

A real-time hot flash detection system for noisy free-living conditions

Andreas Tsiartas

SRI International

Alan Taitz

SRI International

Nicole Arra

SRI International

Fiona Claire Baker

fiona.baker@sri.com

SRI International

Massimiliano de Zambotti

SRI International

Ann Garnier

SRI International

Article

Keywords: Hot flashes, real-time detection, free-living conditions, menopause, wearables

Posted Date: July 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9974484/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: Competing interest reported. The authors declare the following potential competing interest: Massimiliano de Zambotti is currently an employee of Oura Ring Inc. Massimiliano was an employee at SRI International at the time of his contribution to this study, which predates his transition to Oura. Massimiliano de Zambotti's contributions to this work, as well as the statements and opinions expressed, are solely the responsibility of the author and do not represent the official views of Oura Ring Inc. The remaining authors declare no competing interests.

A real-time hot flash detection system for noisy free-living conditions

ANDREAS TSIARTAS, ALAN TAITZ, NICOLE ARRA, FIONA C. BAKER, MASSIMILIANO DE ZAMBOTTI, ANN GARNIER

Abstract — Objective: Hot flashes (HFs) are the most bothersome menopausal symptom. Currently, objective tracking of HFs is limited to research settings, restricting advances in the management of this disruptive symptom. Prior laboratory work from our group developed a machine learning (ML) HF-detection system based on data inputs from sensors commonly found in wearable technology (Forouzanfar et al., 2018; Tsiartas et al., 2021). The current work aims to further investigate and optimize the performance of the novel HF-detection system in real-time and under free-living conditions using data (pulse rate, motion, skin temperature, and skin conductance) collected with a commercially available multi-sensor wrist wearable in a sample of eleven menopausal individuals. The performance of the system was compared against gold-standard expert HF annotations. **Results:** System performance achieved >90% accuracy in real-time detection of HFs, under free-living conditions, including periods of sleep/wake, activity/no activity, and different environmental conditions. Algorithm parameters (e.g., multi-feature resolution) that minimize computations while maintaining >90% accuracy are also provided. **Conclusion:** The HF-detection system can operate in a real-world setting while maintaining high performance, suggesting its feasibility for use in daily life using commercial-grade wrist sensors. **Significance:** An HF-detection system can provide key insights about HF patterns and characterization, both for individuals with HFs and for clinicians seeking to improve the HFs management, as well as to better understand this highly common and disruptive menopausal symptom.

Index Terms — Hot flashes, real-time detection, free-living conditions, menopause, wearables

Andreas Tsiartas (PhD), Alan Taitz (PhD), Nicole Arra (BS), Fiona C. Baker (PhD), SRI International, CA, USA. Ann Garnier (BS), Lisa Health Inc. CA, USA. Massimiliano de Zambotti* (PhD), SRI International and Lisa Health Inc. CA, USA (correspondence e-mail: fiona.baker@sri.com).

*MdZ is currently an employee of Ouraring Inc. MdZ was an employee at SRI International at the time of his contribution to this study, which predates his transition to Oura. MdZ's contributions to this work, as well as the statements and opinions expressed, are solely the responsibility of the author and do not represent the official views of Ouraring Inc.

I. INTRODUCTION

Hot flashes (HFs) are a hallmark of the menopausal transition and post-menopause and are characterized by peripheral vasodilation and sweating (Freedman, 2001). HFs affect upward of 80% of individuals during the menopause transition,

persist for several years, are disruptive to daily life and sleep, and are associated with a cascade of adverse health consequences (Bansal & Aggarwal, 2019). Commercial technology for the measurement of HFs is limited, a major restriction in menopause symptom management.

The expert gold-standard method for measuring HFs is based on human evaluation of increases in sternal skin conductance (SC) recorded using laboratory/ambulatory research-grade devices (Carpenter et al., 1999). Recent advancements in the field introduced automatic algorithms for HF classification that used various methods, including feature engineering such as sternum SC signal processing features (Forouzanfar et al., 2018), as well as various machine learning (ML) methods such as support vector machine and decision trees (Thurston et al., 2009). More recently, Naghavi et al. developed a skin-conductance-based approach for predicting HF onset before subjective perception, using early rising-edge SC features to support just-in-time intervention delivery. Their findings suggest that HF onset prediction from physiological signals is feasible, while also highlighting the need for future multimodal approaches that incorporate additional wearable signals such as activity, temperature, and heart rate. All these systems are developed using research-grade equipment in controlled environments, limiting their real-world applicability, and are based on inputs from the SC signal. In addition, these works used shuffled data in a cross-validation setup. Limited work has been done to validate these systems in unseen sets of women, with separate testing datasets which enable verifying the predictive power of the model and ensuring the model is not memorizing the training or cross-validation data.

Among the few available options, the UFI Biolog system (UFI, Morro Bay, CA) is considered the gold standard ambulatory device for measuring HFs, using a single channel recording of SC levels via wired wet electrodes applied on the chest. The device is bulky, has limited recording capacity, is unsuitable for extensive recordings, and is cost-prohibitive for large-scale use. It also requires expertly trained staff to operate. Usually, SC is analyzed, and HF classification is performed offline, using a desktop computer and specialized software. Additional limitations include discomfort associated with using such devices (e.g., skin irritation and itching (Carpenter et al., 2012)), lack of user interfaces and wireless communication protocols, and lack of consumer-oriented designs and form factors, among others.

Nowadays, multi-sensor wearable technology offers a pathway to large-scale deployment of an automatic system for measuring HFs in free-living conditions, due to increased computational power, wireless communication protocols, upgraded multi-sensor capability (e.g., motion, pulse rate [PR], SC, and temperature sensors), suitable for capturing the physiology of HFs, while also capable of measuring an

individual's bio-behavioral state (e.g., sleep, activity). However, the development of a cost-effective system for measuring and characterizing HFs is challenged by numerous factors beyond technology readiness. Menopause is experienced differently among individuals, with large variability in symptom manifestation and dynamics between and within individuals over time. In addition, there are also between- and within-differences in physiological correlates of HFs, and HF physiological responses are dependent on the context in which HFs occur (e.g., physiological measures around HFs are influenced by arousals during sleep) (Baker et al., 2019).

Our previous laboratory work introduced and demonstrated the feasibility and performance of a novel ML HF-detection system based on integrating multi-sensor features from common signals used by consumer-grade wearables. We also showed how using multi-sensor features is important when HFs occur in different bio-behavioral states (e.g., sleep vs. awake) or across simulated conditions of signal corruption (Tsiartas et al., 2021).

In the current work, we build on our prior research (Tsiartas et al., 2021) and evaluate the performance of a commercially available multi-sensor device, i.e., Empatica E4, in measuring physiological HFs in women in free-living conditions. In addition, we further adapt the algorithm to work in real time, evaluate parameters affecting computational efficiency, and provide an in-depth analysis of the performance under real and simulated free-living conditions (e.g., sensor loss of contact and other types of signal noise). The expert evaluation of HFs based on sternum SC fluctuation collected via a research-grade SC monitor was used as the gold standard for HF annotations.

II. EXPERIMENTAL DESIGN

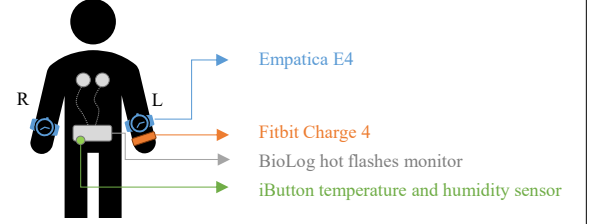
After receiving IRB approval (Pro00053383, Pro00055129), eleven women (40-65 years old) who reported ≥ 5 HFs per 24-h period were recruited from the community (San Francisco Bay Area, CA, USA) via flyers and social media advertisement. Participants underwent a brief phone screen to evaluate major eligibility criteria. All participants reviewed and signed the informed consent form for the study. Participants were free from severe mental health (e.g., major depression) and/or medical (e.g., cardiovascular diseases) conditions. They were not currently taking current medications known to affect sleep and/or the cardiovascular system (e.g., antihypertensives) and were not using hormone therapy. All participants had a body mass index of less than 35 kg/m², and none of the women was pregnant or breastfeeding.

Devices and instructions for use were provided to participants for the at-home data collection protocol, which lasted 48 hours (see data sources and specifications in Table 1 and a sample of multimodal data collection in Fig. 1).

The following devices were used for data collection:

- **3991x/1-SCL BioLog HF monitor** (UFI, Morro Bay, CA), the gold standard device for ambulatory HF monitoring. SC was collected using two 1.5 cm-diameter Ag/AgCl electrodes

Table 1. Data types, sources, and resolution. *SC*, skin conductance; *PR*, pulse rate; *PPG*, photoplethysmography; *ACC*, accelerometer; *BVP*, Blood Volume Pulse; *L*, Left; *R*, right.

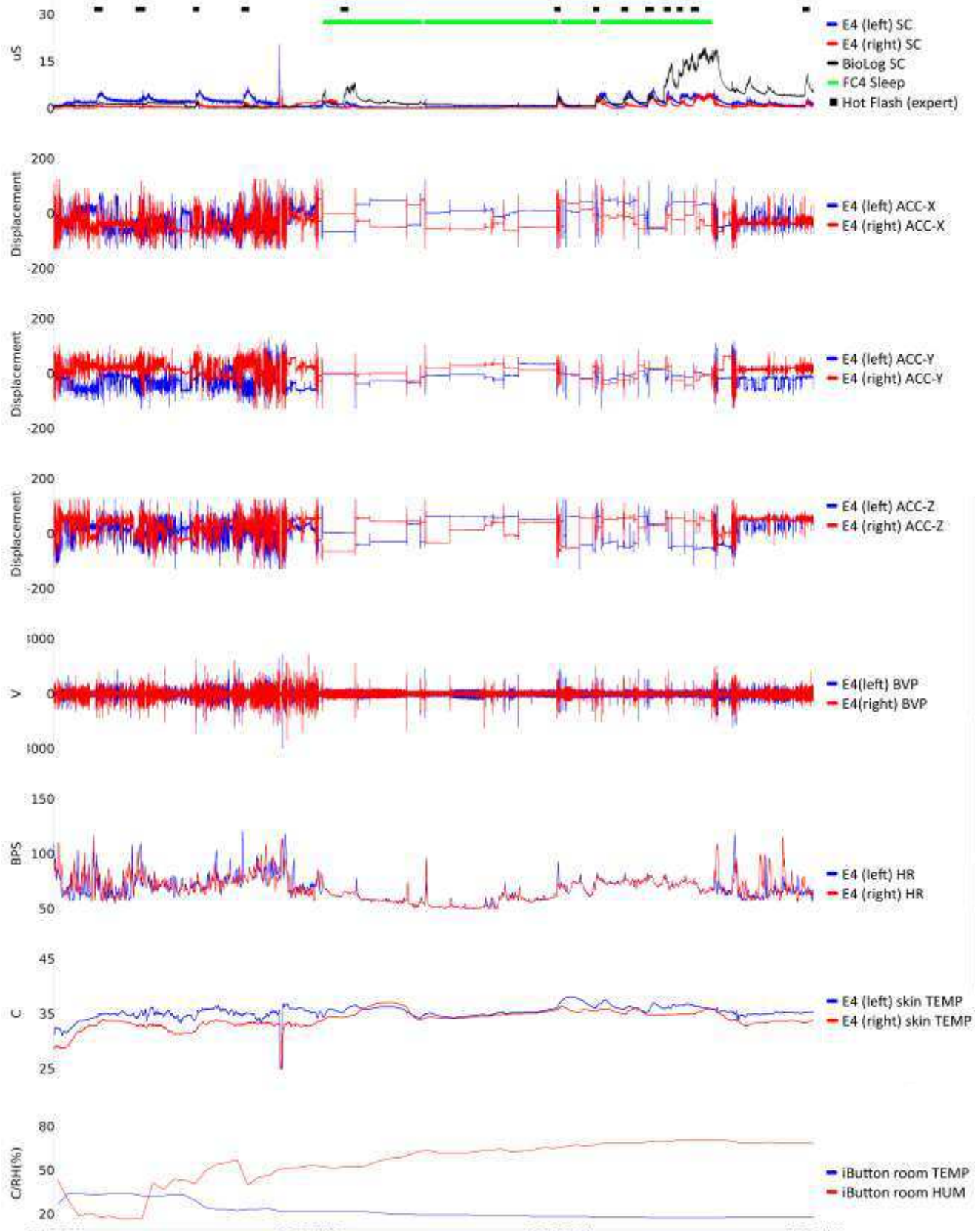


Data type	Data source	Resolution
Physiological	BioLog SC (μ S)	1 Hz
	E4 (L/R) SC (μ S)	4 Hz
	E4 (L/R) PR (bpm)	1 Hz
	E4 (L/R) PPG BVP	64 Hz
Behavioral	FC4 (L) sleep (wake, light, deep, rem)	30 s
	FC4 (L) steps (No.)	1 min
	E4 (L/R) ACC (x, y, z) (1/64g)	64 Hz
	E4 (L/R) tags (timestamp)	-
Environmental	iButton temperature (Celsius)	10 min
	iButton humidity (% relative humidity)	10 min

placed on either side of the sternum (about 4 cm apart; a 0.5-V constant voltage circuit was maintained between them).

- **Fitbit Charge 4** (FC4; Fitbit LLC.), a commercially available, validated, wrist-worn multi-sensor wearable device for sleep measurement (sleep/wake differentiation and sleep stages classification) and activity tracking (e.g., steps, time spent in different activity levels) (de Zambotti et al., 2019). FC4 was placed on the left wrist.
- **Empatica E4** (Empatica Inc.), a research-grade wrist-worn wearable device providing continuous raw data collection from photoplethysmography (PPG), accelerometry (ACC), SC, and skin temperature sensors. The E4 was placed on both wrists (left and right). In addition, this device directly provided PR estimates computed every second.
- **DS1923 Environmental Temp/Humidity Logger iButton** (Maxim Integrated Products Inc.), a miniaturized sensor measuring environmental temperature and humidity. Devices were taped to the pouch containing the BioLog device. For the duration of the study, participants were asked to press an event-marker button on both E4 devices when perceiving an HF. They were also asked to synchronize the FC4 each morning. Participants were instructed to remove the devices while showering, bathing, swimming, or related activities. At the end of the data collection period (~48 h), they were asked to return the equipment via prepaid packages. Participants were compensated for their participation in this study.

Fig. 1 Example of multimodal data collection (Empatica E4, BioLog, Fitbit Charge 4 [FC4], and iButton data) in free-living conditions (12h), from a study participant. *SC*, skin conductance; *PR*, pulse rate; *PPG*, photoplethysmography; *ACC*, accelerometer; *BVP*, Blood Volume Pulse. Black marks at the top indicate HFs, identified from gold-standard SC monitoring, classified by an expert.



III. DATA PROCESSING

Data access. E4 Manager software (provided by Empatica to manage the E4 wristwatch) was used to initialize E4 data collection and retrieve data from the Empatica cloud. Only one E4 device, placed on the left wrist, was used for data processing. The other E4 device was used solely for capturing self-reported HF tags for supporting the expert HF annotations (see below). The Fitabase (Small Steps Labs LLC.) web platform was used to collect and access high-granularity FC4 data. OneWireViewer software was used to initialize the iButton data collection and retrieve the data. UFI FlashTrax 2.1 software was used to initialize the BioLog device and retrieve BioLog data. No personal or identifiable information was collected during the device setup.

HF annotations. Expert HF annotations (timestamp = onset of the flash) were based on manual evaluation of fluctuations in sternal SC ($>1.5\text{-}\mu\text{mho}$ rises within 30-s) in the BioLog SC signal and used as the gold standard for comparison with the algorithm output (Baker et al., 2019). Self-reported annotations (timestamp = user tags) were considered only when tags were present on both E4 devices and were used as supporting information for the evaluation of ambiguous cases. Expert HF annotations were converted to the model sampling grid by matching annotation times to the nearest E4 sampling indices, producing binary HF labels aligned with the extracted feature intervals.

Data alignment and down-sampling. At the beginning of each recording, all devices (FC4, iButton, BioLog, E4) were time synchronized. To enable efficient large-scale storage and processing, we saved the SC and temperature (T) signals down-sampled by a factor 4 and the ACC data down-sampled by a factor of 32 (we found in our analysis that such down-sampling does not result in information loss for HF detection). Then, using the E4 start time and sampling rate, we computed the timestamp of each sample. By combining the annotation time stamps, the start time of the gold-standard device, and the timestamp of each sample, we associated the expert annotation timestamps with all the time-aligned signals from the E4 device for further analysis. Finally, the FC4 sleep annotations (30-s resolution) were time-aligned with both E4 signals and HF annotations using the starting timestamp as a reference time point (Fig. 1). The final starting and ending timestamps used for analysis for each recording are based on manual timestamps to ensure all devices were properly set up. This manual step allowed us to exclude any time points with missing device or annotation data.

IV. SYSTEM DESCRIPTION

The proposed ML HF-detection system aimed to improve the system proposed in our prior work (Tsiartas et al., 2021). To achieve that goal, the following steps were taken: (1) process HFs in real time using sensor data from a commercial wearable device (i.e., Empatica E4), (2) minimize computations while maintaining the system performance at $>90\%$ accuracy, and (3) evaluate the performance of the system under real-world

environmental conditions during free-living recordings. Thus, we modified the context and relative information used in differential features of our prior system (Forouzanfar et al., 2018). We also removed features used in the prior system (Forouzanfar et al., 2018), that require processing the entire HF waveform. In addition, we analyzed the system parameters in depth, fine-tuned the features and parameters, and trained the ML system. In contrast to our prior work, we used the PR estimate provided by the E4 device rather than processing PR ourselves. The ML HF-detection system was retrained using E4 raw sensor data. The classifier computes each feature every 15 seconds (default) using a windowing approach (left and right contexts were used). Below, we describe the processed feature sets from Empatica E4 multi-sensor data:

- **SC features** include (1) the differential area under the curve (AUC) of the SC signal; we calculated the AUC difference between the last X s and the future window ($\pm Y$ s). The number of seconds X and Y were determined experimentally and (2) the derivative described in step 7 (SC slope within a time window from -30 s to +30 s around HF candidates) from our prior work (Forouzanfar et al., 2018). We modified the derivative to use the last X s and the future window ($\pm Y$ s). The SC feature set represents both SC response changes.
- **The temperature feature set** computes the participant's average temperature differential between the prior X and future Y seconds and the current context window (30 seconds past context). The differential temperature feature captures temperature changes around an HF event. X and Y context, in seconds, were determined experimentally.
- **The PPG feature set:** We used the E4 wrist HR estimation (second resolution). We then estimated the differential between the current HR window (30 s past context) and HR window averaged using X past context and Y future context.
- **The motion feature set** captures movements of the subject in the x , y , and z dimensions. We process each dimension separately and extract the absolute maximum displacement (AMD) using window X for past context and Y for future context for each dimension. For the x dimension, we used the raw AMD. In addition, we computed the differential AMD between y - x , z - x , and z - y . The first variable in the differential calculation used a fixed 30-s past window while the second variable has X (seconds) context in the past and Y (seconds) context in the future. X and Y context, in seconds, is determined experimentally.

We time-aligned the features with the expert HF annotations for detection and evaluation (60 seconds in the past and 90 seconds in the future matching window). Since all features are processed independently, the system enables sensor-specific feature selection. The selected features were fed into an ML decision tree classifier, which generated a binary HF/non-HF prediction for each 15-s interval. Interval-level predictions were then post-processed to identify discrete HF detections. A new HF detection was assigned at the onset of a positive prediction sequence, defined as a transition from 0 to 1 in the classifier output. To avoid counting multiple detections within the same physiological HF episode, subsequent detections were suppressed if they occurred within 60 prediction intervals,

corresponding to a 15-min refractory period. The resulting post-processed detections were used for subject-level HF output and evaluation.

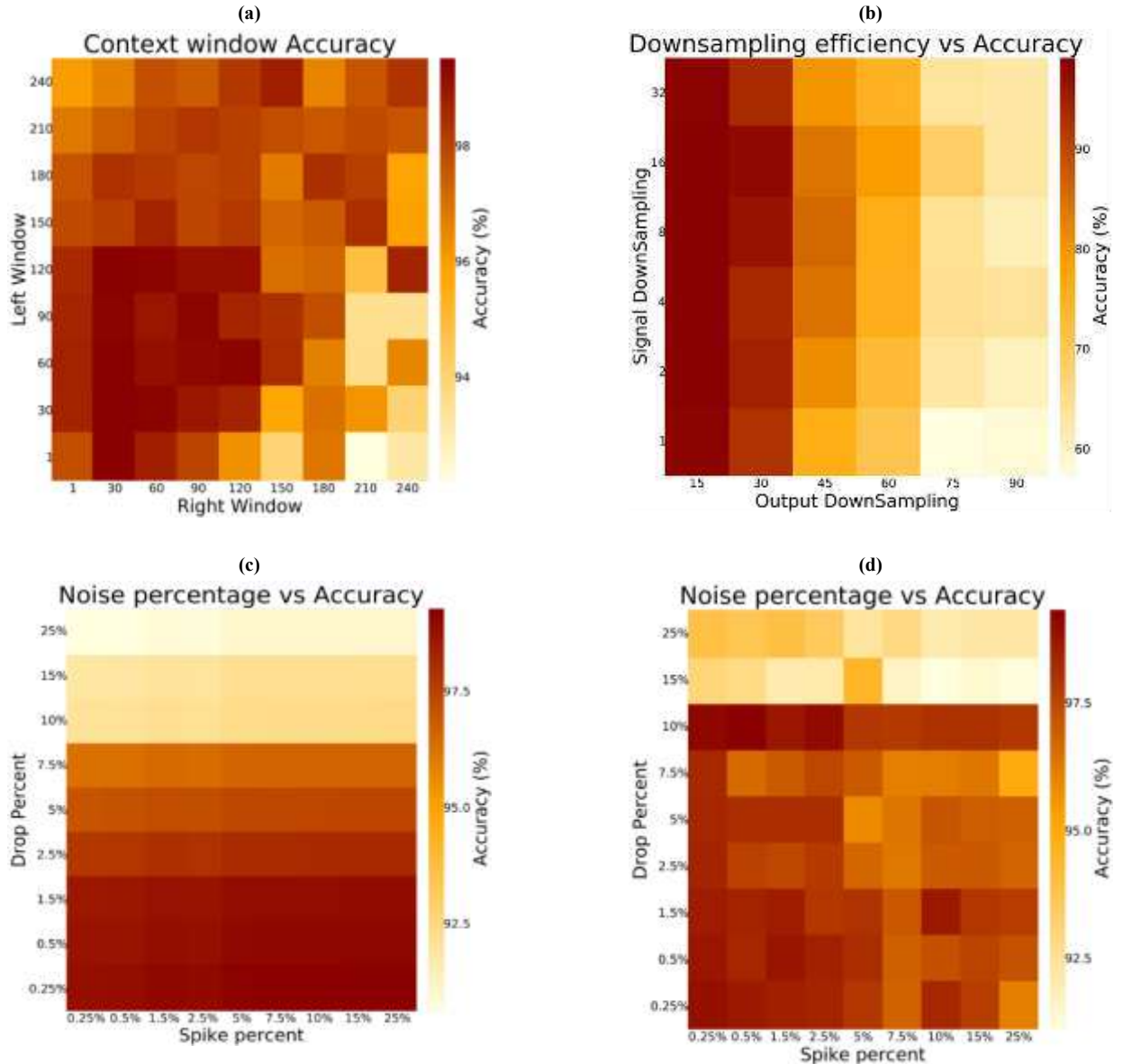
V. ANALYSIS

All analyses were based on a final dataset of 366 physiological HFs recorded in free-living conditions over a combined total of 516.4 hours from eleven women. We split the dataset into training (126 HFs from four women; Age, mean $\pm$ SD: 56 ± 2.16 years; 153.3 hours of data) and testing (240 HFs from seven women; age, mean $\pm$ SD: 56.9 ± 4.2 years; 363.13 hours

of data). Model building and model application were implemented as separate steps. During model building, E4-derived features were aligned with expert HF annotations and used to train a decision tree classifier on the training dataset. The trained model was then serialized and applied to the held-out testing dataset to generate binary HF/non-HF predictions at 15-s intervals. Feature-window and down-sampling parameters were evaluated using the training dataset before final testing.

System performance was evaluated in terms of accuracy, defined as the average sensitivity (percent of true positives) and specificity, (percent of true negatives).

Fig. 2 The 3D intensity plot shows accuracy for the real-time system operation as a function of the amount of context (right and left context) used (a); The 3D intensity plot shows accuracy for the real-time system operation as a function of signal down-sampling (b); The 3D intensity plot shows accuracy for the real-time system operation as a function of percentage of induced noise (signal drops and spikes), without system re-training (c); The 3D intensity plot shows accuracy for the real-time system operation as a function of percentage of induced noise (signal drops and spikes), after retraining the system (d).



A predicted HF was considered a true positive if the predicted HF region overlapped with or occurred within X seconds of an expert-annotated HF onset. Multiple detections within the same matching window were counted as a single detected event. Predicted HF regions not associated with an expert-annotated HF were counted as false positives. Expert-annotated HFs without a corresponding predicted region were counted as false negatives. Analyses were organized by tasks and were designed to answer the following questions:

Task 1- What is the system performance under real-time operation, and what is the optimal window size for past/future information used?

We used the system described in section IV, and evaluated the performance for each context size: 1, 30, 60, 90, 120, 150, 180, 210, and 240 s for past and future context. The usage of the amount of past and/or future information used was kept the same across features, except for the first feature component of the differential value calculation (e.g., SC AUC), which reflects time deviation (30 s past information, i.e., the subtractors was kept constant). We ran a 4-fold cross-validation setup using a training dataset by excluding a subject in each fold (leave one subject out LOSO) and report the average sensitivity and specificity of the system for all the context windows.

Task 2 - What is an optimal computationally efficient configuration that does not compromise the real-time performance found in task 1?

We started with the optimal real-time system identified in task 1 and evaluated the performance for different raw sensor input down-sampling factors (1, 2, 4, 8, 16, and 32) and varied the timing for algorithm computation (15 s, 30 s, 45 s, 60 s, 75 s, 90 s) resulting in 6 by 6 contrasts. We down-sampled all the sensor inputs, except for the HR which is directly processed by Empatica E4 at 1-s resolution. As in task 1, we ran a 4-fold cross-validation setup using the training dataset and report the average sensitivity and specificity of the system for all the computation-reducing parameters.

Task 3 - How much performance loss do we have under common noisy sensor conditions (spikes and signal loss)?

Using the optimal system in task 2 (15 s timing for algorithm computation and down-sampling by a factor 4, resulting in 1-Hz sampling across sensors), we evaluated algorithm performance on the data augmented by synthetic noise by type (“spikes” and “drops”) and amount (0.5%, 1%, 3%, 5%, 10%, 15%, 20%, 30%, 50%), designed to simulate common artifacts. As in task 2, we ran a 4-fold cross-validation setup using the training dataset and report the accuracy of the system for all noise types and parameters.

Task 4 – Does training the system with noisy conditions recover performance loss?

Using the best system parameters from task 2, we trained the algorithm using the synthetically augmented training dataset and retested it for performance. We ran a 4-fold cross-validation setup using the training dataset.

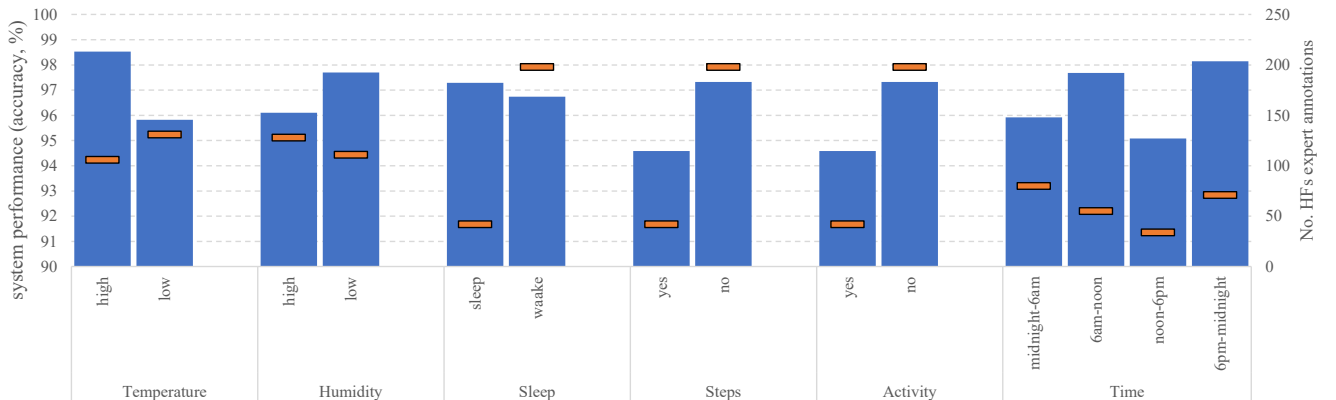
Task 5 - What is the system performance under different bio-behavioral and environmental conditions during free-living recordings?

We used the system trained in task 4 and evaluated algorithm performance across different conditions: sleep vs. wake (classified based on FC4), the occurrence of HFs across 24 h (6-hour window), presence/absence of activity (classified based on Fitbit steps data; steps vs. no steps), median split of external temperature and humidity (from iButton) across the total recording period from the validation dataset. The system was trained on all training data, and the results are reported on the validation dataset.

VI. RESULTS AND DISCUSSION

Task 1. System performance is above 90% accuracy under all real-time parameters evaluated (fixed context windows across all features). The best performance under real-time system operation was achieved using 60 s of left context (Fig. 2a). In

Fig. 3 Performance of the system (accuracy, %) across different bio-behavioral and environmental free-living conditions (temperature, humidity, sleep, steps, activity, time). The number of experts annotated hot flashes (HFs), for each condition, are provided on the secondary axis (dash).



contrast, the best system performance using both left and right context (off-line system operation) was reached at 30 s of past and 30 s of future information use (Fig. 2a). From this analysis, we learned that exceeding 120 s of context across features reduced resolution for predicting HFs. Thus, less than 60 s resolution seems ideal to guarantee optimal system performance. When also using future information, the left and right context needs to be balanced to guarantee optimal system performance.

The final algorithm chosen for the next task used 60 s of left context that enables real-time operation and maximizes performance without using future context. This real-time configuration is also consistent with recent work by Naghavi et al. who showed that early rising-edge SC features can predict HF onset before subjective perception for just-in-time intervention delivery. In contrast to that SC-only approach, the present system evaluates HF identification using a commercial wrist-worn multisensor device that incorporates SC, temperature, pulse-derived, and motion features under free-living conditions.

Task 2. The results indicated that 90% accuracy was met for timing for algorithm computation of <30 s. Down-sampling below 4Hz does not seem to meaningfully impact performance (Fig. 2b).

The final algorithm chosen to progress with the next task used a 15-s algorithm computation interval and down-sampling by a factor of 4, resulting in 1-Hz sampling across sensors. This approach minimizes the impact on real-time delay (a factor of 4, i.e., 1Hz, requires 1-s delay to collect the samples while a factor of 32, i.e., 1/8Hz, requires 8-s delay to collect the samples).

Task 3. Results indicated that 90% accuracy is met for a signal drop of up to 15%. “Spikes” did not substantially affect system performance (Fig. 2c). For higher signal drops (>15%), information loss was significant, resulting in reduced performance.

Task 4. Results indicated that 90% accuracy was maintained for signal drops of up to 15%. Compared with the previous system (Task 3), we achieved an absolute increase of about 6% in accuracy after training the algorithm with the synthetically augmented training dataset (relative reduction in error of about 60%) for a signal drop of 15%. We also evaluated the amount of noise that optimized the performance in the dataset without noise augmentation. The best performance of the algorithm was obtained with a training configuration of 0.25% spike and 0.25% drop, achieving sensitivity, specificity, and accuracy of 99.14%, 99.17%, and 99.16% (Fig. 2d). Thus, we use this trained system configuration as the default system for the next task.

Task 5. We validated our results by applying the final trained system from Task 4 using the validation dataset. Fig. 3 shows the distribution of the physiological HFs (expert annotation) and the system performance (accuracy, %) in 7 women across the different bio-behavioral and environmental conditions during free-living recordings.

Overall, the system is robust across the different free-living conditions analyzed. These results show that the accuracy is

higher than 90% across different conditions and, thus, shows the robustness of the system across a variety of setups, environmental and free-living conditions.

VII. CONCLUSION

In this work, we conducted an in-depth investigation of an ML classifier used to predict HFs in midlife women under free-living conditions using commercial-grade wrist sensors. We conclude that the system can operate in a real-world setting while maintaining high performance (90% accuracy across all conditions tested) supporting its feasibility for measuring HFs during daily life using commercial-grade wrist sensors.

Further work should consider additional critical factors that may limit system performance. For example, elevated body mass index and dark skin tone have been shown to reduce the accuracy of PPG-derived blood volume readings (Fine et al., 2021). These are among the factors that must be accounted for to avoid biases in the HF-detection system performance. HF physiology may change over time, as a function of demographic factors such as age, but also in relation to an individual’s psychophysiological state, such as stress. These factors need to be considered to support reliable performance over time. In the current work, we chose the Empatica E4 as the target wearable device, because it allows access to the raw data required to operate the ML classifier. The sample size was modest; participants had BMI <35 kg/m²; skin tone distribution was not analyzed; the system was evaluated on Empatica E4 rather than multiple consumer wearables; and larger independent validation is needed. Further work should also test the classifier on sensor data from consumer-grade wearables at large scale.

A readily available system for automatic and passive longitudinal measurement of objective HFs should be considered a key milestone for enhancing the management and prevention of HFs, the most common and disruptive menopausal symptom.

DATA AVAILABILITY

The data that support the findings of this study are not publicly available because they contain participant-level physiological and behavioral recordings collected under informed consent and institutional review board restrictions. Deidentified data may be made available from the corresponding author upon reasonable request and subject to applicable institutional approvals, data-use agreements, and restrictions required to protect participant privacy.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF) grant “STTR Phase I: Toward the first novel multi-sensor smart wearable system for hot flash detection” (IIP-2111818; to AG).

AUTHOR CONTRIBUTION

FCB, MdeZ, AG: conceptualization, writing-review & editing
 ATs: formal data analysis, writing-original draft
 ATa: formal data analysis, writing-review & editing
 NA: data collection, writing-review & editing

COMPETING INTEREST

The authors declare the following potential competing interest: Massimiliano de Zambotti is currently an employee of Oura Ring Inc. Massimiliano was an employee at SRI International at the time of his contribution to this study, which predates his transition to Oura. Massimiliano de Zambotti's contributions to this work, as well as the statements and opinions expressed, are solely the responsibility of the author and do not represent the official views of Oura Ring Inc. The remaining authors declare no competing interests.

ETHICS DECLARATIONS

Human ethics declarations

This study involved human participants and was conducted in accordance with the Declaration of Helsinki. The study protocol was approved by Advarra Institutional Review Board (protocol number Pro00072711). All participants provided written informed consent before participation.

VIII. REFERENCES

Baker, F. C., Forouzanfar, M., Goldstone, A., Claudatos, S. A., Javitz, H., Trinder, J., & De Zambotti, M. (2019). Changes in heart rate and blood pressure during nocturnal hot flashes associated with and without awakenings. *Sleep*, 42(11), zsz175.

Bansal, R., & Aggarwal, N. (2019). Menopausal hot flashes: A concise review. *Journal of Mid-Life Health*, 10(1), 6.

Carpenter, J. S., Andrykowski, M. A., Freedman, R. R., & Munn, R. (1999). Feasibility and psychometrics of an ambulatory hot flash monitoring device. *Menopause*, 6(3), 209–215.

Carpenter, J. S., Newton, K. M., Sternfeld, B., Joffe, H., Reed, S. D., Ensrud, K. E., & Milata, J. L. (2012). Laboratory and ambulatory evaluation of vasomotor symptom monitors from the MsFLASH Network. *Menopause (New York, NY)*, 19(6), 664.

Fine, J., Branan, K. L., Rodriguez, A. J., Boonya-Ananta, T., Ramella-Roman, J. C., McShane, M. J., & Coté, G. L. (2021). Sources of inaccuracy in photoplethysmography for continuous cardiovascular monitoring. *Biosensors*, 11(4), 126.

Forouzanfar, M., de Zambotti, M., Goldstone, A., & Baker, F. C. (2018). Automatic detection of hot flash occurrence and timing from skin conductance activity. 1090–1093.

Freedman, R. R. (2001). Physiology of hot flashes. *American Journal of Human Biology*, 13(4), 453–464.

Thurston, R. C., Matthews, K. A., Hernandez, J., & De La Torre, F. (2009). Improving the performance of physiologic hot flash measures with support vector machines. *Psychophysiology*, 46(2), 285–292.

Tsiartas, A., Baker, F. C., Smith, D., & de Zambotti, M. (2021). A novel Hot-Flash classification algorithm via multi-sensor features integration. 2067–2070.

Naghavi N, C. T. (2025). Hot Flash Prediction for the Delivery of Just-In-Time Interventions. *Psychophysiology*, 62(7).