

# (Un)fair devices: Moving beyond AI accuracy in personal sensing

SOFIA YFANTIDOU\*, ATRIDE, Athens, Greece

MARIOS CONSTANTINIDES†, CYENS Centre of Excellence Ltd, Nicosia, Cyprus

DIMITRIS SPATHIS‡, Nokia Bell Labs Cambridge, Cambridge, United Kingdom of Great Britain and Northern Ireland and Google Research, London, United Kingdom of Great Britain and Northern Ireland

ATHENA VAKALI, Aristotle University of Thessaloniki, Thessalonike, Greece

DANIELE QUERCIA, Nokia Bell Labs Cambridge, Cambridge, United Kingdom of Great Britain and Northern Ireland and Politecnico di Torino, Turin, Italy

FAHIM KAWSAR, University of Glasgow, Glasgow, United Kingdom of Great Britain and Northern Ireland

Personal devices are omnipresent in our lives, seamlessly monitoring our activities, from smart rings tracking sleep patterns to smartwatches keeping an eye on missed heartbeats. The rich data streams from such devices fuel advanced Artificial Intelligence (AI) applications. Instead of solely relying on direct sensor measurements, these applications are increasingly leveraging Machine Learning (ML) model estimates to derive insights. But are these estimates biased or not? This literature review delivers compelling evidence about the impact of hidden biases that creep into ML models deployed on personal devices. We discuss critical bias issues drawn from prior work such as racial bias in pulse oximeters, weight bias in optical heart rate sensors, and sex bias in audio-based diagnostics. In response to these challenges, we advocate for a shift from prioritizing performance-oriented evaluations of personal devices to adopting assessments grounded in a human-centered approach. To facilitate this transition, we provide guidelines for the design, development, evaluation, and use of unbiased AI in personal devices, recognizing their potential impact on improving our health, lifestyle, and productivity—more than any other technology.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Applied computing** → *Consumer health*; • **Computing methodologies** → *Artificial intelligence*; • **Social and professional topics** → *Codes of ethics*.

Additional Key Words and Phrases: literature review, survey, machine learning, bias, fairness, responsible artificial intelligence, ubiquitous computing, sensing data

\*Work unrelated to current affiliation. The majority of this work was completed while the author was affiliated with Nokia Bell Labs and the Aristotle University of Thessaloniki, Greece.

†Work partially done at Nokia Bell Labs, United Kingdom.

‡Work unrelated to current affiliation at Google. The majority of this work was completed while the author was affiliated with Nokia Bell Labs.

Authors' Contact Information: Sofia Yfantidou, ATRIDE, Athens, Greece; e-mail: s.yfantidou@gmail.com; Marios Constantinides, CYENS Centre of Excellence Ltd, Nicosia, Cyprus; e-mail: marios.constantinides@cyens.org.cy; Dimitris Spathis, Nokia Bell Labs Cambridge, Cambridge, Cambridgeshire, United Kingdom of Great Britain and Northern Ireland and Google Research, London, United Kingdom of Great Britain and Northern Ireland; e-mail: dimitrios.spathis@nokia-bell-labs.com; Athena Vakali, Aristotle University of Thessaloniki, Thessalonike, Kentrik'ē Makedonia, Greece; e-mail: avakali@csd.auth.gr; Daniele Quercia, Nokia Bell Labs Cambridge, Cambridge, Cambridgeshire, United Kingdom of Great Britain and Northern Ireland and Politecnico di Torino, Turin, Piedmont, Italy; e-mail: daniele.quercia@nokia-bell-labs.com; Fahim Kawsar, University of Glasgow, Glasgow, Scotland, United Kingdom of Great Britain and Northern Ireland; e-mail: fahim.kawsar@nokia-bell-labs.com.



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

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

ACM 2832-0565/2026/2-ART

<https://doi.org/10.1145/3796223>

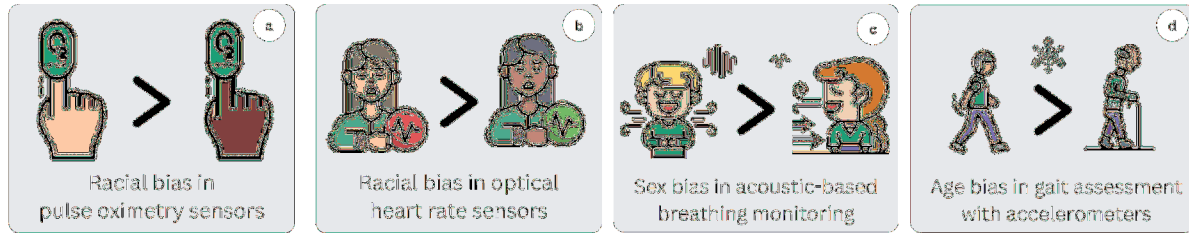


Fig. 1. **Motivating examples of biased personal devices drawn from prior work.** Research efforts have surfaced *racial biases* in (a) pulse oximetry [29] and (b) optical heart rate sensors [8], (c) *sex bias* in acoustic signals [36], and (d) *age bias* in accelerometer and gyroscope measurements [125].

## 1 Introduction

As an extension of ourselves, omnipresent and intimately close, personal devices continuously sense and learn from physiological and behavioral signals of our physical and mental well-being. Their ubiquity is unmistakable, with nearly one-third of adults in the United States using wearable devices to keep tabs on their health and fitness, while globally, the number of connected wearable devices has more than doubled in recent years [31]. Beyond their prevalence, personal devices have shown versatility across various domains, from workplace monitoring to educational contexts. Nevertheless, their most profound impact has been within the healthcare sector, effecting a paradigm shift from reactive to predictive health, with the promise of proactive interventions on the horizon. From detecting arrhythmias through tracking the subtle electrical impulses controlling our hearts to monitoring emotional arousal through the measurement of electrodermal responses, personal devices have evolved into miniature medical instruments, providing a comprehensive understanding of our health right from our pockets.

But here is the stark reality—we cannot assume that personal devices, with their powerful Machine Learning (ML) algorithms, operate equitably for all users. In this literature review, we provide compelling evidence that the data and models fueling these advancements are not immune to biases, possibly rendering algorithmic decisions “unfair”. There are two key notions of fairness: individual fairness and group fairness. Individual fairness ensures similar treatment for similar individuals, irrespective of their demographic characteristics. Conversely, group fairness, the focus of this work, emphasizes equitable outcomes across predefined demographic groups. Real-world cases of “unfair” ML algorithms in personal devices abound (Figure 1). For example, neural network algorithms trained to perform skin lesion classification showed approximately half the original diagnostic accuracy on black patients [47]. At the same time, people of color are consistently misclassified by health sensors such as oximeters as they were scientifically tested on predominantly white populations [29]. Yet, the implications of such biases extend beyond mere measurement errors. Biases in pulse oximetry sensors during the COVID-19 pandemic led to significant delays in recognizing hypoxia among Black and Hispanic patients, possibly contributing to ethnic disparities in COVID-19 outcomes [43]. As technology research prioritizes performance metrics and rapid progress, people become marginalized in the development process.

To systematically corroborate the extent to which research in mobile, wearable, and ubiquitous computing addresses biases in personal devices and to uncover sources of harm within their ML pipelines, we surveyed prominent venues publishing work on systems, human-computer interaction, and ML on personal devices. Surprisingly, according to our findings, the percentage of papers reporting fairness assessments ranged from none to a mere 10% published between 2018 and 2025. However, the lack of awareness transcends academia. A recent white paper on validating the Apple watch estimations of cardio fitness with maximal oxygen uptake omitted any reference to race as a baseline characteristic and refrained from providing performance breakdowns conditioned on demographics, adhering solely to aggregate evaluations [4].

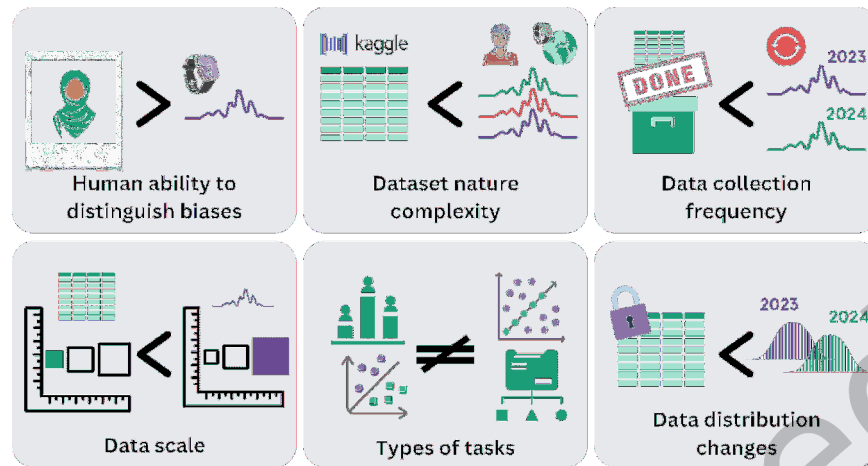


Fig. 2. **Conceptual differences between ML fairness and personal devices data.** Biases are harder to surface, given the complexity of temporal data. Continuous data collection can introduce data drifts, resulting in larger and more dynamic datasets. Additionally, the ML tasks addressed by the personal devices community diverge from those typically explored in fairness research.

Yet, it is time we acknowledged that personal devices are not immune to biases. These devices have evolved beyond mere step counting, now incorporating functionalities ranging from activity recognition and sleep inference to high-stakes applications—detecting Atrial Fibrillation (AFib) [127], diagnosing COVID-19 infections [9], and predicting fertility [70]. Contemplate the ethical and societal repercussions though if an AFib detection algorithm exhibited racial bias or a fertility prediction algorithm proved less reliable for women in non-Western countries. While such applications remain largely unregulated, they fall within the scope of forthcoming regulatory initiatives, such as the EU AI Act, which would mandate comprehensive risk assessments.

While fairness in ubiquitous computing has been acknowledged as a challenge since the 1990s [111], research on fairness in personal devices has received comparatively less attention than in the field of ML. Important works in ML offer foundational insights encompassing broad principles applicable across computer science subdomains [131], as well as specific solutions tied to different stages of the ML pipeline, such as ‘datasheets for datasets’ or ‘model cards for model reporting’ [34, 76]. Additionally, researchers focus on raising domain-specific awareness by identifying challenges and offering recommendations, such as addressing bias on the web [6], in computer vision [10], and in language models [7].

Yet, personal devices’ data and models possess unique characteristics not commonly shared with the broader discourse on AI Ethics (Figure 2). For example, the mobile and wearable community typically deals with small-scale datasets often collected by the authors in-the-lab or in-the-wild during feasibility studies, while the broader ML community frequently utilizes popular, medium- to large-scale benchmark datasets such as UCI Adult [51], German Credit [27], and COMPAS [52]. Such data are collected once and are immutable, in contrast to personal devices’ data that are mutable and longitudinally collected. Contrary to the tabular format of benchmark datasets, such data are mostly sequential/temporal, e.g., sensor data. Then, the biases are harder to surface. In other words, while it is relatively straightforward to distinguish a person’s skin tone from an image, it is much harder to do so from oximetry measurements. Similarly, an image captures the entire semantic meaning of a given data point, whereas a sensor signal is just an observation of the effect we measure; a variation in heart rate cannot be explained by the signal alone as it might be caused by external factors (alcohol, stress). As the vision of ubiquitous computing strives to blend technologies in the background, biases are blended, too. However, it is possible to

build devices that are both accurate and fair. As research evolves, the mobile and wearable computing community needs to recognize that the initial step, awareness, often takes a backseat in our discussions.

## 2 Background and Related Work

First, we present the reader with an introduction to fairness: definitions, measurements, and enhancement mechanisms. Next, we situate our review in previous fairness literature covering three broad areas: *a)* comprehensive surveys; *b)* domain-targeted surveys; and *c)* and surveys focused on over-represented populations.

**Fairness Definitions and Measurements.** Fairness entails the “protection of individuals and groups from discrimination or mistreatment with a focus on prohibiting behaviors, biases and basing decisions on certain protected factors or social group categories” [93]. Quantitative fields (e.g., computer science, statistics) view fairness as a mathematical problem of “equal or equitable allocation, representation, or error rates, for a particular task or problem” [78]. Fairness enhancement mechanisms fall under three categories [11, 83]: *a)* *pre-processing*; *b)* *in-processing*; and *c)* *post-processing* mechanisms. Pre-processing mechanisms involve altering the training data before feeding it into an ML model, while in-processing mechanisms involve modifying the ML algorithms to account for fairness during training. Ultimately, post-processing mechanisms involve altering the output probabilities of an ML model to mitigate biases/imbalances (e.g., enforcing a quota). However, due to the late stage in the learning process in which they are applied, post-processing mechanisms commonly obtain inferior results [112]. They are also considered too invasive or discriminatory since they deliberately damage accuracy for some subjects to compensate others [83]; hence they are less frequently preferred in practice.

**Comprehensive Surveys.** Addressing algorithmic bias in ML has been a longstanding issue [11], despite its recent surge. A number of comprehensive surveys shed light on data and model biases across domains and compared potential mitigation solutions. For example, Caton and Haas [11] and Pessach and Shmueli [83] discussed fairness metrics and categorized mitigation approaches into a widely accepted framework of pre-processing, in-processing, and post-processing methods independently of the application domain. Wan et al. [102] focused exclusively on in-processing modeling methods such as adversarial debiasing, disentangled representations, and fairness-aware data augmentation, while Pessach and Shmueli [83] provided an overview of emerging research trends, including fair adversarial learning, word embeddings, and recommender systems. Along these lines, Le Quy et al. [54] surveyed available datasets for fairness research, including financial, criminological, healthcare, social, and educational datasets. Yet, despite these surveys’ considerable contributions, they tend to be of generic nature and rarely discuss data, models, and applications related to an individual community.

**Domain-targeted Surveys.** Another line of work took a deep dive into well-defined domains, for example, by focusing on fairness for ML for graphs [21], on exploring notions of fairness in clustering [19], and on studying fairness in recommender systems [60]. Another group of works targeted specific unprivileged groups or high-stakes domains. For example, Olteanu et al. [81] reviewed the literature surrounding social data biases, such as biases in user-generated content, expressed or implicit relations between people, and behavioral traces, while in [98], the authors focused specifically on gender bias in Natural Language Processing (NLP). On a different note, Abdul et al. [1] featured emerging trends for explainable, accountable, and intelligible systems within the human-computer interaction community, also discussing notions of fairness. Closer to our work, Mhasawade et al. [75] discussed ML fairness in the domain of public and population health, and Xu et al. [116] explored algorithmic fairness in computational medicine, which only covers a subset of the broad, interdisciplinary mobile, wearable, and ubiquitous computing research domain.

**Surveys Focused on Over-represented Populations (WEIRD Research).** WEIRD research refers to a common criticism in the social sciences that much of the research is conducted on a sample of participants that is **W**estern, **E**ducated, **I**ndustrialized, **R**ich, and **D**emocratic. In particular, a comprehensive study conducted by Henrich et al. [41] in 2010 revealed a significant bias in sample populations, and found that most research samples come

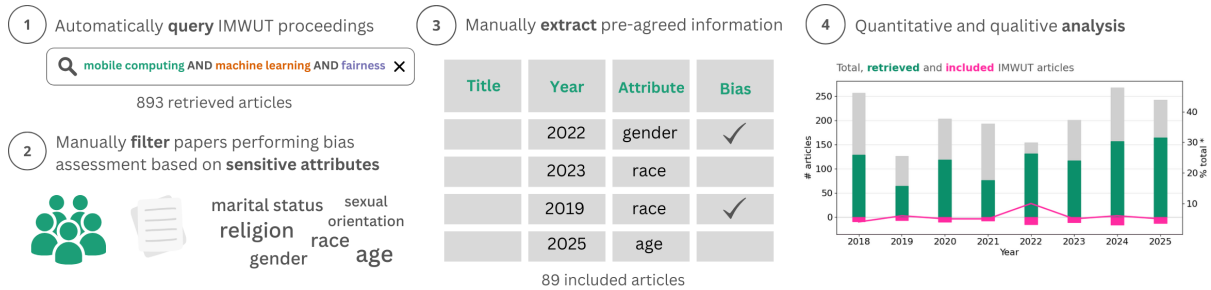


Fig. 3. **Illustration of the literature synthesis methodology.** A high-level overview of the process, including the querying, manual extraction, and filtering, and one of the main results. Note that for readability purposes, we present a simplified version of our query and data extraction. Our query retrieves ~ 58% of all IMWUT publications, while our eligibility assessment filtering ends up with 89 papers (~ 9% of retrieved papers). We notice that only a very small fraction of all IMWUT papers looks at fairness issues, with only a small deviation across years.

from WEIRD populations, which represent only 12% of the global population but account for 96% of research samples. Similarly, Linxen et al. [63] conducted a meta-study on CHI proceedings from 2016 to 2020, reporting that 73% of those papers are based on Western populations. Another recent meta-study on FAccT proceedings from 2018 to 2021 found highlighted concerns with biases in word embeddings and computer vision, and racial disparities, while only a ~ 10% of those papers used original, empirical datasets [53]. Similarly, by analyzing FAccT proceedings between 2018 and 2022, Septiandri et al. [86] found that 84% of the analyzed papers were exclusively based on participants from Western countries, particularly exclusively from the U.S. (63%).

It is evident that research communities other than mobile, wearable, and ubiquitous computing have already started to explore ways of reporting data and models in a fair manner to surface and, ultimately, address encountered biases. Yet, the state of fairness in this community remains unknown, as, at the time of writing, there exists no other survey or position paper in the intersection between personal devices and fairness.

### 3 Methodology

Next, we delineate our methodology for conducting this literature synthesis (§3.1) and provide our positionality statement (§3.2).

#### 3.1 Conducting the Literature Synthesis

We followed Kitchenham and Charters' protocol [50] for conducting systematic reviews to ensure the quality of included works and limit the initial retrieved papers. At least three authors were involved during the eligibility assessment to minimize the effects of bias and priming. In accordance with this protocol, we initially identified the need for a systematic review, as discussed in Section 1, namely to explore where personal device fairness overlaps with traditional fairness definitions, in what ways and to which extend. Figure 3 provides a high-level overview of the process.

**Paper Identification & Screening.** Mobile, wearable, and ubiquitous computing research crosses several fields, ranging from Human-Computer Interaction, Hardware and Software Systems, and Knowledge Discovery and Data Mining. For the scope of this review, we focused on the Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT), a premier journal in the community. For the search process, we utilized the ACM Digital Library, focusing on papers that were published between 2018 and 2025 to capture emerging trends in fairness and personal device research. Apart from year filtering, for the most part, we did not

```

(Abstract: ("mobile*" OR "wearable*" OR "smartwatch*" OR "smartphone*" OR "track*") OR
Title: ("mobile*" OR "wearable*" OR "smartwatch*" OR "smartphone*" OR "track*") OR
Keywords: ("mobile*" OR "wearable*" OR "smartwatch*" OR "smartphone*" OR "track*"))
AND
("deep learning" OR "machine learning" OR "artificial intelligence" OR "classification" OR "regression")
AND
("fair*" OR "bias*" OR "parity" OR "ethic*" OR "responsible AI" OR "RAI" OR "discriminat*" OR "non-discriminat*"
OR "equal*" OR "inclusiv*" OR "transparen*")

```

Fig. 4. **The query utilized for recovering relevant papers from the ACM Digital Library.** Terms related to mobile, wearable, and ubiquitous computing are highlighted in green, ML in orange, and fairness in purple.

limit our search to meta-data, such as titles, keywords, and abstracts, but rather we expanded it to any searchable field, including full text. That excludes the first part of the query, which tries to match terms such as wearable(s) or mobile(s) only in the papers' meta-data, as seen in Figure 4.

**Query Definition.** For the definition of our query, we followed similar terminology with relevant review papers in the fairness literature [11]<sup>1</sup>. Additionally, according to Fjeld et al.'s analysis of prominent AI principles documents, [30], "the fairness and non-discrimination theme is the most highly represented theme in our dataset, with every document referencing at least one of its six principles: "non-discrimination and the prevention of bias", "representative and high-quality data", "fairness", "equality", "inclusiveness in impact", and "inclusiveness in design", mostly included in our query's coverage. To capture the industrial perspective, we consulted the Responsible Artificial Intelligence (RAI) white papers issued by large tech companies. Specifically, Google's<sup>2</sup> and Meta's<sup>3</sup> RAI principles talk about "fairness and inclusion", Amazon's<sup>4</sup> RAI principles promote "diversity, equity, and inclusion" through "detecting bias". Similarly, Nokia's<sup>5</sup> RAI fairness pillar talks about "fairness, non-discrimination, accessibility, and inclusivity". Thus, an iterative refinement process resulted in the query shown in Figure 4. Notably, we omit the term "personal device" from the final query, as an ablation check<sup>6</sup> showed that adding the terms "personal device\*" or "personal sensing" within the inspected time range did not retrieve any additional publications, indicating that the ubiquitous and wearable computing terminology already provides complete coverage of the relevant literature.

**Eligibility Assessment.** To further validate our query, we manually inspected all publications from the latest available IMWUT proceedings at the time of manuscript preparation (Volume 6, Issue 4, published in January 2023) ( $N = 56$ ) to identify eligible papers for inclusion (see inclusion and exclusion criteria below). In total, we identified seven relevant publications, all of which were also returned by our query. This process was irrelevant to our final paper retrieval (pictured in Figure 5) and served validation purposes only. To ensure the high quality and relevance of the included papers, we defined appropriate exclusion criteria that helped us determine the included papers:

<sup>1</sup>The authors use "fairness datasets" as the primary query term along with other terms like "bias," and "discrimination," to narrow down the search.

<sup>2</sup><https://ai.google/responsibilities/responsible-ai-practices/?category=fairness>

<sup>3</sup><https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/>

<sup>4</sup><https://aws.amazon.com/machine-learning/responsible-machine-learning/>

<sup>5</sup><https://www.bell-labs.com/institute/blog/introducing-nokias-6-pillars-of-responsible-ai/>

<sup>6</sup>By ablation, we refer to a controlled query-variation analysis in which additional terms are added to the baseline search query to assess their contribution. Specifically, we augmented the device-related component of the query (Figure 4, green section) with the terms "personal device\*" and "personal sensing" and re-ran the search for the original time range (2018–2022). In both cases, the query returned 523 publications, identical to the baseline query.

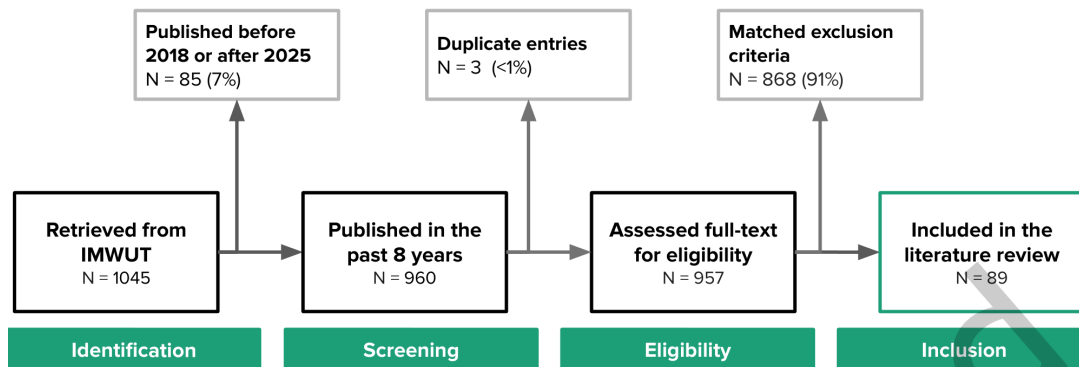


Fig. 5. **PRISMA flow diagram for paper inclusion.** Out of the 957 papers retrieved by our query after the screening, only 9% ( $N = 89$ ) did not check any exclusion criterion and thus were included in the literature synthesis.

- (1) Papers that do not provide a quantitative assessment of at least one empirical or artifact contribution in mobile and wearable technologies (UBI);
- (2) Papers that do not include a quantitative assessment of bias or performance discrepancy in their evaluation with regard to sensitive attributes, specifically age, gender, race, country of origin, language, disability or disease status, marital status, religion, employment status, physiology (i.e., weight, height), socioeconomic status and sexual orientation (FAIR);
- (3) Papers that discuss different domains, such as NLP or computer vision, without incorporating a ubiquitous component (DOM);
- (4) Papers that refer to bias in a different context, such as the bias-variance trade-off or the bias parameter in neural networks (CON).

**Inclusion & PRISMA Statement.** Figure 5 shows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [77] flow diagram. Specifically, the sequential execution of the steps above led to our review’s included papers. Overall, we screened 957 papers after date filtering and duplicate elimination. We then excluded 868 based on our exclusion criteria. Hence, we included 89 papers in our review.<sup>7</sup>

To assess consistency across venues, we reviewed publications from relevant mobile, wearable, and ubiquitous computing conferences and journals from 2022. Specifically, we reviewed ACM SIGMOBILE-sponsored events, including the International Conference on Mobile Computing and Networking (MobiCom,  $N_{2022} = 56$ ), the International Conference on Mobile Systems, Applications, and Services (MobiSys,  $N_{2022} = 38$ ), and the ACM Conference on Embedded Networked Sensor Systems (SenSys,  $N_{2022} = 52$ ), and relevant IEEE journals, including IEEE Pervasive Computing (IEEE Pervasive,  $N_{2022} = 49$ ) and IEEE Transactions on Mobile Computing (IEEE Trans. Mob. Comp.,  $N_{2022} = 319$ ). Following the same methodology, out of 245 retrieved papers, 12 papers complied with our screening and eligibility criteria: none from MobiCom (0% inclusion rate), 3 from MobiSys (8%), 2 from SenSys (4%), 2 from IEEE Pervasive (4%), and 5 from IEEE Trans. Mob. Comp. (2%). Table 1 summarizes our analysis.

Overall, while we acknowledge the limitations of considering a single publication for this review, these findings provided an indication of the generalizability of our results across mobile, wearable, and ubiquitous computing publication venues.

<sup>7</sup>To foster reproducibility, upon acceptance, we intend to release the review data and codebooks

Table 1. **To generalize these results across venues of similar scope, we analyzed recent proceedings (at the time of writing) of ACM MobiCom, MobiSys, and SenSys, and IEEE Trans. of Mobile Computing and Pervasive Computing, and found no deviation from our primary result.** Out of the 668 total papers published in 2022 in these venues, 376 were retrieved by our query, and only 27 complied with our screening eligibility criteria.

Venue	#Published (2022)	#Retrieved (% total)	#Included (% total)
ACM IMWUT	154	131 (85%)	15 (10%)
MobiCom	56	25 (45%)	0 (0%)
MobiSys	38	18 (47%)	3 (8%)
SenSys	52	8 (15%)	2 (4%)
IEEE Pervasive	49	15 (31%)	2 (4%)
IEEE Trans. Mob. Comp.	319	179 (56%)	5 (2%)
<b>Total</b>	<b>668</b>	<b>376 (56%)</b>	<b>27 (4%)</b>

### 3.2 Positionality Statement

Understanding researcher positionality is essential to shed light on our perspective on data collection and analysis [32, 39]. We situate this review paper in a Western country in the 21<sup>st</sup> century, writing as authors who primarily work as academic and industry researchers in the technology sector. We identify as two female and four male authors, and our shared backgrounds include mobile, wearable, interactive and ubiquitous computing, human-computer interaction, and machine learning and artificial intelligence.

## 4 Findings: How bias creeps in

Our perception of the world is shaped by the way our senses and cognitive processes interpret the information around us. Similarly, personal devices rely on sensory input to make decisions. But here’s the catch: these inputs and the algorithms that rely on them are not free from biases (Figure 1).

In summary, of the 957 papers retrieved, only 9% ( $N_{included} = 89$ ) were included in the review, representing just 5% of all IMWUT publications from 2018 to 2025. This underscores the timeliness and necessity of this work. Health emerged as the most prevalent domain, comprising almost one-quarter of the included papers. This was followed by Human-Activity Recognition (20%), Behavioral Sensing & Emotion (15%), Sound, Voice & Hearing (13%) and Privacy & Security (12%). Less common categories included Mobility & Navigation (6%), Motion, Gaze, Gesture & Touch (5%), and Cognition & Attention (5%). The least represented domain was Miscellaneous, accounting for 1% of the included papers.

### 4.1 Are Personal Devices Susceptible to Biases?

Several studies have revealed biases in acoustic signals tied to our body’s unique characteristics. In respiratory monitoring, higher errors have been reported for users with weaker chest motions and smaller body size [56]. Gender-related disparities have likewise been observed, where males’ generally higher respiratory intensity produces stronger abnormal respiratory audio features, while the under-representation of female samples during training limits generalization performance for female users [35]. Similar effects have been reported in breathing monitoring more broadly [36], due to variations of physiological structures (i.e., women’s smaller reflective surfaces encode less information). In speech and audio processing, gender-related biases have been observed in speech enhancement, where higher vocal frequencies in female speech negatively affect enhancement performance [108]. Complementary findings in speech recovery further indicate a slight performance advantage for male speech, attributed to more consistent acoustic characteristics and stronger low-frequency components [105]. Race-related biases have also been identified in speech enhancement systems, where hairstyle-associated physical

characteristics (e.g., African dreadlocks obstructing the acoustic gap created by auxiliary spacers) resulted in lower signal quality metrics [28]. Beyond speech and respiration, biases have been reported in other acoustic sensing and interaction systems. In acoustic heartbeat monitoring, individuals with higher BMI experienced reduced detection performance due to thicker thoracic musculature leading to weak heartbeat signals [107]. Similarly, acoustic interaction systems have shown degraded performance for children under 15, largely due to height constraints below supported operating ranges, as well as for older adults, whose age-related hearing decline increased output energy error, highlighting accessibility limitations tied to age and stature [129]. Finally, prior work on acoustic key generation using bone-conducted vibrations has highlighted age-related physiological differences, showing that variations in bone density affect vibration transmission patterns [109].

Biases have also been reported in radio-frequency-based sensing modalities, including WiFi, Ultra-wideband (UWB), and millimeter-wave radar systems. In contactless sleep monitoring using UWB signals, sleep apnea severity systematically affects performance [59], as signals from individuals with more frequent apnea events provide more informative disturbances that aid sleep-stage prediction, while milder cases are more prone to misclassification. RF-based respiratory monitoring further reports reduced accuracy for larger body types, where increased body size may exceed predefined candidate regions, limiting the capture of respiratory motion [104]. In mmWave radar-based gait recognition, children's more variable and playful walking patterns lead to increased gait instability and reduced recognition performance [103], while wearable mmWave head-gesture sensing is affected by morphological factors such as long hair (a common proxy for gender), which introduces signal noise due to random motion [121]. Finally, WiFi-based body composition sensing shows higher errors for larger individuals due to increased signal attenuation and noise sensitivity, with sex-related differences in abdominal physiology further influencing signal attenuation patterns [82].

Biases in computer vision models have also been encountered. For instance, video biases related to physiology, such as height, and appearance, such as hair or head covering (potentially proxies for gender and religion) [37] or race biases in non-contact photoplethysmography (PPG) measurements [80]. Vision-based physiological sensing has further revealed performance disparities linked to body morphology and skin-related attributes: infrared vein-based dehydration detection exhibits reduced accuracy for heavier individuals due to diminished vein visibility [45], while wearable optical spectroscopy shows accuracy variations across skin tones, largely attributed to under-representation of certain tones in training data [38]. Thermal imaging-based respiration sensing similarly reports slightly higher errors for participants with overweight BMI, though remaining largely BMI-agnostic compared to smartwatch-based baselines [2]. Some of these biases are easier to distinguish as demographic information is integrated into the data. For example, gender may be inferred from video or speech signals. However, sensor signals—the most prevalent data modality used by the community—are more challenging but equally susceptible to data biases.

Biases in sensor signals can be attributed to measurement inaccuracies or concept drift within signal patterns. For instance, research has revealed racial biases in pulse oximeters [29] and PPG sensors for heart rate measurement [8], while gender biases have been linked to variations in electrocardiogram (ECG) quality [119]. Beyond signal acquisition, some studies have shown that the accuracy of certain personal device functions varies with demographic characteristics. Gait detection through accelerometer and gyroscope measurements may be less accurate for elderly users due to age-related changes in gait patterns [125]. Similarly, Li et al. [57] report performance variations in Inertial Measurement Unit (IMU)-based neurological disease screening across sex and age groups, reflecting sex-specific gait biomechanics and age-related gait changes, with reduced sensitivity for younger subjects due to limited training data and for older adults in stroke and traumatic brain injury screening. IMU- and sEMG-based knee osteoarthritis gait re-training also exhibits higher errors for patients with more severe disease, overweight BMI, and male users, attributed to altered gait patterns, increased body-induced signal noise, and greater muscle loading, as well as limited data availability for severe cases [120]. Physiological sensing accuracy is further modulated by demographic characteristics. Sleep staging from ballistocardiogram signals

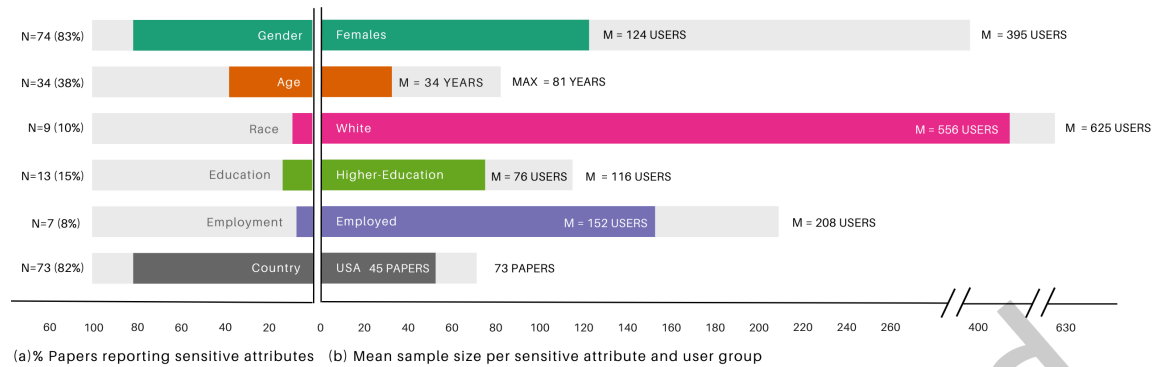


Fig. 6. **Analysis of sensitive attributes and data size.** The bar plots show the percentage of papers reporting certain sensitive attributes (left) and the (mean) sample size in the subset of papers reporting that attribute (right). Sample demographics reporting is not standardized and frequently incomplete, with race, employment status, and education being the least reported sensitive attributes.

is affected by age-related attenuation of sleep spindles, higher BMI-associated sleep apnea prevalence, and documented gender- and race-related differences in sleep architecture [58]. In cardiovascular monitoring, cuffless wrist-worn blood pressure sensing reports reduced signal quality for overweight individuals due to increased tissue depth over the radial artery, as well as variability linked to hypertension-related vascular stiffness [55]. In addition, speech recognition based on accelerometers can be substantially influenced by sex-related physiological variations in individuals' voices (e.g., women tend to have thinner and higher-pitched voices), particularly if these individuals are under-represented in the training data) [61, 97]. IMU and proximity sensors showed performance discrepancies in eating activity detection for users with a large BMI, attributed to factors like changes in the distance of the proximity sensor from the neck [127]. Comparable body-morphology-related effects have also been observed in pressure-based sensing, where human shape estimation from pressure image sequences achieves higher accuracy for subjects closer to average body weight, while performance degrades for lighter or heavier individuals due to distributional shifts in body shape [113]. As models become larger and their training objectives more universal, they learn to generalize to more tasks, including sensitive attributes' inference. Even sensors as seemingly inconspicuous as those used for heart rate and step tracking can be used to predict fine-grained outcomes relevant to health, fitness, and demographic characteristics [95]. However, the challenge with biases in personal devices is that they are often not straightforward to detect. They may remain blended in the background within the signal data or, worse, propagate into high-level inferences, affecting the assessments of health and well-being.

These hidden biases persist due to disparities in demographic representation within commonly employed datasets, alongside the under-reporting of demographic variables. For example, the recruitment of predominantly WEIRD samples is a pressing concern within the mobile and wearable community. Figure 6 shows an analysis of sensitive attributes across papers published in IMWUT. Gender is the most reported sensitive attribute with the mean sample size being 395 users, and the mean number of females being 124 (31%). Evidently, while the community has taken steps in the right direction, there is still plenty of room for improvement. Out of the papers for which the participants' country is reported or can be inferred, almost 6 out of 10 engaged with US samples. Indicatively, age is only reported in less than 4 out of 10 papers, with the mean sample age being 34 years old. For comparison, the maximum mean age reported is 81 years. In a world that is rapidly aging, personal devices are being predominantly developed and tested on young populations. Concerning education, less than 2 in 10 papers reported relevant information, with a median sample size of 116 users, 76 of which were college-educated

(66%). This is perhaps not surprising, as in the early stages of personal device research, participant recruitment frequently occurs within universities. Also, a mere 1 in 10 papers mentioned the participants' race. Among those papers, the sample size is an average of 625 users, with 556 of them being White (89%). These data highlight not only that race is often overlooked but also that non-White populations are underrepresented in the utilized datasets. This poses a real risk that models underperforming for non-White users may go unnoticed. For instance, prior work could not assess the impact of skin tone on AFib detection due to the majority (88.7%) of participants being White [126].

#### 4.2 How do we Mitigate Biases?

It is often infeasible to eliminate all sources of unfairness. Yet, the goal is to surface and mitigate biases as much as possible through fairness enhancement mechanisms.

Within the mobile, wearable, and ubiquitous computing research, preliminary pre-processing mechanisms for fairness enhancement included fair data representation. For instance, Liaqat et al. [62] included both healthy and non-healthy subjects in their dataset for respiratory rate monitoring. Similarly, Zhou et al. [130] employed a fairness-aware client selection mechanism for federated learning to ensure equal representation for subjects with worse connectivity.<sup>8</sup> Su et al. [97] performed data balancing managing to narrow the impact of gender voice differences on their speech recognition model. Along similar lines, Wang et al. [108] explicitly recruited a gender-balanced evaluation cohort to assess and reduce disparities in accelerometer-based speech recognition. Similarly, a strand of work explored data splitting, conditioned on the sensitive attribute (gender, age, BMI, skin tone, country, and health condition) to enable model personalization [48, 66, 74, 97, 125]. More advanced mechanisms suggest modifying feature representations. For example, Wang et al. [106] improved their activity detection model performance by normalizing the window-level features across gender and physiology. Similarly, in line with prior work [68], Su et al. [96] utilized disentangled representations, aiming to isolate activity patterns from redundant noises such as gender, age, and physiological differences.

Regarding in-processing mechanisms, Shahid et al. [88] altered their logistic regression model for Post-traumatic stress disorder (PTSD) screening to include sensitive attributes in its parameters. On a similar note, Sheng and Huber [91] devised a multi-task loss function consisting of activity, subject, and gender loss. Li et al. [59] similarly leveraged multi-task learning, where signals from subjects with more severe sleep apnea provided informative auxiliary supervision that helped correct sleep-stage predictions. Along the same line, Meegahapola et al. [73] have shown that partially personalized country- or continent-specific models outperform country-agnostic models in depression screening, highlighting the role of group-conditioned in-processing strategies in mitigating performance disparities across geographic populations. Finally, in quantifying the causal effect of individual mobility on health status, Zhang et al. [128] considered correlated sensitive attributes, such as age and socioeconomic status, as confounding variables in their causal model. However, they noted that due to dataset privacy constraints in reporting demographic variables, potential unobserved confounding variables might have been missed, highlighting the conflict between fairness and privacy [13]. At the system level, Djebrouni et al. [26] proposed a bias-constrained federated learning aggregation approach that enforces a predefined fairness objective by constraining bias below a threshold while maintaining accuracy.

Finally, post-processing mechanisms aim to mitigate bias by correcting measurements or adapting model outputs after model training. For instance, Hamid et al. [38] applied bias-aware calibration by leveraging skin-tone features and device identifiers to produce corrected oxygen saturation estimates, reducing skin-tone-dependent errors in pulse oximetry. Similarly, Wang et al. [109] employed per-person bone-channel estimation models to compensate for age-related differences in bone density. Li et al. [55] combined single-point calibration with

<sup>8</sup>While Internet connectivity is not a sensitive attribute per se, it has been linked with socioeconomic status, race, nationality, gender, and age, all of which are sensitive attributes [79].

feature-wise attention to assess the importance of each feature for different subjects and adapt blood pressure estimation to individual physiological patterns.

Despite the significant efforts of the aforementioned pioneering works in the community, 3 out of 4 included papers did not report any fairness enhancement mechanism, regardless of the presence of bias in their models. This is partly due to a lack of consideration for fairness-related harms, but it is also connected with the nature of several personal device works: artifact contributions, proof of concept, and early-stage technology development, where performance is prioritized.

### 4.3 How does the Community Capture Alternative Notions of Fairness?

Previously, we have established that only a small fraction of retrieved works (~9%) provide conventional fairness definitions with respect to one or more sensitive attributes. Yet, we believe such definitions do not do full justice to the community’s work, which strives for “fairer” models, perhaps not across sensitive attributes but differing experimental conditions. In particular, we noticed that, in evaluating new personal device systems, the community aims for generalizable and robust models by performing ablations studies, comparing deployment settings, and personalizing models for users and groups (67% of included papers).

**Ablation Studies.** In an ablation study, one or more components of the model are systematically removed or modified, and the performance of the model is evaluated after each change, ensuring its generalizability and robustness. In the included papers, ablation studies take the form of performance evaluation comparisons based on: *user-related*, *device-related*, *environmental*, *experimental*, and *domain-specific components*. In particular *user-related components* include user motion and orientation during data collection in sleep posture monitoring [124], respiratory monitoring [36, 104], gesture recognition [65], user identification [40, 89, 103], indoor localization [17], audio processing and speech generation [87, 108], and heart activity monitoring [115], as well as physiological characteristics and aesthetics, such as neck circumference, hair or clothing in fine-grained activity sensing [56], breathing and vital sign monitoring [36, 107], sleep monitoring [90] and user identification [37, 117]. *Device-related components* include device type, placement, and orientation, sampling rate, and operating system in activity and gaze tracking [44, 56], vital sign monitoring and physiological sensing [67, 107], speech recognition via built-in sensors and speech synthesis [61, 97], audio processing [108] and user behavior sensing [49, 103, 114]. *Environmental components* include ambient noise, light, and temperature that might affect data quality of acoustic [16, 33, 56, 87, 107] or video [37, 67, 101, 110] signals, or random passers-by for human identification [117]. Regarding experimental setup, few included papers studied the effect of equipment placement (e.g., distance, angle) and characteristics (e.g., range) on the model’s robustness in activity sensing [56], audio processing [108] and vital sign monitoring using acoustic signals [36, 107]. Experimental setups also include the effect of hyperparameters, such as data size, number of clients in federated learning, or deep learning architectural choices on performance [26, 94, 114]. Yet, the choice of components to consider in an ablation study is highly domain-dependent. *Domain-specific components* have no limitations and can range from screen size in scrolling interaction experiments [69] to food structure in food-related artifact development [18] and bed size and mattress firmness in sleep posture recognition [64].

**Deployment Setting.** A study’s deployment setting, whether it be in-the-lab or in-the-wild, can significantly impact its outcomes. While laboratory settings can provide controlled environments for experimentation, they may not accurately reflect the complexities of the real world in which the applications are deployed. As a result, in-the-wild (or in-situ) studies have emerged as an alternative, focusing on evaluating the situated design experience of mobile, wearable, and ubiquitous computing. In the included papers, perhaps not surprisingly, in-the-lab studies prevail (~ 54%), which can be explained by the nature of numerous IMWUT papers presenting artifacts or early-stage work. Nevertheless, ~ 29% of included papers conduct in-the-wild studies, and a small fraction of papers (~ 8.5%) report results for both deployments. Interestingly, while 3 out of 10 in-the-lab studies

Table 2. **Fourteen Recommendations for “Fairness by design” in personal devices.** Guidelines for researchers for performing fairness assessments given challenges specific to personal devices, accompanied by examples.

#	Generic Recommendations	Recommendations Tailored to Personal Devices	Example
<b>DATA COLLECTION &amp; ANNOTATION</b>			
1	Identify biases across demographics during design	Consider diverse physiological characteristics beyond legally protected attributes for fair representation	Acoustic sensing exhibits weight bias [107]
2	Recruit adequate sample sizes to ensure diverse representation of different groups	Consider contextual variables specific to the individual’s environment and lifestyle and address variable engagement levels across demographics in personal devices	AFib detectors consider abnormal heart rhythms under a single label due to small sample sizes [127]
3	Ensure that the ground truth data labeling process involves a diverse group of annotators	Train annotators on implicit biases to cultivate fair representation in human-centric personal device data labeling	Personal devices datasets are often self-annotated (by WEIRD authors)
4	Report protected attributes in datasets	Ensure transparency around sensitive attributes, including proxy features that might introduce bias to predictions, considering the “unreadable” nature of the data	Confounding variables may be missed due to privacy constraints in reporting protected attributes in HAR datasets [13]
<b>DATA EXPLORATION &amp; MANIPULATION</b>			
5	Validate data across groups to detect anomalies	Develop data validation methods to normalize across factors like device placement and user activity levels, mitigating biases introduced by in-the-wild variability	Sensor outliers can bias certain groups
6	Investigate sources of measurement error	Examine sources of measurement error from device hardware in model-based estimates to enhance accuracy in downstream tasks	Sensor drift over time can cause discrepancy between the physical state that is measured and the sensor output
7	Be considerate of how data pre-processing methods affect diverse demographics	Consider the impact of windowing on data representation, design choices, and inconsistent data due to device fit and no wear time	Sampling rate or model pruning and compression can bias on-device ML [99]
<b>MODEL DEVELOPMENT &amp; EVALUATION</b>			
8	Consider out-of-the-box fairness enhancement methods to mitigate bias in the data and models	Prioritize lightweight neural architectures and adaptive bias mitigation algorithms, addressing limited software library support for common device ML tasks	Research on bias mitigation for multi-class classification and regression tasks lags behind [3]
9	Consider indirect notions of fairness	Address fairness issues arising from device diversity and connectivity	Client exclusion due to unsupported devices in federated learning [20]
10	Assess model performance using varied fairness metrics for intersectional user groups	Develop fairness metrics suitable for real-time assessment on personal devices and consider the ambiguity in defining fair performance thresholds, especially for regression tasks	Fairness metrics offer only static snapshots, failing to capture real-time fairness dynamics in personal devices
11	Generate diverse synthetic data covering various protected attributes and intersections	Explore timeseries-specific data generation methods to ensure fairness in synthetic personal devices data	Personal devices datasets often lack size to enable fairness assessments [123]
<b>MODEL DEPLOYMENT AND USE</b>			
12	Monitor real-time performance of deployed models	Identify and adjust for data (sensor) and fairness drift	Fall detectors trained on younger populations and deployed on the elderly
13	Reassess the model using collected data after deployment to ensure continued fairness	Account for challenges like lack of labels post-deployment by utilizing unsupervised protocols	In fall detection ground truth data are unavailable without user reporting
14	Provide transparency in model versioning to increase accountability and reproducibility	Ensure transparency in changes to estimate models, as modifications may impact multiple downstream tasks, affecting user experience if unreported	Changes in algorithms cause drifts in unexpected downstream tasks that may disproportionately affect user groups

do not identify biases, this number falls to 0.5 for in-the-wild studies. This confirms our intuition that controlled environments can conceal biases that would emerge once a model is deployed in the real world.

**Personalization.** 1 in 5 papers trained separate personalized models for inference on a single subject [5, 55, 89, 90, 109, 125] or a group of subjects sharing a common characteristic [73, 123]. For instance, Liu et al. [66] built personalized models for different age groups “*illustrating differences in communication patterns across age demographics that can impact model performance*”. Similarly, Mendel et al. [74] utilized age-specific models for

just-in-time mobile safety help, as “*different ages in the sample have a significant influence on supportable moment predictions*”. Liu et al. [67] developed personalized models based on skin tone for camera-based photoplethysmography, as “*previous work had already highlighted [skin tone and gender] issues with the Plane-Orthogonal-to-Skin [method]*”. Su et al. [97] developed gender-specific models for speech recognition, as “*women’s voice is generally thinner and higher in pitch*”, while Zhang et al. [127] explored BMI-based models for detecting eating activities via a multi-sensor necklace, due to “*differences in movement patterns while eating, change in the distance of the proximity sensor from the neck, and difference in posture during the eating activity*”.

Overall, the gains from the interaction between the mobile, wearable, and ubiquitous computing and fairness communities can be bi-directional. On one side, personal device research can learn from other communities’ best practices in fairness assessment and reporting. Still, the latter can build on this community’s approach of rigorous validation, model personalization, and striving for robustness and generalizability through in-the-wild deployments and comprehensive ablation studies.

## 5 The way forward: guidelines for making personal devices fair

Borrowing from “Privacy by design” [12], we propose a “Fairness by design” equivalent, requiring AI developers and researchers to consider data and model fairness from the very beginning of any system design. This proactive approach prioritizes fairness as a core value in the development of personal devices. To guide the mobile and wearable computing community, we offer fourteen recommendations for seamlessly integrating fairness throughout the ML pipeline (Figure 2). These guidelines span across data collection and annotation, data exploration and manipulation, model development and evaluation, and finally, model deployment and use.

**Data Collection & Annotation.** For data collection, researchers should *identify the types of biases and harms* (#1) relevant to their task (e.g., quality-of-service, allocation, and erasure harms [23]). For example, in an AFib detection application, quality-of-service harms could occur if the model had a substantially different performance across ages. In contrast, allocation harms could occur if such difference led to one group unfairly receiving worse care. It is also important to consider group demographics (e.g., historically marginalized groups) relevant to a particular deployment setting that might be harmed. In personal devices, besides legally mandated protected attributes, users’ physiological characteristics like height, weight, or health conditions must also be considered. For example, acoustic sensing for vital sign monitoring applications can be less accurate for overweight and obese people [107]. While comprehensive coverage of sensitive attributes might not be guaranteed, we highly recommend consulting domain experts or conducting feasibility studies in the target application domain.

Ideally, researchers should *collect adequate and diverse sample sizes* (#2) tailored to the application’s needs and target audience to ensure the generalizability of the prediction task. For example, AFib detection systems have shown poor performance in people with abnormal heart rhythms due to limited data diversity in terms of subjects with different types of heart conditions [127]. Best practices in publishing ubiquitous datasets with protected attributes include large-scale datasets such as MyHeart Counts [42], GLOBEM [118], and MIMIC [46] datasets, as well as smaller datasets such as LifeSnaps [122], MotionSense [71], and ExtraSensory [100]. However, while acknowledging the impracticality of recruiting large samples in feasibility studies, researchers should prioritize diversity in terms of relevant protected attributes to mitigate biases in later development stages. Interestingly, factors such as geographic location, daily routines, and cultural practices can significantly impact the data generated by personal devices. Additionally, achieving representative samples in wearable technology may be challenging due to variable engagement levels across demographics. Yet, diverse sampling must balance representing the target audience with maintaining a learnable data distribution to avoid noise and outliers.

For data annotation, researchers should aim for *representation from diverse annotators* (#3) and offer implicit bias training. This holds particularly true in scenarios such as personal device applications where labels characterize human subjects as annotators’ beliefs and identities can lead to biased label assignment [85].

Additionally, researchers should *supplement published datasets with protected attributes* (#4) to facilitate fairness analyses. This requires anonymization and secure sharing through designated platforms such as PhysioNet. In signal data, proxy features may introduce bias to downstream tasks [95], complicating debugging due to the data’s obscured nature. For example, [13] emphasized the trade-off between fairness and privacy, noting how dataset privacy constraints may obscure potential unobserved confounding variables. In such scenarios, researchers should consider “fairness in unawareness” solutions, wherein disparity is assessed even when the protected class remains unobserved [15]. At the same time, the mobile and wearable research community itself could implement a mandatory data statement policy, requiring authors to report sensitive attributes concerning their participant samples in line with recent quests for data excellence [84].

**Data Exploration & Manipulation.** For data exploration, data validation methods should be applied across sensitive attributes to facilitate fairness. For example, *inspecting outliers or missing values* (#5) that systematically fall into particular demographic groups. Unlike controlled laboratory settings, data gathered in-the-wild often contain noise and uncertainty due to factors such as sensor inaccuracies, environmental fluctuations, and user variability. They can also be affected by factors like device placement and user activity levels.

Similarly, *measurement errors* (#6), namely input data containing inaccuracies, for example due to device malfunctions or variations, can affect fairness in downstream tasks. For example, data from wearables might exhibit anomalies due to device fit issues on smaller wrists typically found in women or children. Yet, in such devices, reliance on model-based estimates for various well-being metrics means that these errors can propagate to subsequent tasks. If the original model has not been validated across different groups, it can adversely affect any applications utilizing such data. In contrast, other domains, like computer vision, are less affected by hardware variations; for instance, the camera used typically has minimal impact on model outcomes. For this purpose, visualization tools (e.g., What-If Tool, FairLens, Tensorflow’s Fairness Indicators) can help surface potential data anomalies before they creep into models.

In data manipulation, researchers must *scrutinize pre-processing decisions* (#7). Unlike fields with out-of-the-box datasets, device data often necessitates additional windowing before predictive modeling. For instance, applying a sliding window method can yield numerous samples from a single user’s sensor signal, potentially leading to an overabundance of training data from a limited user pool. As a result, the model may not learn generalizable patterns. Similarly, design choices, such as sampling rate or model pruning and compression, can lead to bias propagation on on-device ML workflows [99]. Illustrative examples of good practice include data balancing techniques to reduce the impact of sex-related voice differences in speech recognition models [97], or data splitting, conditioned on sensitive attributes to enable model personalization [97, 125]. More advanced mechanisms suggest normalizing window-level features across demographics and physiology, or utilizing disentangled representations, aiming to isolate activity patterns from redundant noises such as demographic differences.

**Model Training & Evaluation.** In model training, while we have witnessed a remarkable consolidation of deep learning models and architectures in personal devices, such as Convolutional Neural Networks and Transformers, fairness-preserving mechanisms remain unexplored. However, there is a plethora of fairness toolkits, such as FairLearn, AIF360 and Aequitas, which provide diverse pre-processing, in-processing, and post-processing *bias mitigation algorithms out-of-the-box* (#8) – some applicable regardless of input data types. However, many libraries fail to provide sufficient support for time-series tasks, particularly multi-class classification and regression, which are common in personal device research. Similarly, adaptive bias mitigation algorithms capable of dynamically adjusting to changes in user contexts and device conditions are also lacking. As examples of good practice, in classification tasks, fairness-preserving mechanisms include the incorporation of sensitive attributes as model parameters in Post-traumatic stress disorder (PTSD) screening [88], or the adaptation of the loss function to incorporate activity, subject, and gender loss in human-activity recognition [91]. Other mechanisms also include adversarial debiasing (i.e., training a model to simultaneously classify inputs accurately while minimizing the

ability of an adversary to infer sensitive attributes), and discrimination-aware regularization (i.e., penalizing the model for making decisions based on sensitive attributes) Also, particularly in the context of healthcare, it is important to consider the fairness-utility trade-off, where optimally fair algorithms may result in poorer decisions across groups [22]. More broadly, as personal devices demand lightweight neural architectures, there have been efforts to find neural networks that strike a balance between fairness and accuracy, all while ensuring compliance with hardware specifications [92].

Similarly, we should also *consider indirect notions of fairness* (#9) during training. For example, within the paradigm of federated learning, the resource allocation of participating devices may also reflect the demographic and socio-economic information of owners, which makes the exclusion of such clients unfair in terms of participation. Cheaper devices cannot support the execution of large models and are either excluded or dropped together with their unique data [20]. For example, prior research [130] demonstrates good practices in fair resource allocation in wearable devices of inferior networking conditions (linked with socioeconomic status, race, nationality, gender, and age). Regardless of the training paradigm, one should consider additional constraints, such as device diversity or connectivity.

For model evaluation, assessing fairness in personal devices presents a dual challenge: the definition of a fair performance threshold is ambiguous, and fairness evaluation in regression or multi-class classification is understudied, particularly in online scenarios. Establishing a performance threshold is complex, with statistical tests like the Student's t-test facing limitations in cross-validation. McNemar's Test and  $5 \times 2$  cross-validation and its refinements are better alternatives proposed [25]. In classification scenarios, fairness metrics offer alternatives, but the absence of a universal fairness definition hinders their applicability. In regression, fairness definitions are rare and employ measures like statistical parity and bounded group loss [3]. Comparing outcome distributions through Kullback–Leibler divergence is also used. However, mainstream fairness libraries like AIF360 and FairLearn lack regression-specific metrics. At any rate, single evaluation metrics struggle to reflect the quality of ML models over time. As such, *monitoring and reporting lightweight fairness metrics* (#10) suitable for real-time assessment on ubiquitous devices, considering constraints such as limited computational resources but also drift and online learning should become standard practice. Fairness trees accompanied by domain expertise can help researchers choose appropriate metrics. A practical example can be drawn from [123], where biases are quantified across different stages of the ML lifecycle.

At the same time, we need to account for intersectional fairness, i.e., designing algorithms to account for different social identities that can intersect and impact a person's experiences and outcomes. However, we acknowledge that sometimes it is not feasible to collect data from representative demographics to enable such comparisons, especially for smaller pilot studies. In these cases, researchers can leverage advances in time-series- or sensor-specific generative models to *synthesize data covering multiple protected attributes* (#11) and potential intersections [14, 24]. However, it is essential to acknowledge that these models are currently experimental, and rigorous fairness assessments have not yet been performed.

**Model Deployment & Use.** For model deployment, we should *monitor models' performance in real-time* (#12) ensuring fair predictions despite shifts in input data and demographics. For instance, in fall detection systems, models are typically trained on simulated falls by young individuals, yet are used to detect real falls in elderly. Sensor drift can also introduce imperceptible changes in sensors' outputs over time due to factors like temperature, humidity, and hardware aging. These changes can lead to measurement inaccuracies disproportionately affecting certain regions or demographics.

The impact of data drift can be mitigated during deployment through *ongoing model reassessments and refinements* (#13) tailored to the target population. However, addressing this in personal devices presents challenges, particularly in the absence of post-deployment labels. For example, in fall detection systems, obtaining ground truth data on falls relies on user self-reporting. To address this issue, unsupervised online domain adaptation has

been proposed to prevent models from underperforming when applied to new users [72]. Additionally, to ensure that model behavior is generalizable for a larger group of users, researchers can also employ newly developed cross-dataset benchmarks for longitudinal models [118].

During model use, researchers should consider that model refinements can disproportionately affect certain groups. For example, changes in heart-rate prediction models, even if beneficial for minority groups, can propagate to resting heart rate estimations, affecting the user experience. Thus, researchers should *prioritize transparency in model versioning* (#14) and accompany each version with reports on performance and fairness. This is particularly relevant for personal devices, as they rely on several ML estimates to determine what is presented to users as their body’s ground truth.

## 6 Conclusions

In our technology-laden lives, the pervasive integration of personal devices, diligently monitoring our health and behavior, poses a critical question: Do these digital companions play fair? This article contends that despite their high-stakes applications, personal devices harbor biases, a concern that has been conspicuously overlooked. Drawing attention to the dearth of fairness assessments and the unique characteristics of personal device data, this work advocates for a proactive stance termed “Fairness by design”. As such, it moves beyond traditional accuracy evaluations to ensure personal devices not only deliver accurate insights but also adhere to principles of equity and justice for all users.

## Acknowledgments

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 813162. The content of this paper reflects only the authors’ view, and the Agency and the Commission are not responsible for any use that may be made of the information it contains. The authors acknowledge the use of generative AI tools for assistance with language editing and stylistic refinement. All scientific content, analysis, interpretations, and conclusions are the authors’ own, and the authors take full responsibility for the content of the manuscript.

## References

- [1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (*CHI '18*). Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3173574.3174156
- [2] Rishiraj Adhikary, Maite Sadeh, Nipun Batra, and Mayank Goel. 2024. Jouleseye: Energy expenditure estimation and respiration sensing from thermal imagery while exercising. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–29.
- [3] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*. PMLR, 120–129.
- [4] Apple. 2021. Using Apple Watch to Estimate Cardio Fitness with VO2 max. [https://www.apple.com/healthcare/docs/site/Using\\_Apple\\_Watch\\_to\\_Estimate\\_Cardio\\_Fitness\\_with\\_VO2\\_max.pdf](https://www.apple.com/healthcare/docs/site/Using_Apple_Watch_to_Estimate_Cardio_Fitness_with_VO2_max.pdf)
- [5] Olivier Augereau, Charles Lima Sanches, Koichi Kise, and Kai Kunze. 2018. Wordometer Systems for Everyday Life. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 123 (jan 2018), 21 pages. doi:10.1145/3161601
- [6] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.
- [7] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.
- [8] Brinnae Bent, Benjamin A Goldstein, Warren A Kibbe, and Jessilyn P Dunn. 2020. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ digital medicine* 3, 1 (2020), 18.
- [9] Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. 2020. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) (*KDD '20*). Association for Computing Machinery, New York, NY, USA, 3474–3484. doi:10.1145/3394486.3412865

- [10] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.
- [11] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053* (2020).
- [12] Ann Cavoukian et al. 2009. Privacy by design: The 7 foundational principles. *Information and privacy commissioner of Ontario, Canada* 5 (2009), 12.
- [13] Hongyan Chang and Reza Shokri. 2021. On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 292–303.
- [14] Bhushan Chaudhari, Himanshu Choudhary, Aakash Agarwal, Kamna Meena, and Tanmoy Bhowmik. 2022. FairGen: Fair Synthetic Data Generation. *arXiv preprint arXiv:2210.13023* (2022).
- [15] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*. 339–348.
- [16] Meng Chen, Li Lu, Junhao Wang, Jiadi Yu, Yingying Chen, Zhibo Wang, Zhongjie Ba, Feng Lin, and Kui Ren. 2023. Voicecloak: Adversarial example enabled voice de-identification with balanced privacy and utility. *Proceedings Of The ACM On Interactive, Mobile, Wearable And Ubiquitous Technologies* 7, 2 (2023), 1–21.
- [17] Weiyang Chen, Hongliu Yang, Xiaoyang Bi, Rong Zheng, Fusang Zhang, Peng Bao, Zhaoxin Chang, Xujun Ma, and Daqing Zhang. 2023. Environment-aware multi-person tracking in indoor environments with mmwave radars. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–29.
- [18] Yang Chen, Katherine Fennedy, Anna Fogel, Shengdong Zhao, Chao Zhang, Lijuan Liu, and Chingchiuan Yen. 2022. SSpoon: A Shape-Changing Spoon That Optimizes Bite Size for Eating Rate Regulation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 105 (sep 2022), 32 pages. doi:10.1145/3550312
- [19] Anshuman Chhabra, Karina Masalkovaitė, and Prasant Mohapatra. 2021. An overview of fairness in clustering. *IEEE Access* 9 (2021), 130698–130720.
- [20] Hyunsung Cho, Akhil Mathur, and Fahim Kawsar. 2022. Flame: Federated learning across multi-device environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–29.
- [21] Manvi Choudhary, Charlotte Laclau, and Christine Largeron. 2022. A survey on fairness for machine learning on graphs. *arXiv preprint arXiv:2205.05396* (2022).
- [22] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Halifax, NS, Canada) (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. doi:10.1145/3097983.3098095
- [23] Kate Crawford. 2017. The trouble with bias. keynote at neurips. (2017).
- [24] Jessamyn Dahmen and Diane Cook. 2019. SynSys: A synthetic data generation system for healthcare applications. *Sensors* 19, 5 (2019), 1181.
- [25] Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* 10, 7 (1998), 1895–1923.
- [26] Yasmine Djebrouni, Nawel Benarba, Ousmane Touat, Pasquale De Rosa, Sara Bouchenak, Angela Bonifati, Pascal Felber, Vania Marangozova, and Valerio Schiavoni. 2024. Bias mitigation in federated learning for edge computing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–35.
- [27] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository–German Credit.
- [28] Di Duan, Yongliang Chen, Weitao Xu, and Tianxing Li. 2024. Earse: Bringing robust speech enhancement to cots headphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–33.
- [29] Ashraf Fawzy, Tianshi David Wu, Kunbo Wang, Matthew L Robinson, Jad Farha, Amanda Bradke, Sherita H Golden, Yanxun Xu, and Brian T Garibaldi. 2022. Racial and ethnic discrepancy in pulse oximetry and delayed identification of treatment eligibility among patients with COVID-19. *JAMA internal medicine* 182, 7 (2022), 730–738.
- [30] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication* 2020-1 (2020).
- [31] GMDT Forecast et al. 2019. Cisco visual networking index: global mobile data traffic forecast update, 2017–2022. *Update* 2017 (2019), 2022.
- [32] Hana Frluckaj, Laura Dabbish, David Gray Widder, Huilian Sophie Qiu, and James D. Herbsleb. 2022. Gender and Participation in Open Source Software Development. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov 2022), 31 pages. doi:10.1145/3555190
- [33] Yang Gao, Yincheng Jin, Jagmohan Chauhan, Seokmin Choi, Jiyang Li, and Zhanpeng Jin. 2021. Voice In Ear: Spoofing-Resistant and Passphrase-Independent Body Sound Authentication. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 12 (mar 2021), 25 pages. doi:10.1145/3448113
- [34] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

- [35] Yanbin Gong, Wentao Xie, Chi Xu, Qian Zhang, and Shifang Yang. 2025. SputumLocator: Enhancing Airway Clearance with Auscultation-based Sputum Localization. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 2 (2025), 1–28.
- [36] Yanbin Gong, Qian Zhang, Bobby H.P. NG, and Wei Li. 2022. BreathMentor: Acoustic-Based Diaphragmatic Breathing Monitor System. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 53 (jul 2022), 28 pages. doi:10.1145/3534595
- [37] Erin Griffiths, Salah Assana, and Kamin Whitehouse. 2018. Privacy-Preserving Image Processing with Binocular Thermal Cameras. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 133 (jan 2018), 25 pages. doi:10.1145/3161198
- [38] Tarek Hamid, Patricia Flores, Jane Byun, Elizabeth C Courtney, Kyle C Quinn, and Amanda Watson. 2025. DermaGlow: Objective Quantification of Melanin, Erythema and Skin-tone Using Wearable Optical Spectroscopy. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 2 (2025), 1–28.
- [39] Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2020. Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 107–124. <https://aclanthology.org/2020.gebnlp-1.10>
- [40] Zhixiang He, Fengyuan Ran, Jing Chen, Yangyang Gu, Kun He, Ruiying Du, Ju Jia, and Cong Wu. 2025. HT-Auth: Secure VR Headset Authentication via Subtle Head Tremors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 3 (2025), 1–26.
- [41] Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The Weirdest People in the World? *Behavioral and Brain Sciences* 33, 2-3 (2010), 61–83.
- [42] Steven G Hershman, Brian M Bot, Anna Shcherbina, Megan Doerr, Yasbanoo Moayedi, Aleksandra Pavlovic, Daryl Waggott, Mildred K Cho, Mary E Rosenberger, William L Haskell, et al. 2019. Physical activity, sleep and cardiovascular health data for 50,000 individuals from the MyHeart Counts Study. *Scientific data* 6, 1 (2019), 24.
- [43] Laurens Holmes Jr, Michael Enwere, Janille Williams, Benjamin Ogundele, Prachi Chavan, Tatiana Piccoli, Chinacherem Chinaka, Camillia Comeaux, Lavisha Pelaez, Osatohamwen Okundaye, et al. 2020. Black–White risk differentials in COVID-19 (SARS-COV2) transmission, mortality and case fatality in the United States: translational epidemiologic perspective and challenges. *International journal of environmental research and public health* 17, 12 (2020), 4322.
- [44] Sinh Huynh, Rajesh Krishna Balan, and JeongGil Ko. 2022. IMon: Appearance-Based Gaze Tracking System on Mobile Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 161 (dec 2022), 26 pages. doi:10.1145/3494999
- [45] Towhidul Islam, Md Mahedi Hasan Rigan, Md Olid Hasan Bhuiyan, Tanzima Hashem, and Md Mahbubur Rahman. 2025. H2OPulse: Smartphone-assisted Vein Evaluation for Early Recognition of Dehydration. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 2 (2025), 1–24.
- [46] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.
- [47] Louis Henry Kamulegeya, Mark Okello, John Mark Bwanika, Davis Musinguzi, William Lubega, Davis Rusoke, Faith Nassiwa, and Alexander Börve. 2019. Using artificial intelligence on dermatology conditions in Uganda: A case for diversity in training data sets for machine learning. *BioRxiv* (2019), 826057.
- [48] Mohammed Khwaja, Sumer S. Vaid, Sara Zannone, Gabriella M. Harari, A. Aldo Faisal, and Aleksandar Matic. 2019. Modeling Personality vs. Modeling Personalidad: In-the-Wild Mobile Data Analysis in Five Countries Suggests Cultural Impact on Personality Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 88 (sep 2019), 24 pages. doi:10.1145/3351246
- [49] Minhyung Kim, Inyeop Kim, and Uichin Lee. 2021. Beneficial Neglect: Instant Message Notification Handling Behaviors and Academic Performance. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 18 (mar 2021), 26 pages. doi:10.1145/3448089
- [50] Barbara Kitchenham and Stuart Charters. 2007. *Guidelines for performing systematic literature reviews in software engineering*. Technical Report. Keele University and Durham University.
- [51] Ronny Kohavi and Barry Becker. 1996. Uci machine learning repository: adult data set. Available: <https://archive.ics.uci.edu/ml/machine-learning-databases/adult> (1996).
- [52] Julia Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2018. How We Analyzed the COMPAS Recidivism Algorithm. <https://www.propublica.org/article/how-we-analyzed-the-compass-recidivism-algorithm>
- [53] Benjamin Laufer, Sameer Jain, A. Feder Cooper, Jon Kleinberg, and Hoda Heidari. 2022. Four Years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings, and Future Prospects. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, 401–426. doi:10.1145/3531146.3533107
- [54] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsis. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2022), e1452.
- [55] Chenggao Li, Junyao Peng, Qingyong Hu, Lin Chen, Yandao Huang, Shun Wu, Weibin Cheng, and Qian Zhang. 2025. PracticalBP: Continuous Cuffless Blood Pressure Monitoring with Only One Record for Calibration. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 3 (2025), 1–31.

- [56] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2022. LASense: Pushing the Limits of Fine-Grained Activity Sensing Using Acoustic Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 21 (mar 2022), 27 pages. doi:10.1145/3517253
- [57] Huining Li, Huan Chen, Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Xiaoye Qian, Dongmei Li, Ming-chun Huang, and Wenyao Xu. 2023. Neuralgait: Assessing brain health using your smartphone. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–28.
- [58] Shuzhen Li, Yuxin Chen, Xuesong Chen, Ruiyang Gao, Yupeng Zhang, Chao Yu, Yunfei Li, Ziyi Ye, Weijun Huang, Hongliang Yi, et al. 2024. SleepNetZero: Zero-Burden Zero-Shot Reliable Sleep Staging With Neural Networks Based on Ballistocardiograms. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (2024), 1–25.
- [59] Siheng Li, Beihong Jin, Fusang Zhang, Zhi Wang, Junqi Ma, Xiaoyong Ren, and Haiqin Liu. 2024. Hypnos: A contactless sleep stage monitoring system using UWB signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–27.
- [60] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. 2022. Fairness in recommendation: A survey. *arXiv preprint arXiv:2205.13619* (2022).
- [61] Yunji Liang, Yuchen Qin, Qi Li, Xiaokai Yan, Zhiwen Yu, Bin Guo, Sagar Samtani, and Yanyong Zhang. 2022. AccMyrinx: Speech Synthesis with Non-Acoustic Sensor. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 127 (sep 2022), 24 pages. doi:10.1145/3550338
- [62] Daniyal Liaqat, Mohamed Abdalla, Pegah Abed-Esfahani, Moshe Gabel, Tatiana Son, Robert Wu, Andrea Gershon, Frank Rudzicz, and Eyal De Lara. 2019. WearBreathing: Real World Respiratory Rate Monitoring Using Smartwatches. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 56 (jun 2019), 22 pages. doi:10.1145/3328927
- [63] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How weird is CHI?. In *Proceedings of the 2021 chi conference on human factors in computing systems*. 1–14.
- [64] Chen Liu, Zixuan Dong, Li Huang, Wenlong Yan, Xin Wang, Dingyi Fang, and Xiaojiang Chen. 2024. TagSleep3D: RF-based 3D Sleep Posture Skeleton Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (2024), 1–28.
- [65] Haipeng Liu, Yuheng Wang, Anfu Zhou, Hanyue He, Wei Wang, Kunpeng Wang, Peilin Pan, Yixuan Lu, Liang Liu, and Huadong Ma. 2020. Real-Time Arm Gesture Recognition in Smart Home Scenarios via Millimeter Wave Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 140 (dec 2020), 28 pages. doi:10.1145/3432235
- [66] Tony Liu, Jennifer Nicholas, Max M. Theilig, Sharath C. Guntuku, Konrad Kording, David C. Mohr, and Lyle Ungar. 2020. Machine Learning for Phone-Based Relationship Estimation: The Need to Consider Population Heterogeneity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 145 (sep 2020), 23 pages. doi:10.1145/3369820
- [67] Xin Liu, Yuntao Wang, Sinan Xie, Xiaoyu Zhang, Zixian Ma, Daniel McDuff, and Shwetak Patel. 2022. MobilePhys: Personalized Mobile Camera-Based Contactless Physiological Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 24 (mar 2022), 23 pages. doi:10.1145/3517225
- [68] Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. 2019. On the fairness of disentangled representations. *Advances in neural information processing systems* 32 (2019).
- [69] Li Lu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Minglu Li, and Xiangyu Xu. 2019. I3: Sensing Scrolling Human-Computer Interactions for Intelligent Interest Inference on Smartphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 97 (sep 2019), 22 pages. doi:10.1145/3351255
- [70] Anna Maijala, Hannu Kinnunen, Heli Koskimäki, Timo Jämsä, and Maarit Kangas. 2019. Nocturnal finger skin temperature in menstrual cycle tracking: ambulatory pilot study using a wearable Oura ring. *BMC Women's Health* 19, 1 (2019), 1–10.
- [71] Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. 2019. Mobile Sensor Data Anonymization. In *Proceedings of the International Conference on Internet of Things Design and Implementation* (Montreal, Quebec, Canada) (*IoTDI '19*). ACM, New York, NY, USA, 49–58. doi:10.1145/3302505.3310068
- [72] Alan Mazankiewicz, Klemens Böhm, and Mario Berges. 2020. Incremental Real-Time Personalization in Human Activity Recognition Using Domain Adaptive Batch Normalization. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 144 (dec 2020), 20 pages. doi:10.1145/3432230
- [73] Lakmal Meegahapola, William Droz, Peter Kun, Amalia de Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, Tzolmon Zundui, Carlo Caprini, Daniele Miorandi, Alethia Hume, Jose Luis Zarza, Luca Cernuzzi, Ivano Bison, Marcelo Rodas Britez, Matteo Busso, Ronald Chenu-Abente, Can Günel, Fausto Giunchiglia, Laura Schelenz, and Daniel Gatica-Perez. 2023. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 176 (jan 2023), 32 pages. doi:10.1145/3569483
- [74] Tamir Mendel, Roei Schuster, Eran Tromer, and Eran Toch. 2022. Toward Proactive Support for Older Adults: Predicting the Right Moment for Providing Mobile Safety Help. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 25 (mar 2022), 25 pages. doi:10.1145/3517249

- [75] Vishwali Mhasawade, Yuan Zhao, and Rumi Chunara. 2021. Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence* 3, 8 (2021), 659–666.
- [76] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.
- [77] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, and the PRISMA Group\*. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine* 151, 4 (2009), 264–269.
- [78] Arvind Narayanan. 21. Fairness definitions and their politics. In *Tutorial presented at the Conf. on Fairness, Accountability, and Transparency*.
- [79] United Nations. 2021. With almost half of world’s population still offline, digital divide risks becoming ‘new face of inequality’, Deputy Secretary-General warns general assembly.
- [80] Ewa M Nowara, Daniel McDuff, and Ashok Veeraraghavan. 2020. A meta-analysis of the impact of skin tone and gender on non-contact photoplethysmography measurements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 284–285.
- [81] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data* 2 (2019), 13.
- [82] Chaoyi Pan, Dan Xu, Shiyuan Xu, Fuwei Wang, Chao Zheng, Bingyue Liu, and Xiaojiang Chen. 2024. Perfat: Non-contact Abdominal Obesity Assessment System. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (2024), 1–25.
- [83] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–44.
- [84] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [85] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 5884–5906. doi:10.18653/v1/2022.naacl-main.431
- [86] Akbar Ali Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. 2023. WEIRD FAccTs: How Western, Educated, Industrialized, Rich, and Democratic is FAccT?. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’23)*. ACM. doi:10.1145/3593013.3593985
- [87] Neil Shah, Neha Sahipjohn, Vishal Tambrhalli, Ramanathan Subramanian, and Vineet Gandhi. 2024. Stethospeech: Speech generation through a clinical stethoscope attached to the skin. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–21.
- [88] Farhana Shahid, Wasifur Rahman, M. Saifur Rahman, Sharmin Akther Purabi, Ayesha Seddiqa, Moin Mostakim, Farhan Feroz, Tanjir Rashid Soron, Fahmida Hossain, Nabila Khan, Anika Binte Islam, Nipi Paul, Ehsan Hoque, and A. B. M. Alim Al Islam. 2020. Leveraging Free-Hand Sketches for Potential Screening of PTSD. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 94 (sep 2020), 22 pages. doi:10.1145/3411835
- [89] Muhammad Shahzad and Shaohu Zhang. 2018. Augmenting User Identification with WiFi Based Gesture Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 134 (sep 2018), 27 pages. doi:10.1145/3264944
- [90] Qijia Shao, Junxiao Chen, Ho Man Colman Leung, Meiqi Zhao, Ruoyu Xu, Jiting Liu, Lisa Maria DiSalvo García, Xiaofan Jiang, Marie-Pierre St-Onge, and Xia Zhou. 2025. My Pillow Knows My Sleep: Sleep Monitoring with Computational Fabrics in the Pillowcase. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 3 (2025), 1–26.
- [91] Taoran Sheng and Manfred Huber. 2020. Weakly Supervised Multi-Task Representation Learning for Human Activity Analysis Using Wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 57 (jun 2020), 18 pages. doi:10.1145/3397330
- [92] Yi Sheng, Junhuan Yang, Yawen Wu, Kevin Mao, Yiyu Shi, Jingtong Hu, Weiwen Jiang, and Lei Yang. 2022. The larger the fairer? small neural networks can achieve fairness for edge devices. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*. 163–168.
- [93] Genevieve Smith. 2020. What does “fairness” mean for machine learning systems? [https://haas.berkeley.edu/wp-content/uploads/What-is-fairness\\_-EGAL2.pdf](https://haas.berkeley.edu/wp-content/uploads/What-is-fairness_-EGAL2.pdf)
- [94] Yingjian Song, Haotian Xiang, Zixuan Zeng, Jiayu Chen, Yida Zhang, Zaid Farooq Pitafi, He Yang, Qin Lu, Xiang Zhang, Bradley G Phillips, et al. 2025. Multi-granularity Supervised Contrastive Learning with Online Adaptation for Contactless In-bed Posture Classification. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 2 (2025), 1–32.
- [95] Dimitris Spathis, Ignacio Perez-Pozuelo, Soren Brage, Nicholas J Wareham, and Cecilia Mascolo. 2021. Self-supervised transfer learning of physiological representations from free-living wearable data. In *Proceedings of the Conference on Health, Inference, and Learning*. 69–78.
- [96] Jie Su, Zhenyu Wen, Tao Lin, and Yu Guan. 2022. Learning Disentangled Behaviour Patterns for Wearable-Based Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 28 (mar 2022), 19 pages. doi:10.1145/3517252

- [97] Weigao Su, Daibo Liu, Taiyuan Zhang, and Hongbo Jiang. 2022. Towards Device Independent Eavesdropping on Telephone Conversations with Built-in Accelerometer. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 177 (dec 2022), 29 pages. doi:10.1145/3494969
- [98] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976* (2019).
- [99] Wiebke Toussaint, Akhil Mathur, Aaron Yi Ding, and Fahim Kawsar. 2022. Tiny, always-on and fragile: Bias propagation through design choices in on-device machine learning workflows. *arXiv preprint arXiv:2201.07677* (2022).
- [100] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. 2017. Recognizing detailed human context in the wild from smartphones and smartwatches. *IEEE pervasive computing* 16, 4 (2017), 62–74.
- [101] Raghav H. Venkatnarayan and Muhammad Shahzad. 2018. Gesture Recognition Using Ambient Light. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 40 (mar 2018), 28 pages. doi:10.1145/3191772
- [102] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. 2022. In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)* (2022).
- [103] Dequan Wang, Xinran Zhang, Kai Wang, Lingyu Wang, Xiaoran Fan, and Yanyong Zhang. 2024. Rdgait: A mmwave based gait user recognition system for complex indoor environments using single-chip radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–31.
- [104] Jianyang Wang, Dongheng Zhang, Binbin Zhang, Jinbo Chen, Yang Hu, and Yan Chen. 2024. Rf-gymcare: Introducing respiratory prior for rf sensing in gym environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–28.
- [105] Liang Wang, Aoyang Chang, Boqun Liu, Zhiwen Yu, Pengfei Hu, Jia Liu, Yao Zhang, and Bin Guo. 2025. mmHSE: Enhanced Eavesdropping Attack on Headsets Leveraging COTS mmWave Radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 2 (2025), 1–27.
- [106] Liang Wang, Wen Cheng, Lijia Pan, Tao Gu, Tianheng Wu, Xianping Tao, and Jian Lu. 2018. SpiderWalk: Circumstance-Aware Transportation Activity Detection Using a Novel Contact Vibration Sensor. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 42 (mar 2018), 30 pages. doi:10.1145/3191774
- [107] Lei Wang, Wei Li, Ke Sun, Fusang Zhang, Tao Gu, Chenren Xu, and Daqing Zhang. 2022. LoEar: Push the Range Limit of Acoustic Sensing for Vital Sign Monitoring. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 145 (sep 2022), 24 pages. doi:10.1145/3550293
- [108] Lei Wang, Xingwei Wang, Xi Zhang, Xiaolei Ma, Yu Zhang, Fusang Zhang, Tao Gu, and Haipeng Dai. 2025. AccCall: Enhancing Real-time Phone Call Quality with Smartphone’s Built-in Accelerometer. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 3 (2025), 1–33.
- [109] Yao Wang, Yufei Chen, Haibin Zhang, Minjie Lyu, and Tao Gu. 2025. S2Pair: Secure and Simple Device Pairing Using Human Bone as a Key Generator. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 2 (2025), 1–29.
- [110] Chatchai Wangwiwattana, Xinyi Ding, and Eric C. Larson. 2018. PupilNet, Measuring Task Evoked Pupillary Response Using Commodity RGB Tablet Cameras: Comparison to Mobile, Infrared Gaze Trackers for Inferring Cognitive Load. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 171 (jan 2018), 26 pages. doi:10.1145/3161164
- [111] Mark Weiser. 1993. Some computer science issues in ubiquitous computing. *Commun. ACM* 36, 7 (1993), 75–84.
- [112] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning non-discriminatory predictors. In *Conference on Learning Theory*. PMLR, 1920–1953.
- [113] Ziyu Wu, Fangting Xie, Yiran Fang, Zhen Liang, Quan Wan, Yufan Xiong, and Xiaohui Cai. 2024. Seeing through the tactile: 3d human shape estimation from temporal in-bed pressure images. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–39.
- [114] Yi Xiao, Harshit Sharma, Zhongyang Zhang, Dessa Bergen-Cico, Tauhidur Rahman, and Asif Salekin. 2024. Reading between the heat: Co-teaching body thermal signatures for non-intrusive stress detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–30.
- [115] Chenhan Xu, Huining Li, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Xingyu Chen, Kun Wang, Ming-chun Huang, and Wenyao Xu. 2021. CardiacWave: A MmWave-Based Scheme of Non-Contact and High-Definition Heart Activity Computing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 135 (sep 2021), 26 pages. doi:10.1145/3478127
- [116] Jie Xu, Yunyu Xiao, Wendy Hui Wang, Yue Ning, Elizabeth A Shenkman, Jiang Bian, and Fei Wang. 2022. Algorithmic fairness in computational medicine. *EBioMedicine* 84 (2022), 104250.
- [117] Wei Xu, ZhiWen Yu, Zhu Wang, Bin Guo, and Qi Han. 2019. AcousticID: Gait-Based Human Identification Using Acoustic Signal. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 115 (sep 2019), 25 pages. doi:10.1145/3351273
- [118] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigy Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S Kuehn, Jeremy F Huckins, Margaret E Morris, et al. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the*

- ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–34.
- [119] Joel Xue and Robert M Farrell. 2014. How can computerized interpretation algorithms adapt to gender/age differences in ECG measurements? *Journal of Electrocardiology* 47, 6 (2014), 849–855.
- [120] Baichen Yang, Xinyi Zhang, Jiayi Zhang, Zirui Huang, Qiqi Lu, Jin Zhang, Hai Hu, and Qian Zhang. 2024. KneeGuard: A Calibration-free Wearable Monitoring System for Knee Osteoarthritis Gait Re-training via Effortless Wearing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (2024), 1–29.
- [121] Xiaoying Yang, Xue Wang, Gaofeng Dong, Zihan Yan, Mani Srivastava, Eiji Hayashi, and Yang Zhang. 2023. Header: Sensing head gestures for confirmation dialogs on smartwatches with wearable millimeter-wave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–28.
- [122] Sofia Yfantidou, Christina Karagianni, Stefanos Efstathiou, Athena Vakali, Joao Palotti, Dimitrios Panteleimon Giakatos, Thomas Marchioro, Andrei Kazlouski, Elena Ferrari, and Šarūnas Girdzijauskas. 2022. LifeSnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild. *Scientific Data* 9, 1 (2022), 663.
- [123] Sofia Yfantidou, Pavlos Sermpezis, Athena Vakali, and Ricardo Baeza-Yates. 2023. Uncovering Bias in Personal Informatics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023). doi:10.1145/3610914
- [124] Shichao Yue, Yuzhe Yang, Hao Wang, Hariharan Rahul, and Dina Katabi. 2020. BodyCompass: Monitoring Sleep Posture with Wireless Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 66 (jun 2020), 25 pages. doi:10.1145/3397311
- [125] Hanbin Zhang, Chenhan Xu, Huining Li, Aditya Singh Rathore, Chen Song, Zhisheng Yan, Dongmei Li, Feng Lin, Kun Wang, and Wenyao Xu. 2019. PDMove: Towards Passive Medication Adherence Monitoring of Parkinson’s Disease Using Smartphone-Based Gait Assessment. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 123 (sep 2019), 23 pages. doi:10.1145/3351281
- [126] Hanbin Zhang, Li Zhu, Viswam Nathan, Jilong Kuang, Jacob Kim, Jun Alex Gao, and Jeffrey Olgin. 2021. Towards Early Detection and Burden Estimation of Atrial Fibrillation in an Ambulatory Free-Living Environment. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 86 (jun 2021), 19 pages. doi:10.1145/3463503
- [127] Shibo Zhang, Yuqi Zhao, Dzung Tri Nguyen, Runsheng Xu, Sougata Sen, Josiah Hester, and Nabil Alshurafa. 2020. NeckSense: A Multi-Sensor Necklace for Detecting Eating Activities in Free-Living Conditions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 72 (jun 2020), 26 pages. doi:10.1145/3397313
- [128] Yunke Zhang, Fengli Xu, Tong Xia, and Yong Li. 2022. Quantifying the Causal Effect of Individual Mobility on Health Status in Urban Space. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 193 (dec 2022), 30 pages. doi:10.1145/3494990
- [129] Juntao Zhou, Yijie Li, Yida Wang, Dian Ding, Yu Lu, Yi-Chao Chen, and Guangtao Xue. 2024. Visar: Projecting Virtual Sound Spots for Acoustic Augmented Reality Using Air Nonlinearity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–30.
- [130] Pengyuan Zhou, Hengwei Xu, Lik Hang Lee, Pei Fang, and Pan Hui. 2022. Are You Left Out? An Efficient and Fair Federated Learning for Personalized Profiles on Wearable Devices of Inferior Networking Conditions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 91 (jul 2022), 25 pages. doi:10.1145/3534585
- [131] Matthew Zook, Solon Barocas, Danah Boyd, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, Rachele Hollander, Barbara A Koenig, Jacob Metcalf, et al. 2017. Ten simple rules for responsible big data research. *PLoS computational biology* 13, 3 (2017), e1005399.

Received 27 September 2024; revised 6 January 2026; accepted 24 January 2026