Psychophysiology, ▮▮▮ (2010), 1–7. Wiley Periodicals, Inc. Printed in the USA.
Copyright © 2010 Society for Psychophysiological Research
DOI: 10.1111/j.1469-8986.2010.01155.x

Support Vector Machines to improve physiologic hot flash measures: Application to the ambulatory setting

REBECCA C. THURSTON,[a,b] JAVIER HERNANDEZ,[c] JOSE M. DEL RIO,[c] AND FERNANDO DE LA TORRE[c]

[a]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania
[b]Department of Epidemiology, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania
[c]Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania

Abstract

Most midlife women have hot flashes. The conventional criterion (≥ 2 μmho rise/30 s) for classifying hot flashes physiologically has shown poor performance. We improved this performance in the laboratory with Support Vector Machines (SVMs), a pattern classification method. We aimed to compare conventional to SVM methods to classify hot flashes in the ambulatory setting. Thirty-one women with hot flashes underwent 24 h of ambulatory sternal skin conductance monitoring. Hot flashes were quantified with conventional (≥ 2 μmho/30 s) and SVM methods. Conventional methods had low sensitivity (sensitivity = .57, specificity = .98, positive predictive value (PPV) = .91, negative predictive value (NPV) = .90, F1 = .60), with performance lower with higher body mass index (BMI). SVMs improved this performance (sensitivity = .87, specificity = .97, PPV = .90, NPV = .96, F1 = .88) and reduced BMI variation. SVMs can improve ambulatory physiologic hot flash measures.

Descriptors: Hot flashes, Vasomotor symptoms. Support vector machines, Physiologic measurement

Approximately 70% of midlife women experience hot flashes at some point during the menopausal transition (Gold et al., 2006). Previously thought to persist only during the several years around the menopausal transition, it is now clear that a significant minority of women experience hot flashes well into their 60s and 70s (Barnabei et al., 2002, 2005). Hot flashes are associated with pronounced decrements in quality of life, including physical, social, and emotional functioning (Avis et al., 2003, 2009). They are a consistent predictor of depressed mood during the menopausal transition (Bromberger et al., 2007). More recently, hot flashes have been linked to cardiovascular risk (Thurston, Sutton-Tyrrell, Everson-Rose, Hess, & Matthews, 2008) and bone loss (Crandall et al., 2009). Given findings of potential health risk associated with hormone therapy (Rossouw et al., 2002), the leading treatment for hot flashes, developing an improved understanding of the physiology of hot flashes and new treatments for hot flashes have been of increased scientific interest.

One factor that has limited research on hot flashes has been issues with the physiologic measurement of hot flashes. Physiologic measures of hot flashes in addition to self-report measures are desirable for research on hot flashes given the many factors that influence the reporting of hot flashes (Miller & Li, 2004). These factors include subject adherence, distraction, sleep, and psychological factors such as anxiety, which may increase the likelihood of reporting hot flashes in the absence of a physiologic change (Thurston, Blumenthal, Babyak, & Sherwood, 2005; Thurston, Matthews, Hernandez, & De La Torre, 2009). Physiologic measures address these limitations, allowing quantification of the occurrence of hot flashes without reliance on subject reporting. They have the potential to precisely quantify hot flashes, including their frequency and the exact timing of their occurrence during sleep and wake, not possible when relying solely upon subject reporting.

The most widely used physiologic measure of hot flashes is sternal skin conductance. The standard criterion for the physiologic occurrence of a hot flash is a 2 μmho rise in skin conductance in a 30-s period (Freedman, 1989). However, we previously demonstrated that, although sternal skin conductance reliably changes with hot flashes, this 2-μmho criterion had low sensitivity in the laboratory setting (Thurston et al., 2009). Further, the performance of this criterion varied by subject characteristics, showing particularly poor performance among overweight/obese women. Others have similarly shown issues with the performance of this measure (de Bakker & Everaerd, 1996; Hanisch, Palmer, Donahue, & Coyne, 2007; Sievert et al., 2002). We improved upon this criterion by using the classification approach of Support Vector Machines (SVMs) (Thurston et al., 2009). SVMs are state-of-the-art classification methods

This publication was supported by The Fannie E. Rippel Foundation/American Federation for Aging Research New Investigator Award on Gender Differences in Aging (PI: Thurston), the Caja Madrid Foundation (Hernandez), the National Institutes of Health (K23AG029216), and the Pittsburgh Mind-Body Center (National Institute of Heath grants HL076852/076858).

Address correspondence to: Rebecca C. Thurston, PhD, 3811 O'Hara St., Pittsburgh, PA, 15213. E-mail: thurstonrc@upmc.edu

particularly useful for complicated pattern recognition problems (Guyon, Weston, Barnhill, & Vapnik, 2002; Joachims, 1998; Michel & Kaliouby, 2003). As opposed to applying a single magnitude-based threshold to classify hot flashes, SVMs can characterize the distinct skin conductance patterns associated with hot flashes.

Our prior work considered laboratory-measured hot flashes only. We now extend our prior work to the ambulatory setting, the setting in which much of clinical research on hot flashes occurs. However, measuring hot flashes in the ambulatory setting is more challenging than in the laboratory. The ambulatory setting is characterized by a range of factors increasing error in skin conductance signals, such as baseline drift; artifact producing factors such as physical activity, seat belts, and purses; and a changing electrode-subject interface over longer monitoring periods. Further, hot flash reporting may become less reliable in the ambulatory setting, with increased distractions and greater subject adherence demands for reporting over 24 h or more.

In this investigation, we evaluate the conventional criterion (2 μmho rise in 30-s period) for quantifying hot flashes from sternal skin conductance in the ambulatory setting. We also apply SVMs to sternal skin conductance signals to improve the performance of this measure. For both conventionally quantified and SVM-quantified hot flashes, we evaluate any variation in the algorithms by subject characteristics such as body mass index (BMI), race/ethnicity, and anxiety, factors that have been previously linked to variations in the physiologic detection and/or reporting of hot flashes (Sievert, 2007; Thurston et al., 2009).

Methods

Subjects

Thirty-four African American and Caucasian women between the ages of 40 and 60 were recruited from the surrounding community via newspaper advertisements, fliers in local businesses, and online message board postings. Inclusion criteria included late perimenopausal (amenorrhea last 3–12 months) or postmenopausal (amenorrhea $\geq$ 12 months) status, reporting $\geq$ 4 hot flashes a day, and having a uterus and both ovaries. Women were excluded if having taken hormone therapy (oral or transdermal estrogen and/or progesterone), oral contraceptives, selective serotonin reuptake inhibitors or serotonin norepinepherine reuptake inhibitors, clonidine, methyldopa, bellergal, gabapentin, aromatase inhibitors, selective estrogen receptor modulators in the past 3 months, having taken isoflavone supplements or black cohosh in the past month, currently undergoing acupuncture for the treatment of hot flashes, having reported medical or psychiatric conditions associated with hot flash sensations (panic disorder, pheochromocytoma, leukemia, pancreatic tumor), or having inability to provide informed consent and follow study procedures. Of these 34 women, two women were excluded due to equipment failure during the session, and one woman withdrew from the study for a final sample of 31 women.

Procedures

Participants underwent measurement of height, weight, and waist circumference, and completed a battery of questionnaires for assessment of medical, demographic, and psychological characteristics. Participants were equipped with an ambulatory sternal skin conductance monitor and an electronic diary, and instructed to wear the monitor over 24 h. They were instructed to avoid heavy physical activity, swimming, and showering while wearing the monitor. At each experience of a hot flash, participants were instructed to press two event mark buttons on the monitor and complete the electronic diary entry, both of which provide date- and time-stamped subjective hot flash reports. For five subjects, problems with electrode adhesion compromised data quality; these subjects conducted a make-up monitoring session with all usable monitoring data reported here. All study procedures were approved by the University of Pittsburgh Institutional Review Board, and all participants provided written informed consent.

Measures

Sternal skin conductance was recorded via the Biolog ambulatory monitor (model 3991/2-SCL; UFI, Morro Bay, CA), a portable device worn in a pouch around the waist. The Biolog measures sternal skin conductance during daily life, sampled at 1 Hz from the sternum via a 0.5-volt constant voltage circuit passed between two Ag/AgCl electrodes (Vermed, Bellows Falls, VT) filled with 0.05M KCL Velvachol/glycol paste (Dormire & Carpenter, 2002). The Biolog has two event mark buttons providing a date/time-stamped subjective event report. During waking monitoring hours, participants completed a portable electronic diary (Palm Zire, Palm, Inc., Sunnyvale, CA) when experiencing a hot flash, with questions about the experience of the hot flash programmed with Palm-compatible software (Satellite Forms; Thacker Network Technologies, Lacombe, AB, Canada). Height and weight were measured via a fixed stadiometer and a calibrated balance beam scale, respectively. Menstrual history, parity, education, marital status, alcohol use, and smoking status were assessed by standard demographic and medical history questionnaires. Depressive symptoms were assessed via the Center for Epidemiologic Studies Depression Survey (Radloff, 1977) and state and trait anxiety via the State-Trait Anxiety Inventory (Spielberger, 1983).

Data Reduction

Physiologic hot flashes were classified with two methods: (1) 2 μmho rise in 30 s (conventional criterion), and (2) SVM-defined hot flashes.

Conventional criterion. For the conventional criterion, skin conductance increases of 2 μmho in 30 s (Freedman, 1989) were flagged automatically by custom software and edited for artifact using standard methods (Carpenter, Andrykowski, Freedman, & Munn, 1999). A 20-m lockout period was implemented after the start of the flash, after which no hot flashes were coded.

Support Vector Machines. Building an SVM model, or algorithm, consists of both a training phase, in which the model is developed, and a testing phase, in which model performance is tested. Data were first preprocessed to normalize skin conductance signals and to remove noise. Since we have previously found dramatic differences in the magnitude of skin conductance rises associated with hot flashes (Thurston et al., 2009), signals were scaled between zero and one within subjects. Momentary voltage drops and baseline drift were removed, and exponential smoothing ($\alpha = 0.08$) (Holt, 2004) was applied to reduce signal noise.

Next, a dataset is created in which all hot flash events are labeled. Given the presence of reported hot flashes accompanied by no skin conductance changes and the known psychological influences on hot flash reporting limiting the validity of reports (Thurston, Blumenthal, Babyak, & Sherwood, 2005), and consistent with our prior work (Thurston et al., 2009), all data were labeled via expert-defined physiologic hot flashes. Skin conductance changes associated with hot flashes show a sharp and rapid rise following by a sloping return to baseline, or "swishy tail," that can be distinguished from the "sawtooth" pattern characteristic of activity or other sweating-related artifact (Carpenter et al., 1999). Thus, all data were visually reviewed and hot flash and non-hot flash intervals labeled, based upon the characteristic skin conductance shape of the hot flash rather than solely on the magnitude of the rise. The reliability of this manual labeling from two separate coders was $\kappa = .86$. All available skin conductance training data is then divided into smaller segments belonging to one of two classes (hot flash or non-hot flash).

The SVM model is then trained, learning the characteristics associated with hot flash and non-hot flash segments. Key features of the hot flash-associated skin conductance rises versus the non-hot flash-associated skin conductance are extracted, and a non linear decision boundary is used to separate hot flash and non-hot flash events (see online Appendix). Once the SVM model has been trained, the performance of this model is evaluated on a new set of testing sequences. The same preprocessing steps, data segmentation, and feature extraction techniques are used for the testing set. We implemented a leave-one-subject-out strategy, useful when limited quantities of data are available for training and testing (Witten & Frank, 2005). The leave-one-subject-out cross-validation approach uses a single subject's session from the original sample as testing, and the remaining sessions (other subjects) as the training data from which the SVM is learned. This is repeated with a retraining of the SVM in each round, such that each session is used only once as testing, and the remainder as training data. The SVM model classifies each new segment of skin conductance data in this testing set as a hot flash or non-hot flash, and the performance of the algorithm is calculated.

LIBSVM, a publicly available MATLAB library, was used to implement the core of SVM training and testing (C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines, 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm/). Although specialized expertise is required to build an SVM, this publicly available library provides SVM tutorials, sample programming code, and program code libraries that form the core of training and testing of the SVM algorithm. See the on-line Appendix for more technical details of preprocessing, segmentation, features extracting, training, and testing of the SVM.

Calculation of performance indices. True positive (TP), false positive (FP), false negative (FN), and true negative (TN) hot flashes were scored relative to self-reported hot flashes and expert-defined hot flashes. Expert label was the primary referent, given the known psychological and related influences on hot flash reporting (Thurston et al., 2005, 2009), although the referent of self-report was also calculated. For models using self-reported hot flashes as the referent, only waking hours of data were used, given the low reliability of hot flash reporting during sleep (Thurston et al., 2006). Given the tendency for participants to report hot flashes multiple times via multiple methods, hot flash reports within 15 min were grouped as one report. Physiologic hot flashes were defined both by the conventional criterion as well as SVM. A physiologic hot flash was considered concordant with a report/expert label if it was met by a report/label within 5 min before and 20 min after the physiologic hot flash. The 5-min pre-flash interval is consistent with the published literature (Carpenter et al., 1999); the 20-min post-flash interval was selected for consistency with the traditional inter-flash lockout period and to account for the duration over which skin conductance changes associated with a single hot flash occur. A TP was a physiologic hot flash prediction met by an expert-labeled/self-reported hot flash. A FP was a physiologic hot flash prediction not met by an expert-labeled/self-reported hot flash. A FN was an expert label/self-report not followed by a physiologic hot flash prediction. A TN was all 20-min intervals lacking both an expert label/self-report and a physiologic hot flash. For each physiologic hot flash detection method (conventional, SVM), sensitivity (TP/TP+FN), specificity (TN/FP+TN), positive predictive value (PPV; TP/TP+FP), negative predictive value (NPV; TN/TN+FN), and F1 (2 TN/(2 TN+FP+FN)) were calculated. Sensitivity corresponds to the percentage of labeled/reported hot flashes accompanied by a hot flash prediction, specificity to the percentage of 20-min segments without a labeled/reported hot flash also lacking a hot flash prediction, PPV to the percentage of hot flash predictions also accompanied by a labeled/reported hot flash, NPV to the percentage of 20-min segments without a hot flash prediction that also lack a labeled/reported hot flash. F1 provides an overall index of performance, a summary index unbiased by the large number of true negatives in the sample (Jardine & van Rijsbergen, 2002).

Data Analysis

Comparisons of performance indices by subject characteristics were performed using Spearman's rho, Pearson correlation coefficients, and linear and logistic regression, with transformation of rates as necessary. Comparisons between SVM and conventional criterion performance were conducted using paired *t*-tests. Due to skew, transformations were applied to meet model assumptions of normality: all SVM performance indices were exponentiated for models with the referent of expert label, specificity was exponentiated for SVM models using the referent of self-report, specificity was exponentiated for conventional criterion for both referents, and NPV and PPV exponentiated for the conventional criterion with the referent of expert label. Statistical comparisons for PPV were limited by a zero denominator for multiple PPV values; these observations were set to missing, and sample sizes noted when deviating from the full sample. Variations in performance by the subject characteristics of race/ethnicity, BMI, and state and trait anxiety were evaluated in linear regression models, with each performance index considered separately. Since state and trait anxiety were highly correlated (0.67, $p < .0001$), and findings for trait anxiety were stronger than those for state anxiety, only trait anxiety findings are reported here. Analyses were performed using SAS v.9.2 (SAS Institute, Cary, NC) and MATLAB v.7.0 (MathWorks, Natick, MA). Tests were 2-sided with $\alpha = .05$.

Results

Participants were on average 53 years old, postmenopausal, and overweight (Table 1). Women reported on average 10 hot flashes during the waking hours of the ambulatory monitoring. An average of 24.62 ($SD = 2.86$) monitoring hours were conducted per woman.

Table 1. *Subject Characteristics*

N	31
Age, *M* (*SD*)	53.0 (4.8)
Race, *n* (%)	
White	17 (54.8)
Black	14 (45.2)
Education, *n* (%)	
High school or less	9 (29.0)
Vocational/some college	13 (41.9)
College or higher	9 (29.0)
Menopausal status, *n* (%)	
Postmenopausal	19 (61.29)
Late perimenopausal	12 (38.71)
BMI, *M* (*SD*)	28.7 (6.0)
Marital Status, *n* (%)	
Never married	7 (22.6)
Married	17 (54.8)
Divorced/widowed	7 (22.6)
Current smoker, *n* (%)	8 (25.81)
Parous, *n* (%)	24 (77.4)
State anxiety, *M* (*SD*)	29.4 (7.2)
Trait anxiety, *M* (*SD*)	32.5 (6.8)
CESD, *M* (*SD*)	5.0 (3.6)
Reported hot flashes (waking), Median (IQR)†	10 (7)
Conventionally defined hot flashes, Median (IQR)‡	10 (18)
SVM-defined hot flashes, Median (IQR)‡	14 (10)

†Waking hours only.
‡Waking and sleeping hours.

Conventional Criterion

The conventional 2-μmho criterion showed low sensitivity relative to both expert-labeled and subject-reported hot flashes (Table 2). Further, the performance of the conventional criterion significantly varied by BMI. Consistent with prior findings in the laboratory (Thurston et al., 2009), when compared to expert labeled hot flashes, the conventional criterion showed lower sensitivity ($b(SE) = -2.44$ (1.13), $p = .04$) and a lower F1 score ($b(SE) = -2.45$ (1.09), $p = .03$) as BMI, considered as a continuous variable, increased (Table 3). Using the referent of subject reports, with higher BMI, sensitivity of the conventional criterion was lower ($b(SE) = -2.80$ (1.14), $p = .02$), specificity$_{\text{exp}}$ higher ($b(SE) = 511.59$ (210.20), $p = .02$), and NPV$_{\text{exp}}$ lower ($b(SE) = -510.31$ (192.44), $p = .01$). Thus, the conventional criterion differentially missed both the expert-labeled and subject-reported hot flashes among heavier women. Comparisons by obesity status category are presented in Table 3. There were no differences in the performance of the conventional criterion for either referent by anxiety or race/ethnicity.

Similar to laboratory findings (Thurston et al., 2009), the magnitude of the 30-s skin conductance rises associated with hot flashes significantly and inversely varied by BMI ($r = -.36$, $p = .045$). Particularly low rises were observed among the obese women ($M(SD) = 1.62$ (1.28)) in contrast to normal weight ($M(SD) = 2.93$ (1.05)) and overweight women ($M(SD) = 2.68$ (1.41); $F(2,28) = 3.73$, $p = .04$). Thus, the attenuated skin conductance rises associated with obesity were likely misclassified using the conventional single threshold criterion.

Table 2. *Performance of Sternal Skin Conductance 2 μmho/30 s Criterion to Classify Hot Flashes*

	Referent: expert-defined hot flashes	Referent: subject-reported hot flashes
Sensitivity	.57	.52
Specificity	.98	.92
PPV	.91	.59
NPV	.90	.92
F1	.60	.44

Note: $N = 31$ for all models except PPV for expert-defined hot flashes referent ($N = 24$), and PPV for subject-reported hot flash referent ($N = 23$).

Table 3. *Performance of Conventional Criterion (2 μmho/30 s) to Classify Hot Flashes from Sternal Skin Conductance by Obesity Status*

	Referent: expert-defined hot flashes			Referent: subject-reported hot flashes		
	Normal ($N = 10$)	Overweight ($N = 7$)	Obese ($N = 14$)	Normal ($N = 10$)	Overweight ($N = 7$)	Obese ($N = 14$)
Sensitivity	.75	.63	.41	.69	.68	.32*
Specificity	.98	.99	.99	.86	.91	.96*
PPV	.94	.93	.86	.51	.64	.63
NPV	.93	.90	.88	.96	.93	.88*
F1	.77	.68	.44	.53	.60	.30

Note: Normal weight BMI < 25 (referent), Overweight BMI 25–29.9, Obese BMI $\geq$ 30; Low state anxiety $\leq$ 27.5 (referent), high state anxiety > 27.5; *N*s are as noted except PPV for expert-defined hot flashes referent (normal weight $N = 9$, overweight $N = 6$, obese $N = 9$) and PPV for subject-reported hot flashes referent (normal weight $N = 9$, overweight $N = 6$, obese $N = 8$).
*$p < .05$, obesity status comparison.

SVM

Application of SVMs improved the performance of sternal skin conductance-assessed hot flashes when compared to both expert-labeled hot flashes and self-reported hot flashes (Table 4). These improvements were statistically significant when compared to the conventional criterion in the case of sensitivity ($t(30) = 4.47$, $p < .0001$), NPV ($t(30) = 3.92$, $p = .0005$) and F1 score ($t(30) = 4.40$, $p = .0001$) for the referent of expert labeled hot flashes, and the F1 score ($t(30) = 3.12$, $p = .004$) and to a lesser extent sensitivity ($t(30) = 1.88$, $p = .07$) for the referent of subject-reported hot flashes.

No differences were observed in the performance of SVM models by BMI for the expert label referent. The one condition in which SVM performance varied by BMI was for NPV when using the referent of subject report ($b(SE) = -0.34$ (0.16), $p = .04$). Thus, more of the 20-min intervals containing a hot flash report among higher BMI women were not accompanied by an SVM hot flash. Comparisons by obesity status category are presented in Table 5.

Table 4. *Performance of Sternal Skin Conductance Using SVM Methods to Classify Hot Flashes*

	Referent: expert-defined hot flashes	Referent: subject-reported hot flashes
Sensitivity	.87	.61
Specificity	.97	.91
PPV	.90	.61
NPV	.96	.92
F1	.88	.56

Note: $N = 31$ for all models except PPV for subject-reported hot flash as referent ($N = 28$).

Table 5. *Performance of SVM to Classify Hot Flashes from Sternal Skin Conductance by Obesity Status*

	Referent: expert-defined hot flashes			Referent: subject-reported hot flashes		
	Normal ($N = 10$)	Overweight ($N = 7$)	Obese ($N = 14$)	Normal ($N = 10$)	Overweight ($N = 7$)	Obese ($N = 14$)
Sensitivity	.92	.89	.84	.71	.72	.48
Specificity	.98	.95	.98	.88	.90	.94
PPV	.95	.88	.88	.55	.66	.63
NPV	.97	.97	.96	.95	.92	.90
F1	.93	.88	.85	.60	.62	.50

Note: Normal weight BMI < 25 (referent), Overweight BMI 25–29.9, Obese BMI $\geq$ 30; *N*s are as noted except PPV for subject-reported hot flashes as referent (normal weight $N = 10$, overweight $N = 6$, obese $N = 12$).

Similar to prior work (Thurston et al., 2005, 2009), for the referent of self-reported hot flashes only, women higher in trait anxiety had a somewhat lower F1 score for SVM models ($b(SE) = -1.36(0.68)$, $p = .05$). There were no differences in the performance of SVM models by anxiety for the expert label referent, nor by race/ethnicity for either referent.

Additional Analyses

We compared the two referents, self-report and expert label, to understand how differently these two referents characterized the data. Relative to expert labels, self-report had a sensitivity of .57, a specificity of .96, a PPV of .80, a NPV of .87, and an F1 score of .63. This pattern can be interpreted as a tendency to under-report hot flashes, similar to prior work (Carpenter, Monahan, & Azzouz, 2004). Further, given our prior work showing associations between anxiety and the reporting of hot flashes (Thurston et al., 2005, 2009), we evaluated the association between anxiety and reporting of hot flashes by comparing reported hot flashes to our referent of expert labeled hot flashes. These models indicated that higher trait anxiety was associated with lower $\text{specificity}_{\text{exp}}$ ($b(SE) = 445.83$ (130.00), $p = .003$), lower PPV ($b(SE) = -1.76\,(0.72)$, $p = .02$), and a lower F1 score ($b(SE) = -1.49$ (0.63), $p = .03$) of subject reporting of hot flashes when compared to expert labeled hot flashes. Further, higher trait anxiety was associated with the tendency to report hot flashes accompanied by no discernable change in skin conductance (OR (95% CI) = 1.22 (1.03–1.45), $p = .02$). Thus, when subject reports were compared to our primary referent of expert labels, trait anxiety was associated with elevated hot flash reporting, even in the absence of any skin conductance change.

Discussion

The present investigation showed that the conventional criterion had low sensitivity. Moreover, the performance of this conventional criterion varied by BMI, showing particularly low sensitivity among women with higher BMI. Applying SVMs improved the performance of sternal skin conductance in detecting hot flashes. Further, variations by obesity status were reduced with use of SVMs.

This investigation extends our prior work in the laboratory (Thurston et al., 2009) to the ambulatory setting. While the ambulatory setting is the primary setting in which clinical research on hot flashes is carried out, particularly clinical trials, the ambulatory setting is a more difficult setting for physiologic hot flash measurement, characterized by greater baseline shifts, variations in signal quality, and artifact from factors such as activity, which can produce skin conductance elevations that meet the conventional threshold. SVMs are a class of machine learning models that are particularly well suited to these types of complex pattern recognition problems (Guyon et al., 2002; Joachims, 1998; Michel & Kaliouby, 2003). They present several advantages over the conventional criterion. In contrast to the conventional criterion's application of a single threshold, SVMs characterize the shape of the skin conductance changes associated with the hot flashes versus artifact. This approach not only more carefully quantifies the hot flash, but can reduce systematic variation by subject characteristics such as BMI, which may reduce the magnitude of hot flash-associated skin conductance rises. Further attenuating systematic variation by subject characteristics is SVMs' ability to be developed to vary by key subject characteristics, or ultimately, by woman. Multiple sources of information, such as subject characteristics or dynamic physiologic or subjective information, can also be used in training SVM models (Bundele & Banerjee, 2009; Sommer, Golz, Trutschel, & Edwards, 2008). Finally, SVMs will ultimately classify hot flashes in a largely automated fashion, reducing the extensive visual editing characteristic of the conventional criterion approach. A valid, automated coding approach for physiologic hot flash measures is essential to the use of these measures with larger samples and by a wider range of researchers (Miller & Li, 2004).

Similar to our prior findings in the laboratory (Thurston et al., 2009) variations in the performance of the conventional criterion were observed by BMI. Lower performance of the conventional criterion, particularly low sensitivity, was observed with higher BMI, with strikingly poor performance among obese women. Skin conductance recordings among higher BMI women are often characterized by greater artifact (ambient sweating). However, similar to our findings in the laboratory, the magnitude of hot flash-associated skin conductance rises were also attenuated among higher BMI women. The conventional threshold's single magnitude-based threshold frequently misclassified these smaller magnitude hot flashes among higher BMI women as nonevents. Thus, with use of the conventional criterion, bias by BMI is introduced. This phenomenon is particularly worrisome as obesity is a leading risk factor for hot flashes (Gold et al., 2006; Thurston, Sowers, Chang, Chang, Gold, Johnston, & Matthews, 2008), and the majority of women in the U.S. are overweight or obese (Flegal, Carroll, Ogden, & Curtin, 2010). These BMI differences in performance were reduced with application of SVMs. Notably, SVMs characterize the pattern, as opposed to only the amplitude of the rise, of sternal skin conductance changes associated with hot flashes.

In our previous work, we showed that elevated anxiety assessed at the time of the laboratory testing was associated with poorer performance of conventional and SVM models when compared to self-report, largely due to elevated reporting of hot flashes lacking any physiologic change among more anxious women (Thurston et al., 2009). In this ambulatory investigation, anxiety showed less pronounced and consistent associations with hot flash reporting than in the laboratory setting. However, trait anxiety was associated with elevated hot flash reporting when self-reports were compared to expert labels, the leading referent. Trait anxiety was also associated with the tendency to report hot flashes in the absence of any discernible skin conductance changes. Finally, higher trait anxious women had somewhat lower F1 scores for SVM models when compared to self-report. Notably, anxiety has been shown to be a leading predictor of hot flash reporting in epidemiologic investigations (Freeman et al., 2005; Gold et al., 2006) as well as elevated ambulatory hot flash reporting when compared to physiologically assessed hot flashes (Thurston et al., 2005). These findings are consistent with a large body of literature linking anxiety to elevated symptom reporting (Cohen et al., 1995; Pennebaker, 1982).

Two referents were considered here: expert labeled hot flashes and self-reported hot flashes. While self-reported hot flashes are traditionally considered as a referent, the many reporting biases make this a less ideal referent, including subject non-adherence, inability to reliably report hot flashes during sleep, as well as the psychological factors discussed above. Reporting may be particularly unreliable in the ambulatory setting, characterized by multiple distractions, attentional demands, and settings in which hot flash reporting may be impossible (e.g., driving). Thus, key performance indices may be somewhat lower with the referent of self-report in the ambulatory versus laboratory setting (Thurston et al., 2009). No other gold standard measures of hot flashes exist. Coders can be trained to visually distinguish hot flashes from skin conductance signals with high reliability as demonstrated here. Therefore, consistent with our prior work (Thurston et al., 2009), we present findings considering both self-report and expert-labeled hot flashes as a referent. However, due to its demand for rigorous training and extensive personnel coding time, this approach is not feasible for larger investigations or for longer monitoring durations, underscoring the need for development of a more automated approach.

Several limitations deserve mention. First, the study sample was relatively small. While the SVM models were stable as the number of hot flashes across women was high, certain secondary comparisons by subject characteristics involved fairly small numbers, limiting power. In calculating PPV, particularly for the conventional criterion, zero denominator problems (no physiologic hot flashes) further limited sample size PPV. Further, while SVM libraries and documentation are publicly available, building SVM models from raw data requires specialized experience. Finally, this SVM model was a preliminary model for the ambulatory setting requiring further elaboration, refinement, and further testing in independent samples. With further testing, a final model can be made available to the wider scientific community.

This study has several strengths. This study was a detailed analysis of subjectively reported and physiologically recorded hot flashes over 24 h, allowing a careful comparison of these indices. It included two racial/ethnic groups, allowing validation of these ambulatory measurement approaches across these groups. The assessment of multiple physiologic and psychological characteristics, allowed comparison of the performance of these measurement approaches by these subject characteristics. Most importantly, this study is the first to apply SVMs to quantifying hot flashes physiologically in the ambulatory setting. This approach improved the detection of hot flashes, further advancing the physiologic measurement of hot flashes.

The present research showed that the standard single threshold criterion to quantify hot flashes physiologically had poor performance, particularly among higher BMI women, which may have attenuated skin conductance rises with hot flashes. This research indicates the importance of using more sophisticated pattern recognition models, SVMs, to characterize hot flashes from physiologic signals. These models improved the overall performance of this index and reduced variation by BMI. This improved method to quantify hot flashes has the potential to further advance research on hot flashes, supporting research aimed at better understanding the physiology of hot flashes as well as the development of new treatments for hot flashes.

REFERENCES

Avis, N. E., Colvin, A., Bromberger, J. T., Hess, R., Matthews, K. A., Ory, M., & Schocken, M. (2009). Change in health-related quality of life over the menopausal transition in a multiethnic cohort of middle-aged women: Study of Women's Health Across the Nation. *Menopause*, *16*, 860–869.

Avis, N. E., Ory, M., Matthews, K. A., Schocken, M., Bromberger, J., & Colvin, A. (2003). Health-related quality of life in a multiethnic sample of middle-aged women: Study of Women's Health Across the Nation (SWAN). *Medical Care*, *41*, 1262–1276.

Barnabei, V. M., Cochrane, B. B., Aragaki, A. K., Nygaard, I., Williams, R. S., McGovern, P. G., et al. (2005). Menopausal symptoms and treatment-related effects of estrogen and progestin in the Women's Health Initiative. *Obstetrics and Gynecology*, *105*, 1063–1073.

Barnabei, V. M., Grady, D., Stovall, D. W., Cauley, J. A., Lin, F., Stuenkel, C. A., et al. (2002). Menopausal symptoms in older women and the effects of treatment with hormone therapy. *Obstetrics and Gynecology*, *100*, 1209–1218.

Bromberger, J. T., Matthews, K. A., Schott, L. L., Brockwell, S., Avis, N. E., Kravitz, H. M., et al. (2007). Depressive symptoms during the menopausal transition: The Study of Women's Health Across the Nation (SWAN). *Journal of Affective Disorders*, *103*, 267–272.

Bundele, M., & Banerjee, R. (2009). An SVM classifier for fatigue-detection using skin conductance for use in the BITS-Lifeguard Wearable Computing System. *International Conference on Emerging Trends in Engineering & Technology* (pp. 934–939). Nagpur, Maharashtra, India.

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

Carpenter, J. S., Monahan, P. O., & Azzouz, F. (2004). Accuracy of subjective hot flush reports compared with continuous sternal skin conductance monitoring. *Obstetrics and Gynecology*, *104*, 1322–1326.

Cohen, S., Doyle, W. J., Skoner, D. P., Fireman, P., Gwaltney, J. M. Jr., & Newsom, J. T. (1995). State and trait negative affect as predictors of objective and subjective symptoms of respiratory viral infections. *Journal of Perspectives in Social Psychology*, *68*, 159–169.

Crandall, C. J., Zheng, Y., Crawford, S. L., Thurston, R. C., Gold, E. B., Johnston, J. M., & Greendale, G. A. (2009). Presence of vasomotor symptoms is associated with lower bone mineral density: A longitudinal analysis. *Menopause*, *16*, 239–246.

de Bakker, I. P., & Everaerd, W. (1996). Measurement of menopausal hot flushes: Validation and cross-validation. *Maturitas*, *25*, 87–98.

Dormire, S. L., & Carpenter, J. S. (2002). An alternative to Unibase/glycol as an effective nonhydrating electrolyte medium for the measurement of electrodermal activity. *Psychophysiology*, *39*, 423–426.

Flegal, K. M., Carroll, M. D., Ogden, C. L., & Curtin, L. R. (2010). Prevalence and trends in obesity among US adults, 1999–2008. *Journal of the American Medical Association*, *303*, 235–241.

Freedman, R. R. (1989). Laboratory and ambulatory monitoring of menopausal hot flashes. *Psychophysiology*, *26*, 573–579.

Freeman, E. W., Sammel, M. D., Lin, H., Gracia, C. R., Kapoor, S., & Ferdousi, T. (2005). The role of anxiety and hormonal changes in menopausal hot flashes. *Menopause*, *12*, 258–266.

Gold, E., Colvin, A., Avis, N., Bromberger, J., Greendale, G., Powell, L., et al. (2006). Longitudinal analysis of vasomotor symptoms and race/ethnicity across the menopausal transition: Study of Women's Health Across the Nation (SWAN). *American Journal of Public Health*, *96*, 1226–1235.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*, 389–422.

Hanisch, L. J., Palmer, S. C., Donahue, A., & Coyne, J. C. (2007). Validation of sternal skin conductance for detection of hot flashes in prostate cancer survivors. *Psychophysiology*, *44*, 189–193.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, *20*, 5–10.

Jardine, N., & van Rijsbergen, C. (2002). The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, *7*, 217–240.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the European Conference on Machine Learning*. Berlin: Springer.

Michel, P., & Kaliouby, R. E. (2003). Real time facial expression recognition in video using support vector machines. *Proceedings of the 5th International Conference on Multimodal Interfaces* (pp. 258–264). New York: Association for Computing Machinery.

Miller, H. G., & Li, R. M. (2004). Measuring hot flashes: Summary of a National Institutes of Health workshop. *Mayo Clinical Procedures*, *79*, 777–781.

Pennebaker, J. (1982). *The psychology of physical symptoms*. New York: Springer-Verlag.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385–401.

Rossouw, J. E., Anderson, G. L., Prentice, R. L., LaCroix, A. Z., Kooperberg, C., Stefanick, M. L., et al. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the Women's Health Initiative randomized controlled trial. *Journal of the American Medical Association*, *288*, 321–333.

Sievert, L. L. (2007). Variation in sweating patterns: Implications for studies of hot flashes through skin conductance. *Menopause*, *14*, 742–751.

Sievert, L. L., Freedman, R. R., Garcia, J. Z., Foster, J. W., del Carmen Romano Soriano, M., Longcope, C., & Franz, C. (2002). Measurement of hot flashes by sternal skin conductance and subjective hot flash report in Puebla, Mexico. *Menopause*, *9*, 367–376.

Sommer, D., Golz, M., Trutschel, U., & Edwards, D. (2008). Assessing driver's hypovigilance from biosignals. *4th European Conference of the International Federation for Medical and Biological Engineering*, *22*, 152–155.

Spielberger, C. D. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto: Consulting Psychologists Press.

Thurston, R. C., Blumenthal, J. A., Babyak, M. A., & Sherwood, A. (2005). Emotional antecedents of hot flashes during daily life. *Psychosometric Medicine*, *67*, 137–146.

Thurston, R. C., Blumenthal, J. A., Babyak, M. A., & Sherwood, A. (2006). The association between hot flashes, sleep complaints, and psychological functioning among healthy menopausal women. *International Journal of Behavioural Medicine*, *13*, 163–172.

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*, 285–292.

Thurston, R. C., Sowers, M. R., Chang, Y., Chang, B., Gold, E. B., Johnston, J. M., & Matthews, K. A. (2008). Adiposity and reporting of vasomotor symptoms among midlife women: The Study of Women's Health Across the Nation. *American Journal of Epidemiology*, *167*, 78–85.

Thurston, R. C., Sutton-Tyrrell, K., Everson-Rose, S. A., Hess, R., & Matthews, K. A. (2008). Hot flashes and subclinical cardiovascular disease: Findings from the Study of Women's Health Across the Nation Heart Study. *Circulation*, *118*, 1234–1240.

Witten, I., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann Publishers.

Supporting Information

Additional supporting information may be found in the online version of this article:

Appendix S1: This section provides further details about procedures implemented for hot flash detection using SVMs.

Figure S1: Illustration of the features extracted from each segment of skin conductance to characterize hot flashes.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

(Received May 13, 2010; Accepted September 29, 2010)

Appendix

This section provides further details about procedures implemented for hot flash detection using SVMs.

Preprocessing. Signal processing techniques were used to normalize skin conductance signals across subjects and to remove noise. Since we have previously found dramatic differences in the magnitude of skin conductance rises associated with hot flashes (Thurston, Matthews, Hernandez, & De La Torre, 2009), signals were scaled between zero and one within subject to attenuate between-subject variability. In addition, momentary voltage drops were removed, and baseline drift was removed with a sliding window based approach. Finally, exponential smoothing (α=0.08) was applied to reduce signal noise.

Segmentation. A segmentation algorithm was applied to automatically select segments of skin conductance potentially including a hot flash. A segment was defined as an interval that contained the highest intensity (apex) over a 10-minute period, and its beginning (onset) and end (offset) intensities were lower than 5% of the maximum subject's intensity. While several such onset and offset boundaries are possible, we choose those that are closest to the apex position, and at most 20-minutes and 30- minutes from the apex for the onset and offset, respectively. Short segments (< 4 minutes) and segments with low apex values (< 25% of the average subject's apex values) were automatically flagged as non-hot flashes and discarded. All other segments were associated to one of the two classes based on the proximity to a labeled hot flash.

Feature extraction. The following features were extracted from each segment to characterize the shape and intensity of physiologic changes associated with hot flashes: (1) apex value, (2) relative apex position to the segment duration, (3) number of high intensity peaks that are separated by at least one thirtieth (1/30) of the segment duration, (4) 2 µmho based features,

(5) intensities of 10 equally separated points normalized by the onset intensity, and (6) sorted intensities of 10 equally separated points. The 2 µmho-based features were characterized by: (a) maximum intensity increase in 30 seconds, and (b) binary predicates answering the question (1-yes, 0-no): “Is the maximum intensity increase in 30 seconds larger than a certain set of thresholds?” (where the set of thresholds ranged from the minimum intensity to the maximum intensity in increments of 0.01). Figure 1 illustrates these features for a hot-flash segment. Features were extracted for all of the segments and standardized to have a mean of 0 and a standard deviation of 1.

Training phase. We used a Radial Basis Function (RBF) kernel to allow non-linear decision boundaries between classes. The weights of the SVM were set as the class priors (i.e. percentage of training samples for each class) to give the same relevance to both classes, independently of which class contains more samples in the training set. The penalization factor (c) and the kernel width (g) were learned by choosing the best pair of c and g that minimize the empirical error. That is, for each combination of the parameters ($c \in \log_2([-5:2:9])$) and ($g \in \log_2([-15:1:3])$), we chose the pair that minimized the empirical error. In our case, we used the following estimation of the detection error:

$$\text{Error} = K_+ * \text{FP} + K_- * \text{FN}$$

where FP (false positives) and FN (false negatives) are the misclassifications of our model; K_+ is the class prior of the positive class; and K_- is the class prior of the negative class. Critical to the success of SVM is the selection of the parameters that avoid overfitting to the training set. When overfitting occurs, the learned model fits the training set very well but shows poor generalization to unobserved data. To address this problem, the detection error was computed using 5-subject cross validation. The 5-subject cross validation technique divides the training data in N (in our experiments N = 6) groups of 5 subjects, trains N SVM models excluding one of the groups when learning each of the N models, and it computes the empirical error for each excluded group. The final detection error is then estimated by averaging the N computed errors, and the pair of parameters c and g that have less average error are selected

Testing phase. During the testing phase, the same pre-processing steps, data segmentation and feature extraction techniques were used to generate the testing set. We implemented a leave-one-subject-out strategy, useful when limited quantities of data are available for training and testing (Witten & Frank, 2005). The leave-one-subject-out cross-validation approach uses a single subject's session from the original sample as validation data to be tested, and the remaining sessions as the training data from which the SVM is trained. This is repeated with a re-training of the SVM in each round, such that each session is used only once as validation data, and the remainder as training data.

References

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**, 285-292.

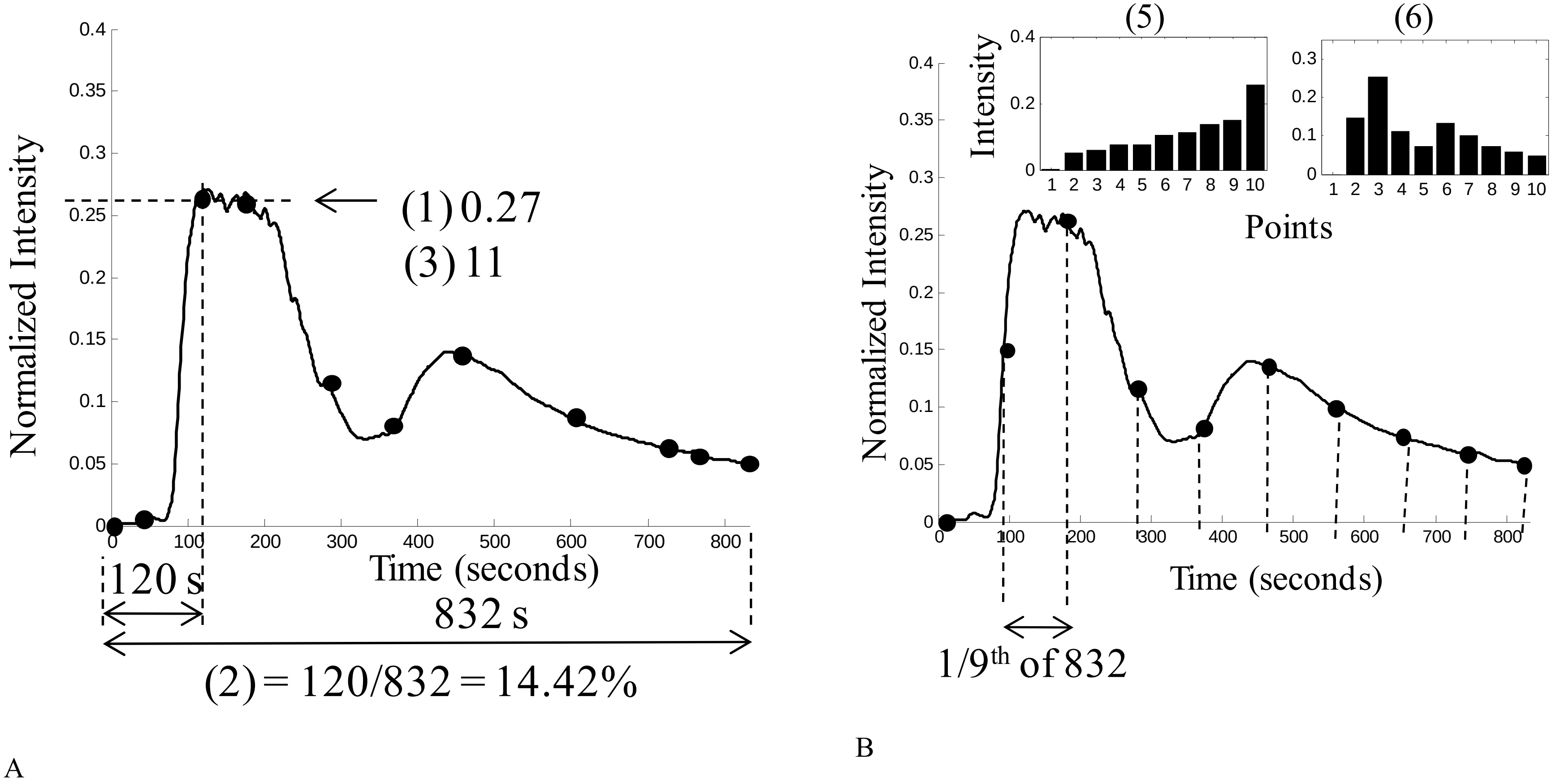
Witten, I., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques.* San Francisco: Morgan Kaufmann Publishers.

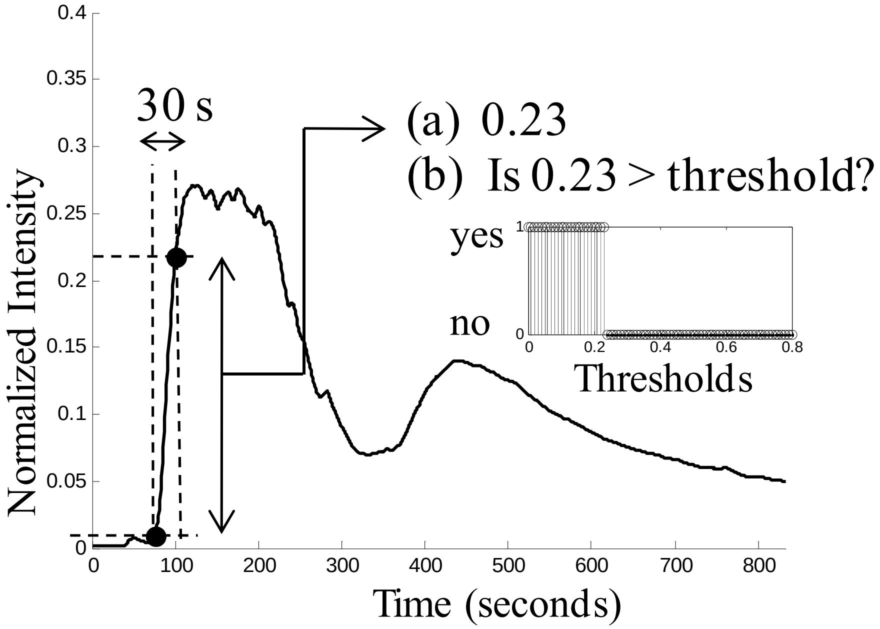
Figure 1: Illustration of the features extracted from each segment of skin conductance to characterize hot flashes.

A (1) Apex value, (2) Relative apex position, and (3) Number of intensity peaks

B (5) Normalized intensities and, (6) Sorted intensities of 10 equally separated points

C 2 µmho-based features: (a) maximum intensity increase in 30 seconds and (b) binary predicates


Normalized Intensity
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
100
200
300
400
500
600
700
800
Time (seconds)
(1) 0.27
(3) 11
120 s
832 s
(2) = 120/832 = 14.42%
A
(5)
(6)
Intensity
Points
1 2 3 4 5 6 7 8 9 10
1/9th of 832
B



30 s
(a) 0.23
(b) Is 0.23 > threshold?
yes
no
1
0
0
0.2
0.4
0.6
0.8
Thresholds
Normalized Intensity
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
100
200
300
400
500
600
700
800
Time (seconds)

C