

# Improving the performance of physiologic hot flash measures with support vector machines

REBECCA C. THURSTON,<sup>a,b</sup> KAREN A. MATTHEWS,<sup>a,b,c</sup> JAVIER HERNANDEZ,<sup>d</sup> AND FERNANDO DE LA TORRE<sup>d</sup>

<sup>a</sup>Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA;

<sup>b</sup>Department of Epidemiology, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania, USA;

<sup>c</sup>Department of Psychology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

<sup>d</sup>Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

## Abstract

Hot flashes are experienced by over 70% of menopausal women. Criteria to classify hot flashes from physiologic signals show variable performance. The primary aim was to compare conventional criteria to Support Vector Machines (SVMs), an advanced machine learning method, to classify hot flashes from sternal skin conductance. Thirty women with  $\geq 4$  hot flashes/day underwent laboratory hot flash testing with skin conductance measurement. Hot flashes were quantified with conventional ( $\geq 2$   $\mu\text{mho}$ , 30 s) and SVM methods. Conventional methods had poor sensitivity (sensitivity = 0.41, specificity = 1, positive predictive value (PPV) = 0.94, negative predictive value (NPV) = 0.85) in classifying hot flashes, with poorest performance among women with high body mass index or anxiety. SVM models showed improved performance (sensitivity = 0.89, specificity = 0.96, PPV = 0.85, NPV = 0.96). SVM may improve the performance of skin conductance measures of hot flashes.

**Descriptors:** Menopause, Anxiety, Behavioral medicine, Sensation/perception, Age, Electrodermal

Hot flashes are considered the “classic” menopausal symptom, with up to 75% of women transitioning through menopause reporting hot flashes (Gold et al., 2006; Kronenberg, 1990). Hot flashes, episodes of intense heat accompanied by flushing and sweating, are associated with significant impairments in quality of life (Avis et al., 2003), mood (Bromberger et al., 2003), and subjective sleep quality (Kravitz et al., 2003) among midlife women. In 2002, the Women’s Health Initiative hormone therapy arms were terminated due to findings of increased health risk associated with hormone therapy (Rossouw et al., 2002), the leading treatment for hot flashes. Thus, the measurement, etiology, and treatment of menopausal hot flashes have been of particular research and clinical interest.

One important area of research has been the physiologic measurement of hot flashes (Miller & Li, 2004). Physiologic hot flash measures have the advantage of avoiding many of the memory and reporting processes that may influence subjective hot flash reports. For example, affective factors such as anxiety are strong correlates of self-reported hot flashes (Freeman et al., 2005; Gold et al., 2006), particularly the reporting of hot

flashes lacking simultaneous physiologic evidence (Thurston, Blumenthal, Babyak, & Sherwood, 2005). In addition, physiologic measures allow for measurement of hot flashes during sleep when women are less likely to subjectively report hot flashes (Miller & Li, 2004; Thurston, Blumenthal, Babyak, & Sherwood, 2006).

Sternal skin conductance is currently the most widely used physiologic measure of hot flashes. Early reports cited the optimal physiologic measure of hot flashes to be sternal skin conductance. One study showed a sternal skin conductance threshold of  $\geq 2$   $\mu\text{mho}$  rise in a 30-s period to have high sensitivity and specificity in classifying hot flashes (Freedman, 1989). While subsequent studies confirmed the utility of sternal skin conductance (de Bakker & Everaerd, 1996), multiple laboratory (de Bakker & Everaerd, 1996; Hanisch, Palmer, Donahue, & Coyne, 2007; Sievert, 2007; Sievert et al., 2002) and ambulatory studies found low sensitivity of the 2  $\mu\text{mho}/30$  s criterion (Hanisch, Palmer, Donahue, & Coyne, 2007; Thurston, Blumenthal, Babyak, & Sherwood, 2005). However, there has not been widespread agreement on an alternate definition in classifying hot flashes from sternal skin conductance, and the 2  $\mu\text{mho}/30$  s criterion continues in widespread use.

One limitation of the 2  $\mu\text{mho}/30$  s criterion is that it applies a single threshold to classify hot flashes across all women. Thus, it cannot accommodate between-subject variability in hot flash-associated skin conductance rises. Moreover, this approach fails to capture the characteristic shape of hot flash skin conductance

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

Address reprint requests to: Rebecca C. Thurston, Ph.D., 3811 O’Hara St., Pittsburgh, PA 15213. E-mail: thurstonrc@upmc.edu

events, classifying artifactual 2  $\mu$ mho rises as hot flashes and necessitating manual visual editing.

A more sophisticated approach for classifying hot flashes are Support Vector Machines (SVMs) (Boser, Guyon, & Vapnik, 1992; Shawe-Taylor & Cristianini, 2000). SVMs are state-of-the-art classification methods that have shown excellent performance in complicated pattern recognition problems (Guyon, Weston, Barnhill, & Vapnik, 2002; Joachims, 1998; Michel & Kaliouby, 2003). The goal of SVM in this setting is to classify segments of data (e.g., skin conductance signals) as a hot flash or as not a hot flash. However, rather than implementing a single threshold across women, SVM characterizes the shape of the hot flash-associated skin conductance change. To do so, SVM algorithms learn model parameters that provide maximum discrimination between hot flashes and non-hot flash skin conductance patterns, typically projecting data points into high dimensional space to achieve this property. SVMs are notable for their ability to find a unique optimal solution efficiently, their flexibility to incorporate multiple types of data, their ability to model nonlinear patterns, and the generalization of an SVM algorithm developed on one sample to previously unobserved samples. SVMs have a well-founded theoretical foundation, and although a discussion of which is beyond the scope of the present paper, we refer the reader to excellent introductions (Burges, 1998; Hearst, 1998; Noble, 2006) as well as more advanced material (Shawe-Taylor & Cristianini, 2004) about SVM.

SVMs have several advantages in detecting hot flashes over conventional hot flash classification methods. First, they are able to characterize patterns of change in the skin conductance signal, as opposed to a single magnitude-based threshold. Second, SVMs allow for automated discrimination of hot flashes from artifact. Moreover, SVM can be developed for group-specific or woman-specific models. Finally, SVMs allow for simultaneous use of multiple forms of input data in classifying hot flashes. These multiple inputs can include multiple physiologic indices as well as subject characteristics.

Previous investigators have emphasized the importance of considering subject characteristics in physiologic measures of hot flashes. Finding poor performance of conventional physiologic hot flash measures in ethnically diverse samples, Sievert and colleagues have questioned whether the performance of and optimal skin conductance sampling site for hot flashes may vary across racial/ethnic groups (Sievert et al., 2002). However, there has not been formal laboratory validation of physiologic measures of hot flashes across racial/ethnic groups. Notably, in studies using questionnaire measures, pronounced differences in hot flash reporting by race/ethnicity are apparent, with African American women reporting more frequent and bothersome hot flashes versus non-Hispanic Caucasian women (Gold et al., 2006; Thurston et al., 2008). These racial/ethnic differences are not well understood, but have been attributed in part to differences in body size or composition between groups (Crawford, 2007) and need to be investigated with physiologic hot flash measures. However, the validity of physiologic measures across women of varying body sizes or race/ethnicities is not understood.

The primary aim of this study was to compare the performance of conventional criteria (2  $\mu$ mho/30 s) and state-of-the-art machine learning methods (SVMs) to detect hot flashes from sternal skin conductance signals. Secondary aims were to compare the performance of these measures 1) across key subject characteristics (race/ethnicity, obesity status, anxiety), and 2) across sampling sites (arm, sternum, upper trapezius).

## Methods

### Subjects

Thirty-four African American and Caucasian women between the ages of 40 and 60 were recruited. Inclusion criteria included late perimenopausal (amenorrhea of 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 norepinephrine reuptake inhibitors, clonidine, methyl dopa, bellergal, gabapentin, aromatase inhibitors, or 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, reporting medical or psychiatric conditions associated with hot flash sensations (panic disorder, pheochromocytoma, leukemia, pancreatic tumor), or the inability to provide informed consent and follow study procedures. Of these 34 women, one woman was excluded due to equipment failure during the session, and three women were excluded due to no reported nor physiologically detected hot flashes during the laboratory session, for a final sample of 30 women. While two of the 30 women did not report hot flashes during the session, they were included here due to evidence of skin conductance changes potentially consistent with a hot flash.

### Procedures

Laboratory testing procedures took place in a 24 ( $\pm$  1) $^{\circ}$ C temperature controlled room between the hours of 12:00 and 19:00, given the circadian rhythm of hot flashes with peak frequency occurring during the late afternoon hours (Freedman, Norton, Woodward, & Cornelissen, 1995; Thurston, Blumenthal, Babyak, & Sherwood, 2005). Participants underwent testing in a seated position and wore a light-weight cotton hospital “scrub” top and pants.

Participants underwent four laboratory tasks: observation, mental stress, heating, and a cold pressor task. Heating has been previously shown to induce hot flashes among symptomatic women (Freedman, 1989; Freedman & Dinsay, 2000; Germaine & Freedman, 1984), and some suggestive evidence indicates that mental stressors may be associated with increased hot flash occurrence (Swartzman, Edelberg, & Kemmann, 1990). Speech and cold pressor tasks have been shown to reliably provoke a stress response, including among postmenopausal women (Saab, Matthews, Stoney, & McDonald, 1989). During the observation session, participants sat quietly for 30 m, during which time they read magazines or completed study self-report instruments. For the speech task, participants prepared (2 m) and delivered (3 m) a speech about an assigned topic while being observed by an experimenter, followed by a 5-m rest period. For the heat provocation, a 22"  $\times$  14.5" electric heating pad maintained at a temperature of 42 ( $\pm$  1) $^{\circ}$ C was placed on the participant's torso for 30 m. For the cold pressor task, an ice pack was placed on the participant's forehead for up to 60 s. Participants subjectively reported hot flashes via pressing a time and date stamped event marker and rating the severity of the hot flash (mild (1)–severe (4)). Skin conductance, skin temperature, and heart rate were measured continuously throughout testing.

Participants also underwent measurement of height, weight, and waist circumference, and completed a battery of question-

naires for assessment of medical, demographic, and psychological characteristics. All study procedures were approved by the University of Pittsburgh Institutional Review Board, and all participants provided written informed consent.

### Measures

Skin conductance, the main physiologic index of hot flashes, was recorded from the sternum, upper trapezius, and lateral deltoid of the left arm with a 0.5 V constant voltage circuit sampling from two silver/silver chloride electrodes (Vermed Inc, Bellows Falls, VT, USA) at each site filled with 0.05 M KCL Velvachol/glycol paste (Dormire & Carpenter, 2002). Skin temperature and heart rate were also recorded. Skin temperature was recorded with Yellow Springs 400 series thermistor probes (YSI, Yellow Springs, OH) taped to the pad and dorsal surface of the distal phalanx of the third finger (de Bakker & Everaerd, 1996; Freedman, 1989; Germaine & Freedman, 1984; Tataryn et al., 1981). Heart rate was measured by ECG via three silver/silver chloride electrodes (Kendall, Syracuse, NY, USA) in a standard 3-lead configuration. Skin conductance, skin temperature, and heart rate signals were recorded via Grass polygraph (model 7D, skin conductance adaptor SCA1, temperature probe adaptor TPA, Grass Technologies, Astro-Med Inc., West Warwick, RI, USA) and digitized at 1 KHz by an analogue to digital converter.

Height and weight were measured via a fixed stadiometer and a calibrated balance beam scale, respectively. Waist circumference was measured via tape measure at the level of the natural waist or the narrowest part of the torso from the anterior aspect; if a waist narrowing was difficult to identify, the measure was taken at the smallest horizontal circumference between the ribs and iliac crest. 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), state and trait anxiety via the Spielberger State Trait Anxiety Inventory (Spielberger, 1983), and perceived stress via the 10-item Perceived Stress Scale (Cohen, Kamarck, & Mermelstein, 1983). In addition, somatization was assessed via the somatization subscale of the Symptom Checklist-90 (Derogatis, 1983), symptom sensitivity via the symptom sensitivity scale (Barsky, Goodson, Lane, & Cleary, 1988), and physical activity via the Paffenbarger scale (Paffenbarger, Wing, & Hyde, 1978).

### Data Reduction

Physiologic hot flashes were classified in two ways: 1)  $\geq 2$   $\mu\text{mho}$  rise in 30 s (conventional criterion), and 2) SVM-defined hot flashes. For both models, a 5-m reporting window and 10-m lockout period was implemented, after the start of the flash, after which no hot flashes were coded. For the conventional criterion, skin conductance increases of  $\geq 2$   $\mu\text{mho}$  in 30 s (Freedman, 1989) were flagged automatically by custom software and edited for artifact using standard methods (Carpenter, Andrykowski, Freedman, & Munn, 1999).

Building an SVM model consists of both a training phase, where the model is developed, and a testing phase, where model performance is tested. In the training phase, the entire sample of hot flashes and non-hot flash skin conductance data segments are labeled. From this labeled data, the SVM function learns the characteristics of a hot flash-associated and non-hot flash associated skin conductance changes. While the primary data input into the SVM here is skin conductance, the SVM function can

also include additional data inputs such as subject characteristics known to modify the skin conductance signal or hot flash reporting. Once the SVM function is trained, a testing phase ensues in which SVM classifies each new segment of skin conductance data as a hot flash or not a hot flash and the accuracy of SVM classification calculated. The SVM-LIB, a publicly available Matlab library, is used to implement the core of SVM training and testing (Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001, <http://www.csie.nyu.edu.tw/~cjlin/libsvm/>). Although specialized expertise is required to build an SVM for optimal performance, this publicly available library provides SVM tutorials, sample programming code, and program code libraries that form the core of training and testing the SVM algorithm.

For this specific classification problem, the detection of hot flashes using an SVM was formulated as a binary classification problem (yes/no hot flash). 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), in creating the labeled dataset to train SVM, 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, Andrykowski, Freedman, & Munn, 1999). Thus, all data were visually reviewed and hot flash and non-hot flash intervals labeled, with hot flashes labeled based upon their characteristic skin conductance shape rather than solely the magnitude of the rise. Next, several signal pre-processing steps were performed on all skin conductance data to remove noise and temporally normalize the signal. First, an exponential smoothing filter ( $\alpha = 0.08$ ) was used to average noise in the signal. Second, given the wide variability in hot flash-associated skin conductance rises and duration across subjects, the skin conductance amplitude was normalized between zero and one within each subject’s session, and skin conductance segment durations were linearly re-scaled to the maximum, minimum, and mean duration of training hot flashes. The SVM was trained from this normalized skin conductance signal. Finally, several subject characteristics previously documented to be associated with hot flashes (e.g., age, race, body mass index [BMI], menopausal status, smoking) (Gold et al., 2006; Thurston, Blumenthal, Babyak, & Sherwood, 2005), were considered for inclusion as additional data inputs into SVM models. Those improving model performance were included; thus all SVM models were optimized for race/ethnicity and obesity status, and additionally anxiety for models with self-reported hot flashes as the reference.

For the testing phase, using SVM-LIB, we implemented a leave-one-out strategy, a useful method when limited quantities of data are available for training and testing, as in the present study (Witten & Frank, 2005). The leave-one-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 built. This is repeated, with re-estimation of SVM each time, such that each session in the sample is used once as validation data, and the remainder as training data. The final SVM includes data from all subjects for maximum generalizability to other samples.

True positive (TP), false positive (FP), false negative (FN), and true negative (TN) hot flashes were scored relative to self-

reported hot flashes, and for SVM also to expert-defined hot flashes. A TP was scored when a physiologic hot flash was accompanied by a hot flash report, a FP when a physiologic hot flash was not met by a hot flash report, a FN when a hot flash report was not accompanied by a physiologic hot flash, and a TN for all 10-m intervals lacking both a hot flash report and physiologic hot flash. For each individual and hot flash definition (2  $\mu$ mho, SVM), the sensitivity (TP/TP+FN), specificity (TN/FP+TN), positive predictive value (PPV; TP/TP+FP), and negative predictive value (NPV; TN/TN+FN) was calculated. Thus, in the case of self report as the criterion, sensitivity corresponds to the percentage of hot flash reports also accompanied by a physiologic hot flash, specificity to the percentage of 10-m segments without a hot flash report also lacking a physiologic hot flash, PPV to the percentage of physiologic hot flashes that are also reported, and NPV to the percentage of segments with no physiologic hot flash that also lack a hot flash report.

Skin temperature and heart rate were also examined together with skin conductance using previously established criteria (Germaine & Freedman, 1984) and SVM. Since results were not improved for any model with inclusion of these indices, they are not presented here. SVM models were also performed for alternate skin conductance sites, but since results were comparable (arm) to or worse (trapezius) than the sternal site, they are not presented here.

#### Data Analysis

Comparisons of subject characteristics by race/ethnicity were conducted using *t*-test and chi-square tests. Comparisons of reporting rates by subject characteristics were performed using Spearman's rho and linear regression, with transformation of rates as necessary. Statistical comparisons for specificity and PPV using the conventional 2  $\mu$ mho criterion were limited by the uniformly high specificity (resulting in limited range), and low rate of hot flash reports physiologically detected with this criterion (resulting in a zero denominator for multiple PPV values), respectively. Sensitivity using the conventional criterion was log transformed and all SVM indices were square root transformed for analysis. Characteristics considered included age, race/ethnicity, BMI, waist circumference, menopausal stage, parity, education, marital status, physical activity, alcohol use, smoking status, depression symptoms, state and trait anxiety, perceived stress, somatization, and symptom sensitivity. While there was one extreme observation on the Paffenbarger scale (activity value = 19055), exclusion of this observation did not alter conclusions. The Paffenbarger scale was log transformed for analysis. The area under the receiver operating characteristic (ROC) curve (AUC) was calculated for both conventional and SVM definitions of hot flashes (Swets, 1988). Analyses were performed using SAS v.8.02 (SAS Institute, Cary, NC, USA) and MATLAB v.7.0 (MathWorks, Natick, MA, USA). Tests were 2-sided with  $\alpha = 0.05$ .

#### Results

Participants on average were 53 years old, postmenopausal, and overweight or obese (Table 1). African American women were more likely to be smokers, and had higher levels of depressive symptoms, state and trait anxiety, and perceived stress than Caucasian women. Participants retrospectively reported on average 7 daily hot flashes at the time of screening.

A total of 62 hot flashes were reported during the laboratory session, 25 of which occurred during observation, 13 during the speech task or subsequent 5-m rest period, 19 during heating, 1 during cold pressor, and 4 during periods preceding or following laboratory tasks. The majority of hot flashes were given ratings of "1 (mild)" ( $n = 23$ , 37%) or "2" ( $n = 22$ , 35%), with 24% rated as "3" ( $n = 8$ , 13%) or "4 (severe)" ( $n = 7$ , 11%) (severity ratings were unavailable for two reported hot flashes). Neither the number of reported or physiologically detected hot flashes (by the conventional criterion), nor their mean intensity rating differed by ethnic group.

#### Conventional Criterion

The sensitivity of the conventional 2  $\mu$ mho/30 s criterion to characterize hot flashes was low across both racial/ethnic groups (Table 2). A similar pattern of low sensitivity was observed across skin conductance sampling sites (arm: 0.37, trapezius: 0.44), which did not vary by race/ethnicity. Results were also examined using an alternate criterion suggested by other investigators (1.78  $\mu$ mho/45 s) (Hanisch, Palmer, Donahue, & Coyne, 2007), but this criterion also showed low sensitivity (sensitivity = 0.50, specificity = 0.99, PPV = 0.94, NPV = 0.87). Results were also examined for the conventional criterion relative to expert-defined hot flashes, with similar low sensitivity (sensitivity = 0.47, specificity = 0.99, PPV = 0.94, NPV = 0.87). Sensitivity of the conventional criterion relative to reported hot flashes was particularly low among hot flashes given a rating of "1, mild" (sensitivity = 0.20), versus those rated as "2," "3," or "4, severe," (sensitivity = 0.46).

We next examined additional participant characteristics associated with the performance of the 2  $\mu$ mho criterion. Among available psychological, medical, and demographic factors, higher state ( $r = -.47$ ,  $p = .01$ ) and trait anxiety ( $r = -.38$ ,  $p = .04$ ) was associated with lower sensitivity of the measure. Higher state anxiety ( $r = -.38$ ,  $p = .03$ ) was also related to lower NPV. Higher BMI and waist circumference were marginally associated with lower sensitivity ( $r = -.33$ ,  $p = .08$ ) and NPV ( $r = -.31$ ,  $p = .09$ ), respectively. When BMI and state anxiety were included together in a regression model in relation to sensitivity (log transformed), anxiety remained significantly associated with sensitivity ( $b = -0.02$ ,  $p = .04$ ) whereas BMI was somewhat attenuated ( $b = -0.01$ ,  $p = .11$ ). Table 3 shows the variation of the performance of the measure by obesity category and state anxiety.

Notably, pronounced variability in the magnitude of skin conductance rises associated with hot flash reports was apparent (0–6.26  $\mu$ mho/30 s). Of the 62 hot flash reports, the majority (65%,  $n = 40$ ) were associated with <2  $\mu$ mho sternal skin conductance rise (range 0–1.95  $\mu$ mho/30 s), and a subset of these ( $n = 14$ , 79% of which were rated as mild) across 5 women were associated with no apparent skin conductance rise. Thus, false negatives included both "submaximal" rises as well as hot flash reports accompanied by no skin conductance rise. Further investigation of the 48 hot flash reports accompanied by some skin conductance rise revealed a pattern in which BMI tended to be inversely related to the magnitude of sternal skin conductance rises with hot flash reports ( $r = -.33$ ,  $p = .09$ ). Conversely, higher state anxiety ( $r = .37$ ,  $p = .04$ ) and perceived stress ( $r = .40$ ,  $p = .03$ ) were associated with a greater number hot flash reports accompanied by a complete absence of a sternal skin conductance rise.

**Table 1.** Subject Characteristics

	Total Sample ( <i>N</i> = 30)	African American ( <i>N</i> = 15)	Caucasian ( <i>N</i> = 15)	<i>P</i> value*
Age (M, SD)	52.5 (5.3)	51.5 (6.5)	53.5 (3.8)	.30
BMI				
< 25	11 (36.7)	6 (40.0)	5 (33.3)	.76
25–< 30	7 (23.3)	4 (26.7)	3 (20.0)	
30 (severity ratings were unavailable for two reported hot flashes)	12 (40.0)	5 (33.3)	7 (46.7)	
Education ( <i>n</i> , %)				
≤ High school	9 (30.0)	4 (44.4)	5 (55.6)	.13
Some college/votech	13 (43.3)	9 (69.2)	4 (30.8)	
College (severity ratings were unavailable for two reported hot flashes)	8 (26.7)	2 (25.0)	6 (75.0)	
Waist circumference (cm, M, SD)	36.6 (6.7)	38.3 (6.5)	34.9 (6.6)	.17
Menopausal status ( <i>n</i> , %)				
Late perimenopause	12 (40.0)	7 (58.3)	5 (41.7)	.46
Postmenopause	18 (60.0)	8 (44.4)	10 (55.6)	
Current smoker ( <i>n</i> , %)	9 (30.0)	8 (88.9)	1 (11.1)	.005
Depressive symptoms (M, SD)	5.0 (3.7)	6.5 (4.0)	3.6 (2.9)	.04
State anxiety (M, SD)	29.3 (7.2)	32.4 (7.8)	26.2 (5.0)	.02
Trait anxiety (M, SD)	33.0 (7.2)	36.1 (7.6)	29.9 (5.3)	.02
Number of laboratory reported hot flashes (median, interquartile range)	2 (2)	2 (2)	2 (2)	.99
Number of physiologically detected hot flashes (median, interquartile range)†	1 (1)	1 (2)	2 (1)	.83
Mean intensity rating (M, SD)‡	2.0 (0.8)	2.0 (0.8)	2.0 (0.8)	.90

\*For ethnicity comparison.

†2  $\mu\text{mho}/30$  s criteria.

‡On a scale of 1 (mild) – 4 (severe).

### Support Vector Machine Models

SVM-defined hot flashes showed improved performance in defining hot flashes across women, with no differences by ethnicity (Table 4). While there remained some variation by obesity status and state anxiety (Table 5), the performance of this measure was improved across all groups. Consistent with a role of anxiety in hot flash reporting (Thurston, Blumenthal, Babyak, & Sherwood, 2005), anxiety differences remained only among models using self-reported hot flashes as the reference. For SVM models, sensitivity was slightly higher among hot flashes given a severity rating of “2” or higher (sensitivity = 0.84) versus those rated “1, mild” (sensitivity = 0.76). The AUC of the SVM models were 0.92 and 0.89 for expert defined and subject report references, respectively, in comparison to 0.75 for conventional criterion.

### Discussion

This study indicated that conventional widely used criteria to classify hot flashes had low sensitivity in the laboratory. The performance of this measure was similar across Caucasian and African American women and was not improved implementing alternate sampling sites. Moreover, it showed pronounced vari-

**Table 2.** Performance of Sternal Skin Conductance (2  $\mu\text{mho}/30$  s Criterion) across Ethnic Groups

	Total Sample	African American	Caucasian
Sensitivity†	0.41	0.43	0.41
Specificity‡	1.00	0.99	1.00
PPV§	0.94	0.88	1.00
NPV¶	0.85	0.86	0.85

Note. All ethnicity comparisons nonsignificant.

†*N* = 28, ‡*N* = 30, §*N* = 16, ¶*N* = 30.

ability across key participant characteristics, with particularly low sensitivity observed among women with high BMI or anxiety. Finally, application of a sophisticated pattern classification approach, SVM, improved the performance of sternal skin conductance in classifying hot flashes across women.

This study found the conventional 2  $\mu\text{mho}/30$  s threshold to have low sensitivity in classifying hot flashes. While one important investigation found strong performance of this threshold (Freedman, 1989), subsequent studies have failed to replicate this finding. For example, de Bakker and colleagues showed the conventional criterion had a laboratory sensitivity of 0.37 (de Bakker & Everaerd, 1996). Several more recent investigations have also found low sensitivity of this threshold (Hanisch, Palmer, Donahue, & Coyne, 2007; Sievert et al., 2002). One study among men with prostate cancer suggested a 1.78  $\mu\text{mho}/45$  s threshold (Hanisch, Palmer, Donahue, & Coyne, 2007), which also had low sensitivity in the present investigation. Although these studies varied in devices and patient populations, there is a consistency in findings of low sensitivity of the conventional criterion.

A single threshold may not be adequate to characterize hot flashes across women. Sensitivity and specificity exist in a dynamic balance, and lowering the skin conductance threshold increased sensitivity, but also decreased specificity. This property is particularly important given the wide variation in skin conductance rises associated with hot flashes, many of which were well under 1  $\mu\text{mho}$ . Notably, hot flash-associated skin conductance rises show a characteristic shape that can be distinguished from artifactual rises (Carpenter, Andrykowski, Freedman, & Munn, 1999) that frequently meet the 2  $\mu\text{mho}$  threshold and must be removed by manual editing. Therefore, not only the initial rise, but the pattern of skin conductance changes characterizes hot flashes. Algorithms that capture this pattern, rather than applying a single threshold, have the potential to optimally classify hot

**Table 3.** Performance of Sternal Skin Conductance (2 μmho/30 s criterion) by Obesity Status and State Anxiety

	Obesity Status			State Anxiety	
	Normal (N = 11)	Overweight (N = 7)	Obese (N = 12)	Low (N = 15)	High (N = 15)
Sensitivity†	0.61	0.52	0.20*	0.63	0.16**
Specificity‡	0.99	1.00	1.00	1	0.99
PPV§	0.88	1.00	1.00	1	0.80
NPV¶	0.91	0.90	0.77*	0.89	0.81

Note: Normal weight BMI <25 (reference), overweight BMI 25–29.9, obese BMI ≥30; low state anxiety ≤27.5 (reference), high state anxiety >27.5. †N = 28, ‡N = 30, §N = 16, ¶N = 30, \*p < .05, \*\*p < .01.

flashes. These models should also have the ability to maximally discriminate hot flashes from artifact in an automated fashion as well as allow optimization by subject characteristics.

SVM, a state-of-the-art classification method, meets these criteria, imparting it with several advantages over the single threshold approach. First, SVM can learn a pattern of hot flashes versus artifact events to optimally characterize the dynamics of physiologic signals associated with hot flashes, allowing greater classification accuracy. The need for visual editing, with associated possible introduction of bias, is reduced. In addition, SVM can learn parameters to adapt the algorithm to group-specific or woman-specific situations. This feature is important given the potential for the signal to vary with subject characteristics. Moreover, the SVM can be retrained as more training data are available, allowing for continued improvement of algorithms. Further, SVM can integrate different types of information, including other physiologic signals, into the classification process. This feature has the potential to be important in quantifying sources of artifact, particularly in the ambulatory environment. In light of these capabilities, SVM outperformed the conventional approach in the present investigation.

The performance of sternal skin conductance in classifying hot flashes varied by key subject characteristics, with particularly low sensitivity among women with a higher BMI or reporting high state anxiety at the time of testing. There were several factors driving this low sensitivity, including hot flash reports accompanied by “submaximal” rises failing to meet the conventional threshold, as well as hot flash reports accompanied by no apparent skin conductance rise. Exploratory analyses suggested that BMI tended to be inversely related to the magnitude of skin conductance rises with hot flash reports. Conversely, state anxiety was most consistently related to the reporting of hot flashes associated with no apparent skin conductance changes. Given the post-hoc nature of and limited power for these analyses, they should be interpreted with caution. However, they may suggest physiologic differences by body

composition influencing the skin conductance signal, such as subcutaneous adipose tissue modifying the density or functioning of sweat glands or acclimatization (Havenith & van Middendorp, 1990). Conversely, anxiety may influence the perception and reporting of hot flashes. The role of affective factors in symptom reporting is well established (Cohen et al., 1995; Watson & Pennebaker, 1989), anxiety is the most consistent predictor of questionnaire-reported hot flashes (Freeman et al., 2005; Gold et al., 2006), and anxiety has previously been linked to the increased reporting of hot flashes lacking physiologic evidence (Thurston, Blumenthal, Babyak, & Sherwood, 2005). Consistent with this body of literature, anxiety differences remained only for SVM models using reported hot flashes as the reference.

Previous investigators have noted the potential for variation in skin conductance measures of hot flashes by race/ethnicity. Finding low sensitivity sternal skin conductance in Mexican women, Sievert and colleagues suggested that alternate skin conductance sampling sites may more optimally classify hot flashes across different racial/ethnic groups (Sievert et al., 2002). However, they later found little compelling support for this hypothesis (Sievert, 2007). Similarly, we did not find evidence of improved or dramatically different performance across the alternate sites and across the African American and non-Hispanic Caucasian women in this investigation. This was true of the conventional criterion as well as with SVM (data not shown).

The study results should be interpreted in the context of several limitations. First, the study sample, although consistent with previous validation studies, was small, and power may have been limited for secondary/exploratory analyses. Moreover, an older Grass polygraph was used, potentially increasing error and exacerbating the poor sensitivity of the threshold. However, poor performance was noted only for sensitivity, not all performance indices as would be expected with error introduced by equipment, and the present results are broadly consistent with prior laboratory and ambulatory studies measuring skin conductance using a range of devices. Furthermore, the manufacturer of the

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

	SVM (reference, expert defined hot flashes)‡			SVM (reference, subject reported hot flashes)		
	Total Sample	African American	Caucasian	Total Sample	African American	Caucasian
Sensitivity†	0.89	0.95	0.85	0.84	0.81	0.87
Specificity‡	0.96	0.97	0.96	0.95	0.94	0.96
PPV§	0.85	0.81	0.88	0.84	0.78	0.88
NPV¶	0.96	0.97	0.95	0.93	0.92	0.95

Note: All ethnicity comparisons nonsignificant. ‡Optimized by race and obesity status. ||Optimized by race, obesity status, and anxiety. †N = 28, ‡N = 30, §N = 30, ¶N = 30.

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

	SVM (reference, expert defined hot flashes)¶					SVM (reference, subject reported hot flashes)				
	Obesity Status			State Anxiety		Obesity Status			State Anxiety	
	Normal (N = 11)	Overweight (N = 7)	Obese (N = 12)	Low (N = 15)	High (N = 15)	Normal (N = 11)	Overweight (N = 7)	Obese (N = 12)	Low (N = 15)	High (N = 15)
Sensitivity#	1.00	1.00	0.74*	0.95	0.81	0.94	0.83	0.77	0.95	0.71*
Specificity‡	0.97	0.98	0.94	0.96	0.96	0.95	0.98	0.92	0.96	0.93
PPV§	0.82	0.92	0.84	0.92	0.77	0.77	0.93	0.85	0.92	0.77‡
NPV¶	1.00	1.00	0.91*	0.97	0.95	0.98	0.94	0.87*	0.97	0.89

Note: Normal weight BMI <25 (reference), overweight BMI 25–29.9, obese BMI ≥30; low state anxiety ≤27.5 (reference), high state anxiety >27.5.

¶Optimized by race and obesity status.

||Optimized by race, obesity status, and anxiety.

#N = 28, ‡N = 30, §N = 30, ¶N = 30, †p < 0.10, \*p < 0.05.

electrodes used for skin conductance measurement of hot flashes has recently changed, although the size of electrodes has remained consistent, and those used in the present study are the current standard for hot flash measurement. Moreover, building SVM models from raw skin conductance signals requires specific expertise in building and testing SVM models. Although one strength of SVMs are their generalizability across independent samples (Borges, 1998), the results from SVM models presented here should be regarded as preliminary, requiring further development, testing among independent samples, and extension to the ambulatory environment.

This study has several notable strengths. Although variations in the performance of the skin conductance measure between ethnic groups have been previously noted (Sievert, 2007; Sievert, et al., 2002), this study is among the first to formally validate skin conductance measures across different African American and non-Hispanic Caucasian women. Moreover, this study included examination of multiple skin conductance sampling sites as well as multiple physiologic indices. Furthermore, this study assessed multiple psychological, physiological, and demographic subject

characteristics, and is among the few to examine variations in the performance of this measure by these characteristics. Finally, this study proposed use of a new, innovative method for characterizing hot flashes from skin conductance signal using advanced machine learning methods. Use of these methods has the potential to further advance the physiologic measurement of hot flashes.

This study indicated that standard methods to classify hot flashes may have unacceptably low sensitivity. This low sensitivity may be particularly acute among women who have key characteristics such as higher BMI or anxiety. This study proposed application of SVM methods to classify hot flashes from physiologic signals, allowing a more sophisticated characterization of not only the rise, but the pattern of skin conductance changes with hot flashes. Those models also allow optimization by subject characteristics. This improved method to characterize hot flashes has the potential to improve the functioning of physiologic measures for more widespread use in investigations aimed at understanding the etiology and developing new treatments for hot flashes.

## REFERENCES

- 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.
- Barsky, A. J., Goodson, J. D., Lane, R. S., & Cleary, P. D. (1988). The amplification of somatic symptoms. *Psychosom Med, 50*, 510–519.
- Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144–152). New York: ACM Press.
- Bromberger, J. T., Assmann, S. F., Avis, N. E., Schocken, M., Kravitz, H. M., & Cordal, A. (2003). Persistent mood symptoms in a multiethnic community cohort of pre- and perimenopausal women. *American Journal of Epidemiology, 158*, 347–356.
- Borges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*, 121–167.
- 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 Personality and Social Psychology, 68*, 159–169.
- Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior, 24*, 385–396.
- Crawford, S. L. (2007). The roles of biologic and nonbiologic factors in cultural differences in vasomotor symptoms measured by surveys. *Menopause, 14*, 725–733.
- de Bakker, I. P., & Everaerd, W. (1996). Measurement of menopausal hot flushes: validation and cross-validation. *Maturitas, 25*, 87–98.
- Derogatis, L. R. (1983). *SCL-90-R, Administration, Scoring, and Procedures Manual for the Revised Version* (2nd ed). Towson, MD: Clinical Psychometric Research.
- 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.
- Freedman, R. R. (1989). Laboratory and ambulatory monitoring of menopausal hot flashes. *Psychophysiology, 26*, 573–579.
- Freedman, R. R., & Dinsay, R. (2000). Clonidine raises the sweating threshold in symptomatic but not in asymptomatic postmenopausal women. *Fertility and Sterility, 74*, 20–23.
- Freedman, R. R., Norton, D., Woodward, S., & Cornelissen, G. (1995). Core body temperature and circadian rhythm of hot flashes in menopausal women. *Journal of Clinical Endocrinology and Metabolism, 80*, 2354–2358.
- 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.
- Germaine, L. M., & Freedman, R. R. (1984). Behavioral treatment of menopausal hot flashes: Evaluation by objective methods. *Journal of Consulting and Clinical Psychology, 52*, 1072–1079.
- Gold, E., Colvin, A., Avis, N., Bromberger, J., Greendale, G., Powell, 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.
- Havenith, G., & van Middendorp, H. (1990). The relative influence of physical fitness, acclimatization state, anthropometric measures and gender on individual reactions to heat stress. *European Journal of Applied Physiology and Occupational Physiology*, *61*, 419–427.
- Hearst, M. (1998). Support vector machines. *IEEE Intelligent Systems*, *13*, 18–28.
- Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the European Conference on Machine Learning*. Berlin: Springer.
- Kravitz, H. M., Ganz, P. A., Bromberger, J., Powell, L. H., Sutton-Tyrrell, K., & Meyer, P. M. (2003). Sleep difficulty in women at midlife: A community survey of sleep and the menopausal transition. *Menopause*, *10*, 19–28.
- Kronenberg, F. (1990). Hot flashes: Epidemiology and physiology. *Annals of the New York Academy of Sciences*, *592*, 52–86; discussion 123–133.
- 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 Clinic Proceedings*, *79*, 777–781.
- Noble, W. (2006). What is a support vector machine? *Nature Biotechnology*, *24*, 1565–1567.
- Paffenbarger, R. S., Jr., Wing, A. L., & Hyde, R. T. (1978). Physical activity as an index of heart attack risk in college alumni. *American Journal of Epidemiology*, *108*, 161–175.
- 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. *JAMA*, *288*, 321–333.
- Saab, P. G., Matthews, K. A., Stoney, C. M., & McDonald, R. H. (1989). Premenopausal and postmenopausal women differ in their cardiovascular and neuroendocrine responses to behavioral stressors. *Psychophysiology*, *26*, 270–280.
- Shawe-Taylor, J., & Cristianini, N. (2000). *Support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.
- Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press.
- 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.
- Spielberger, C. D. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto: Consulting Psychologists Press.
- Swartzman, L. C., Edelberg, R., & Kemmann, E. (1990). Impact of stress on objectively recorded menopausal hot flushes and on flush report bias. *Health Psychol*, *9*, 529–545.
- Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*, 1285–1293.
- Tataryn, I. V., Lomax, P., Meldrum, D. R., Bajorek, J. G., Chesarek, W., & Judd, H. L. (1981). Objective techniques for the assessment of postmenopausal hot flashes. *Obstetrics and Gynecology*, *57*, 340–344.
- Thurston, R., Bromberger, J., Joffe, H., Avis, N., Hess, R., Crandall, C., et al. (2008). Beyond frequency: Who is most bothered by vasomotor symptoms? *Menopause*, *15*, 841–847.
- Thurston, R. C., Blumenthal, J. A., Babyak, M. A., & Sherwood, A. (2005). Emotional antecedents of hot flashes during daily life. *Psychosomatic 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 Behavioral Medicine*, *13*, 163–172.
- Watson, D., & Pennebaker, J. W. (1989). Health complaints, stress, and distress: Exploring the central role of negative affectivity. *Psychological Review*, *96*, 234–254.
- Witten, I. H., & Frank, E. (2005). Credibility: Evaluating what's been learned. *Data mining: Practical machine learning tools and techniques* (pp. 149–151). San Francisco: Morgan Kaufmann Publishers Inc.

(RECEIVED March 4, 2008; ACCEPTED June 15, 2008.)