

left earbud	right earbud	wristband	headband	Raw	Feature	Lester et al. [32]	Cornelius et al. [14]	BioQ
✓	✓	×	×	0.567±0.03	0.627±0.02	0.765±0.01	0.737±0.01	**0.863**±0.01
✓	×	✓	×	0.527±0.02	0.500±0.03	0.531±0.00	0.512±0.01	**0.751**±0.01
✓	×	×	✓	0.514±0.01	0.545±0.01	0.562±0.01	0.546±0.01	**0.784**±0.01
×	✓	✓	×	0.503±0.01	0.491±0.01	0.529±0.01	0.512±0.01	**0.751**±0.00
×	✓	×	✓	0.478±0.03	0.546±0.01	0.559±0.01	0.548±0.01	**0.788**±0.01
×	×	✓	✓	0.459±0.05	0.528±0.03	0.587±0.00	0.564±0.01	**0.754**±0.01
✓	✓	✓	×	0.571±0.01	0.543±0.04	0.610±0.00	0.589±0.00	**0.767**±0.01
✓	✓	×	✓	0.518±0.03	0.575±0.02	0.629±0.00	0.615±0.00	**0.803**±0.00
✓	×	✓	✓	0.492±0.04	0.537±0.01	0.557±0.00	0.539±0.00	**0.756**±0.00
×	✓	✓	✓	0.508±0.04	0.534±0.01	0.556±0.00	0.536±0.00	**0.754**±0.01
✓	✓	✓	✓	0.510±0.04	0.554±0.01	0.590±0.00	0.574±0.00	**0.759**±0.00

Table 11: Matching performance from GYR-based bio-cues with various device combinations.

left earbud	right earbud	headband	F1 Macro			EER		
			Raw	Feature	BioQ	Raw	Feature	BioQ
✓	✓	×	0.256±0.01	0.565±0.05	**0.882**±0.01	0.501±0.00	0.360±0.04	**0.100**±0.01
✓	×	✓	0.293±0.10	0.519±0.05	**0.762**±0.01	0.498±0.01	0.370±0.03	**0.226**±0.00
×	✓	✓	0.253±0.01	0.504±0.05	**0.764**±0.01	0.500±0.00	0.401±0.04	**0.226**±0.01
✓	✓	✓	0.363±0.10	0.523±0.04	**0.798**±0.01	0.509±0.03	0.384±0.02	**0.183**±0.01

a wristband yields lower results, likely due to the independent motion of the wrist or challenges in generating high-quality bio-cues from isolated movements. Nonetheless, BioQ's FDR for all device combinations remains above 1, indicating clear distinguishability, whereas baselines consistently produce FDR below 1.

Device-specific analysis for GYR (shown in Tables 9 and 11) also show similar trends. Since GYR data is unavailable for wristbands, they were excluded from the GYR experiments. Across both bio-cue generation and matching, BioQ performs notably better when earbuds are involved. For example, bio-cue generation achieves a high FDR of 5.743 across users and 2.293 for the same user over time, along with a matching F1 of 0.882 and an EER as low as 0.100. Including more devices in the system further stabilizes results, as training benefits from exposure to diverse device location combinations (e.g., both earbuds and headbands). This suggests that device-agnostic training is not only efficient for training and testing but also enhances accuracy stability in such scenarios.

4.5 Performance under Similar User Conditions

To assess BioQ's effectiveness in real-world scenarios and its robustness to subtle variations in user conditions, we conducted experiments under diverse user states. For instance, FatigueSet includes labels for baseline (seated at rest while baseline data is collected), physical activity (walking, jogging), and mental fatigue (dual letter and number task). Meanwhile, WESAD contains conditions such as baseline (neutral state), amusement (watching humorous video clips), and stress (public speaking and mental arithmetic tasks). In WESAD, roughly half of the participants performed these tasks while standing, and the remaining participants were seated, resulting in relatively minor variability in physical movement.

We consider Raw and Feature baselines for comparison methods since Lester et al. [32] and Cornelius et al. [14] are limited to only accelerometer data. Table 12 shows the results for cue generation. BioQ consistently outperforms these baselines across all user states and activities—achieving improvements of 13.5× for WESAD and 4.5× for FatigueSet in the across-user evaluation, and 5.6× (WESAD) and 5.4× (FatigueSet) in the same-user (across-time) setting compared to the best-performing baseline. We also provide bio-cue

Table 12: Bio-cue generation under similar user conditions.

Dataset	User State	Raw	Feature	BioQ
FDR (across users) ↑				
WESAD	Baseline	0.025±0.02	0.059±0.01	**0.465**±0.14
	Stress	0.002±0.00	0.047±0.03	**0.337**±0.05
	Amusement	0.010±0.01	0.019±0.01	**0.892**±0.22
FatigueSet	Baseline	0.039±0.02	0.836±0.15	**3.526**±0.76
	Activity	0.013±0.02	0.533±0.27	**1.511**±0.71
	Fatigue	0.053±0.04	0.545±0.14	**3.649**±0.89
FDR (same user, across time) ↑				
WESAD	Baseline	0.007±0.00	0.045±0.01	**0.268**±0.08
	Stress	0.004±0.00	0.048±0.03	**0.255**±0.08
	Amusement	0.000±0.00	0.004±0.00	**0.016**±0.01
FatigueSet	Baseline	0.042±0.02	0.352±0.10	**2.073**±0.95
	Activity	0.016±0.02	0.266±0.07	**1.247**±0.29
	Fatigue	0.104±0.05	0.312±0.11	**1.717**±0.28

Table 13: Bio-cue matching under similar user conditions.

Dataset	User State	Raw	Feature	BioQ
F1 ↑				
WESAD	Baseline	0.490±0.02	0.509±0.03	**0.636**±0.01
	Stress	0.413±0.02	0.549±0.03	**0.618**±0.02
	Amusement	0.490±0.05	0.568±0.02	**0.632**±0.02
FatigueSet	Baseline	0.324±0.07	0.664±0.08	**0.840**±0.04
	Activity	0.517±0.13	**0.782**±0.03	0.780±0.04
	Fatigue	0.271±0.02	0.767±0.04	**0.828**±0.04
EER ↓				
WESAD	Baseline	0.496±0.01	0.445±0.02	**0.367**±0.01
	Stress	0.503±0.00	0.450±0.02	**0.346**±0.03
	Amusement	0.477±0.02	0.419±0.03	**0.356**±0.01
FatigueSet	Baseline	0.501±0.04	0.304±0.08	**0.160**±0.04
	Activity	0.424±0.08	0.196±0.03	**0.180**±0.03
	Fatigue	0.490±0.01	0.200±0.01	**0.179**±0.04

Table 14: Runtime system cost.

Operation	Feat. Ext.	Emb. Gen.	Data Trans.	Matching
RPI4B				
Latency (ms)	25.3	3.2	6.2	2.1
CPU (%)	6.5	70	0	0
Energy (mJ)	18.72	8.26	2.21	5.46
RPIZW				
Latency (ms)	68.5	6.9	61.0	4.5
CPU (%)	13.6	18	0	0
Energy (mJ)	37.64	11.81	1.81	7.64
Pico W				
Latency (ms)	108.0	260.0	61.3	341.0
Energy (mJ)	22.32	54.63	19.55	71.88

matching results in Table 13 where BioQ demonstrates similarly robust gains. BioQ improves the F1 score by 1.2× on WESAD and 1.1× on FatigueSet while reducing the EER by 0.08 and 0.06, respectively, compared to the best-performing baseline. Notably, the Feature baseline is relatively effective in FatigueSet when users are engaged in more dynamic activities like walking or jogging. However, its performance drops in the less dynamic baseline and fatigue states, suggesting it struggles to capture subtler physiological changes. In contrast, BioQ remains robust across all conditions, highlighting its ability to adapt to varying motion levels and psychological states.

4.6 System Cost

Runtime cost: As BioQ is designed for wearable devices, we focus on three embedded platforms: Raspberry Pi 4B (RPI4B), Raspberry Pi Zero W (RPIZW), and Raspberry Pi Pico W (Pico W). These devices were chosen because commercial wearables do not yet support the direct execution of custom code. They also represent a range of computational capabilities, with the RPI4B modeling high-performance devices like smart earbuds and the Pico W representing ultra-low-power, compact wearables like smart rings or patches. We measured latency, CPU usage, and energy (using Monsoon Power Monitor) for each operation in BioQ.

The main runtime operations of BioQ are as follows: each wearable device generates (1) features and (2) bio-cues from available sensors. When matching is required, the devices send the embedding values to the host device, where (3) the host performs the matching. Table 14 shows the results when PPG signals are processed. We omit the CPU usage of Pico W because it executes instructions sequentially on a single core in a bare-metal environment, making CPU usage inapplicable.

The core operations of embedding generation and matching incur modest latencies and low energy demands on both the RPI4B and RPIZW. For instance, embedding generation completes in 3.2 ms (8.26 mJ) on the RPI4B and 6.9 ms (11.81 mJ) on the RPIZW, while matching requires only 2.1 ms (5.46 mJ) and 4.5 ms (7.64 mJ), respectively. The Pico W, by contrast, exhibits substantially longer latencies for these steps (e.g., 260 ms for embedding generation), primarily because it lacks optimized libraries and runs bare-metal code. Yet, this also indicates an opportunity: custom, hardware-specific implementations on the Pico W could significantly lower both latency and energy usage. Meanwhile, on the RPI4B and RPIZW, feature extraction and bio-cue transmission emerge as the dominant latency components, reflecting sensor- and device-specific operations rather than overhead from BioQ itself. On the Pico W, we limit feature extraction to simple statistical measures of PPG due to restricted library support, thereby reducing complexity in that particular stage. Finally, although the RPI4B draws more instantaneous power than the RPIZW, its shorter execution times lead to lower overall energy consumption for CPU-heavy tasks. Altogether, these results underline that BioQ can operate efficiently on resource-constrained devices, with headroom for further optimizations on ultra-low-power MCUs like the Pico W.

Server-side cost: On the server side, model training is required using a global dataset. To investigate its cost, we measured training time and model size using a desktop-scale server equipped with NVIDIA GeForce RTX 3090 GPUs. For the bio-cue generation model, using all available sensor modalities with a pair of devices from the FatigueSet (ACC, GYR, PPG) and WESAD (ACC, EDA, TEM) datasets, the training process took 103 and 316.7 seconds for bio-cue generation and 55.8 and 159.7 seconds for bio-cue matching, correspondingly. The bio-cue generation model sizes were 13.31KB and 18.50KB for the WESAD and FatigueSet datasets, respectively, while the matching model size was the same at 18.50 KB. Note that we do not generalize the training time as it depends on the amount of the global dataset.

5 Related Work

Support for wearable collaboration. Extensive research efforts have explored a variety of foundational technologies critical for enabling collaboration among wearable devices. Sensor fusion [13, 19]

combines data from multiple sensors to better understand user context. To overcome the constraint of a single device's processing capability, distributed inference techniques [2, 18] enable more efficient data processing by distributing tasks across devices. Dynamic device discovery protocols [27, 42] enhance the ability of devices to identify and communicate with one another in ad hoc networks without human intervention. Likewise, advancements in ad hoc networking [12, 23] ensure that wearables can quickly form flexible networks, essential for dynamic environments. Despite these advancements in wearable technology, less focus has been placed on managing these devices, particularly in scenarios where they are shared or used in public environments.

Detecting if devices are worn on the same body. Several studies have explored methods to determine whether devices are carried by the same person. Early works used accelerometers to compare walking patterns using the coherence function, identifying devices worn on the same individual when positioned closely to each other (e.g., on the waist) [32]. However, accuracy diminished when devices were farther apart. This was improved by training a Support Vector Machine (SVM) using coherence features from accelerometers [14], enhancing detection across varied body placements. Another study [47] compared footstep signatures from accelerometers to determine if devices were personally collocated, offering robust performance but mainly during walking activities. A common limitation of these approaches is their reliance on periodic movements, such as walking, which reduces their effectiveness in stationary scenarios or during non-walking activities.

There is extensive research on key establishment between devices, focusing on scenarios when two devices can securely generate a shared key when worn by the same person. Shaking two devices together can mutually authenticate the devices securely based solely on the shaking motion [36]. While intuitive and effective, this method relies on close physical proximity between the devices. For example, a user would need to remove their wearables (such as smart earbuds) and shake them near each other to enable pairing, which can be inconvenient. Alternatively, ongoing research leverages users' walking patterns (gait) detected via accelerometers for secure device interactions [45, 53]. Gait-Key [53] generates a shared key from synchronized gait data, while BANDANA [45] authenticates devices by analyzing gait patterns. Although these methods show strong results, they depend on user movements, limiting their applicability in static or non-repetitive scenarios.

6 Discussions and limitations

Beyond co-location detection: BioQ's unobtrusive and automatic detection of device co-location is critical for the scenarios in §2.2, especially for small form-factor wearables where even manual authentication is not feasible. BioQ achieves this by leveraging vital signs that can be monitored by wearable devices and their characteristics of being consistent throughout the body. However, to realize the scenarios in real-life situations, further operations such as policy management and executions are needed. For example, in case a user wants to mute confidential notifications when they lend one earbud to a friend to listen to music together, a user (or a system) needs to add a notification policy, which then must be coordinated on the corresponding devices, e.g., changing the notification message to be spoken on earbuds, based on the device co-location status. This becomes more challenging when wearable devices lack sufficient resource capabilities, which would necessitate a more sophisticated resource orchestration scheme, e.g., offloading heavy operations to a smartphone. We leave this as future work.

Limitation in data and generalizability. We acknowledge that our current datasets were not originally collected to capture subtle, real-world phenomena such as minimal-movement scenarios or shared-device use. We thus may not fully encompass the breadth of behaviors encountered outside controlled experimental settings. Nevertheless, our analyses across various user states and activities demonstrate that the proposed method maintains robustness despite these limitations, suggesting good potential for generalization. In the future, larger-scale data collections could be conducted, specifically targeting challenging conditions (e.g., shared earbuds or minimal-movement activities) to more comprehensively evaluate and enhance the system's capacity to detect subtle yet meaningful behavioral cues. By incorporating these tailored datasets, such future investigations would be better positioned to extend the approach's applicability to diverse, real-world deployment contexts.

Support for cross-modality. Currently, BioQ relies on devices sharing the same sensor modalities (e.g., a PPG channel present on all devices). In practice, however, a wristband might only have an accelerometer, while an earbud might only sense heart rate. Bridging these heterogeneous signals is nontrivial, as each sensor type exhibits distinct noise profiles, sampling rates, and domain-specific features. Future research could investigate cross-sensor embeddings or domain adaptation pipelines, where each modality is mapped into a common latent space. This would allow BioQ to function even when devices have disjoint sensor sets—a highly relevant scenario for real-world wearable ecosystems.

Model architectures and scalability. Finally, BioQ uses margin-based contrastive learning and a lightweight model for embedding generation. With recent advances in self-supervised learning and multimodal learning, other training pipelines might be effective for bio-cue generation. Furthermore, bigger models tend to have better capacity in extracting detailed features, and given that different mobile devices have different computing capabilities, we might afford to deploy bigger models for more capable devices. However, challenges in handling model heterogeneity would need to be addressed.

7 Conclusion

We introduced BioQ, a system designed to unobtrusively detect when multiple devices are co-located on the same body at a given time. BioQ is a key enabler for context-aware multi-device collaboration, facilitating use cases such as reliable health tracking, adaptive notifications, resource optimization, and enhanced privacy. BioQ is both device- and activity-agnostic, functioning seamlessly with various wearables positioned on different body parts and extending beyond motion-sensing data. It flexibly integrates physiological and inertial sensing data to generate and match bio-cues, which are sensor embeddings derived through contrastive representation learning. Experiment results confirm that BioQ consistently outperforms baseline methods in bio-cue generation and matching while maintaining cost-effectiveness.

References

[1] Khaled Alanezi, Xinyang Zhou, Lijun Chen, and Shivakant Mishra. 2015. Panorama: A framework to support collaborative context monitoring on co-located mobile devices. In *Mobile Computing, Applications, and Services: 7th International Conference, MobiCASE 2015, Berlin, Germany, November 12–13, 2015, Revised Selected Papers 7*. Springer, 143–160.

[2] Ananta Narayanan Balaji and Li-Shiuan Peh. 2023. AI-On-Skin: Towards Enabling Fast and Scalable On-body AI Inference for Wearable On-Skin Interfaces. *Proceedings of the ACM on Human-Computer Interaction* 7, EICS (2023), 1–34.

[3] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*. 144–152.

[4] Jordan Li Cahoon and Luis Antonio Garcia. 2023. Continuous Stress Monitoring for Healthcare Workers: Evaluating Generalizability Across Real-World Datasets. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. 1–5.

[5] G Clifford Carter. 1987. Coherence and time delay estimation. *Proc. IEEE* 75, 2 (1987), 236–255.

[6] G Clifford Carter. 1993. Tutorial overview of coherence and time delay estimation. *Coherence and Time Delay Estimation* (1993), 1–27.

[7] Marie Chan, Daniel Estève, Jean-Yves Fourniols, Christophe Escriba, and Eric Campo. 2012. Smart wearable systems: Current status and future challenges. *Artificial intelligence in medicine* 56, 3 (2012), 137–156.

[8] Youngjae Chang, Akhil Mathur, Anton Isopoussu, Junehwa Song, and Fahim Kawsar. 2020. A systematic study of unsupervised domain adaptation for robust human-activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–30.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[10] Bhawana Chhaglani, Utku Günay Acer, Si Young Jang, Fahim Kawsar, and Chulhong Min. 2023. Cocoon: On-Body Microphone Collaboration for Spatial Awareness. In *Proceedings of the 24th International Workshop on Mobile Computing Systems and Applications* (Newport Beach, California) *(HotMobile '23)*. Association for Computing Machinery, New York, NY, USA, 89–95. doi:10.1145/3572864.3580340

[11] Davide Chicco. 2021. Siamese neural networks: An overview. *Artificial neural networks* (2021), 73–94.

[12] Imrich Chlamtac, Marco Conti, and Jennifer J-N Liu. 2003. Mobile ad hoc networking: imperatives and challenges. *Ad hoc networks* 1, 1 (2003), 13–64.

[13] Maria Cornacchia, Koray Ozcan, Yu Zheng, and Senem Velipasalar. 2016. A survey on activity detection and classification using wearable sensors. *IEEE Sensors Journal* 17, 2 (2016), 386–403.

[14] Cory T Cornelius and David F Kotz. 2012. Recognizing whether sensors are on the same body. *Pervasive and Mobile Computing* 8, 6 (2012), 822–836.

[15] F John Dian, Reza Vahidnia, and Alireza Rahmati. 2020. Wearables and the Internet of Things (IoT), applications, opportunities, and challenges: A Survey. *IEEE access* 8 (2020), 69200–69211.

[16] Andrea Ferlini, Alessandro Montanari, Chulhong Min, Hongwei Li, Ugo Sassi, and Fahim Kawsar. 2021. In-ear PPG for vital signs. *IEEE Pervasive Computing* 21, 1 (2021), 65–74.

[17] Ronald A Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics* 7, 2 (1936), 179–188.

[18] Taesik Gong, Si Young Jang, Utku Günay Acer, Fahim Kawsar, and Chulhong Min. 2024. Synergy: Towards On-Body AI via Tiny AI Accelerator Collaboration on Wearables. arXiv:2401.08637 [cs.DC] https://arxiv.org/abs/2401.08637

[19] Raffaele Gravina, Parastoo Alinia, Hassan Ghasemzadeh, and Giancarlo Fortino. 2017. Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges. *Information Fusion* 35 (2017), 68–80.

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[21] Headphonesty. 2021. Sharing Headphones: Is It Bad For You? (2021). https://www.headphonesty.com/2021/01/sharing-headphones/ Accessed: 2024-09-12.

[22] Yash Jain, Chi Ian Tang, Chulhong Min, Fahim Kawsar, and Akhil Mathur. 2022. Collossl: Collaborative self-supervised learning for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–28.

[23] Emil Jovanov. 2019. Wearables meet IoT: Synergistic personal area networks (SPANs). *Sensors* 19, 19 (2019), 4295.

[24] Manasa Kalanadhabhatta, Chulhong Min, Alessandro Montanari, and Fahim Kawsar. 2021. FatigueSet: A multi-modal dataset for modeling mental fatigue and fatigability. In *International Conference on Pervasive Computing Technologies for Healthcare*. Springer, 204–217.

[25] Seungwoo Kang, Youngki Lee, Chulhong Min, Younghyun Ju, Taiwoo Park, Jinwon Lee, Yunseok Rhee, and Junehwa Song. 2010. Orchestrator: An active resource orchestration framework for mobile context monitoring in sensor-rich mobile environments. In *2010 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 135–144. doi:10.1109/PERCOM.2010.5466982

[26] Dean M Karantonis, Michael R Narayanan, Merryn Mathie, Nigel H Lovell, and Branko G Celler. 2006. Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. *IEEE transactions on information technology in biomedicine* 10, 1 (2006), 156–167.

[27] Kasem Khalil, Khalid Elgazzar, Mohamed Seliem, and Magdy Bayoumi. 2020. Resource discovery techniques in the internet of things: a review. *Internet of Things* 12 (2020), 100293.

[28] Diederik P Kingma, Max Welling, et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning* 12, 4 (2019), 307–392.

[29] Sylvia D Kreibig. 2010. Autonomic nervous system activity in emotion: A review. *Biological psychology* 84, 3 (2010), 394–421.

[30] Youngki Lee, Seungwoo Kang, Chulhong Min, Younghyun Ju, Inseok Hwang, and Junehwa Song. 2015. CoMon+: A cooperative context monitoring system for multi-device personal sensing environments. *IEEE Transactions on Mobile Computing* 15, 8 (2015), 1908–1924.

[31] Youngki Lee, Chulhong Min, Younghyun Ju, Seungwoo Kang, Yunseok Rhee, and Junehwa Song. 2014. An Active Resource Orchestration Framework for PAN-Scale, Sensor-Rich Environments. *IEEE Transactions on Mobile Computing* 13, 3 (2014), 596–610. doi:10.1109/TMC.2013.68

[32] Jonathan Lester, Blake Hannaford, and Gaetano Borriello. 2004. "Are you with me?"–using accelerometers to determine if two devices are carried by the same person. In *International Conference on Pervasive Computing*. Springer, 33–50.

[33] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644* (2015).

[34] Dominique Makowski. 2021. Neurophysiological Data Analysis with NeuroKit2. NeuroKit. https://neuropsychology.github.io/NeuroKit/

[35] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods* 53, 4 (feb 2021), 1689–1696. doi:10.3758/s13428-020-01516-y

[36] Rene Mayrhofer and Hans Gellersen. 2009. Shake well before use: Intuitive and secure pairing of mobile devices. *IEEE Transactions on Mobile Computing* 8, 6 (2009), 792–806.

[37] Chulhong Min, Akhil Mathur, Alessandro Montanari, and Fahim Kawsar. 2023. SensiX: A System for Best-Effort Inference of Machine Learning Models in Multi-Device Environments. *IEEE Transactions on Mobile Computing* 22, 9 (2023), 5525–5538. doi:10.1109/TMC.2022.3173914

[38] Chulhong Min, Alessandro Montanari, Akhil Mathur, and Fahim Kawsar. 2019. A closer look at quality-aware runtime assessment of sensing models in multi-device environments. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 271–284.

[39] Bayan Al Muhander, Jason Wiese, Omer Rana, and Charith Perera. 2020. Privacy-aware internet of things notices in shared spaces: A survey. *arXiv preprint arXiv:2006.13633* (2020).

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[42] T Poongodi, Anu Rathee, R Indrakumari, and P Suresh. 2020. IoT sensing capabilities: Sensor deployment and node discovery, wearable sensors, wireless body area network (WBAN), data acquisition. *Principles of internet of things (IoT) ecosystem: Insight paradigm* (2020), 127–151.

[43] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592* (2020).

[44] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 400–408. doi:10.1145/3242969.3242985

[45] Dominik Schürmann, Arne Brüsch, Stephan Sigg, and Lars Wolf. 2017. BANDANA—Body area network device-to-device authentication using natural gAit. In *2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 190–196.

[46] Society for Human Resource Management. n.d.. New uses for wearable devices in the workplace. (n.d.). https://www.shrm.org/topics-tools/news/technology/new-uses-wearable-devices-workplace Accessed: 2024-09-12.

[47] Animesh Srivastava, Jeremy Gummeson, Mary Baker, and Kyu-Han Kim. 2015. Step-by-step detection of personally collocated mobile devices. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*. 93–98.

[48] Zhiquan Tan, Yifan Zhang, Jingqin Yang, and Yang Yuan. 2023. Contrastive learning is spectral clustering on similarity graph. *arXiv preprint arXiv:2303.15103* (2023).

[49] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, and Cecilia Mascolo. 2020. Exploring contrastive learning in human activity recognition for healthcare.

arXiv preprint arXiv:2011.11542 (2020).

[50] Fang-Ching Tseng, Zih-Yun Chiou, Ho-Hsuan Chuang, Li-Ting Su, Yong-Han Lin, Yu-Rou Lin, Yi-Chi Lee, Peng-Jui Wang, Uei-Dar Chen, and Yung-Ju Chang. 2023. Multiple Device Users' Actual and Ideal Cross-Device Usage for Multi-Stage Notification-Interactions: An ESM Study Addressing the Usage Gap and Impacts of Device Context. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[51] Paul Van Gent, Haneen Farah, Nicole Van Nes, and Bart Van Arem. 2019. HeartPy: A novel heart rate algorithm for the analysis of noisy signals. *Transportation research part F: traffic psychology and behaviour* 66 (2019), 368–378.

[52] Diana A. Vasile, Fahim Kawsar, and Chulhong Min. 2024. Emerging Paradigms in Wearable Security: Adaptable and Secure Sandboxing for On-the-Fly Collaboration Among Wearables. *IEEE Security & Privacy* 22, 6 (2024), 30–39. doi:10.1109/MSEC.2024.3440198

[53] Weitao Xu, Chitra Javali, Girish Revadigar, Chengwen Luo, Neil Bergmann, and Wen Hu. 2017. Gait-key: A gait-based shared secret key generation protocol for wearable devices. *ACM Transactions on Sensor Networks (TOSN)* 13, 1 (2017), 1–27.

[54] Jerrold H Zar. 2014. Spearman rank correlation: overview. *Wiley StatsRef: Statistics Reference Online* (2014).

[55] Hao Zhang, Zheng Li, Jiahui Yang, Xin Wang, Caili Guo, and Chunyan Feng. 2023. Revisiting Hard Negative Mining in Contrastive Learning for Visual Understanding. *Electronics* 12, 23 (2023), 4884.

[56] Zhishuai Zhang, Wei Shen, Siyuan Qiao, Yan Wang, Bo Wang, and Alan Yuille. 2020. Robust face detection via learning small faces on hard images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1361–1370.