EDITOR: Erik Cambria, cambria@ntu.edu.sg

DEPARTMENT: AFFECTIVE COMPUTING AND SENTIMENT ANALYSIS

Affect Detection From Wearables in the "Real" Wild: Fact, Fantasy, or Somewhere In between?

Sidney K. D'Mello and Brandon M. Booth, *University of Colorado Boulder, Boulder, CO, 80309, USA*

Affect detection from wearables in the "real" wild—where people go about their daily routines in heterogeneous contexts—is a different problem than affect detection in the lab or in the "quasi" wild (e.g., curated or restricted contexts). The U.S. government recently supported a program to develop and evaluate the performance of contemporary affect detection systems in the real-wild along the dimensions of accuracy, robustness, and generalizability. Evaluations by an independent testing team revealed that none of the performing teams met the aspirational performance metrics. Alarmingly, performance was near zero for several cases. This article is the result of soul searching to reconcile the chasm between expected and achieved performance in light of past successes of the field. We discuss the major challenges faced, their implications for future research, and suggest a path forward.

The ability to autonomously, ubiquitously, accurately, and robustly infer affect as people go about their daily lives is one of the holy grails of affective computing. This vision was largely a fantasy in the early days of the field (roughly 1995–2010), when researchers were grappling with foundational theoretical (e.g., what is an emotion?), technical (e.g., how to measure facial expressions?), and methodological (e.g., how to represent ground truth?) issues.

Consequently, early affect detection systems mainly focused on detecting acted affect (i.e., posed expressions), which was followed by the efforts to detect spontaneous affect (i.e., nonacted but elicited in response to a stimulus). This research was mainly conducted in the lab but would occasionally occur in quasi (e.g., YouTube videos) or restricted (e.g., classrooms) real-world settings.[1] Smartphones and wearable sensing ushered forth by fitness trackers changed everything. Suddenly, researchers were able to record various the aspects of human behavior and physiology (e.g., heart rate, activity, locations visited, phone use, and social media use) as it unfolded in the real-world and across a variety of contexts, which we refer to as the "real-wild." When combined with cost-effective computing and advances in deep learning, the vision of real-world affect detection from wearables was suddenly within reach. Accordingly, the past decade (2010 and beyond) has yielded numerous efforts toward fully automated affect detection in the real-wild. Literature surveys suggest impressive accuracies, such as 65%–97%,[2] 60%–99%,[3] and 78%–97%[4] for stress detection and 65%–81% for other affective states (anxiety, positive affect, etc).

Notwithstanding that a lack of standardized approaches to validate systems and report results complicates independent verification and comparison, the promising results have garnered significant attention beyond the affective computing community—who would not be excited by a 97% accuracy of stress detection? For example, the smart health community is interested in being able to track a person's emotions because this can have profound implications for diagnosis and treatment of numerous mental health

1541-1672 © 2023 IEEE
Digital Object Identifier 10.1109/MIS.2022.3221854
Date of current version 13 February 2023.

conditions including depression, anxiety, and bipolar disorder. Whereas wearable devices can provide continuous monitoring of physiological signals, converting these raw values into emotion estimates is a game changer. Similarly, industrial and organizational psychologists, who study the methods to improve occupational outcomes (e.g., decreasing absenteeism and improving task performance) are keenly interested in automatically tracking stress and early warning indicators of burnout, and then there is the military, where automatic measurement of workload, trust, and other affect-related constructs are important components for next-generation teams of humans and intelligent machines.

But a nagging issue persists in the midst of this enthusiasm. The ability to detect a complex psychological construct like affect from commodity sensors as people engage in everyday activities (e.g., working, sleeping, leisure, commuting) without restriction (i.e., people are moving, dancing, laughing), in varied physical (e.g., while skiing, meditating, dancing) and social (e.g., along, with friends, work colleagues) contexts with high (or even moderate) accuracies seems too good to be true. It is also inconsistent with psychological research questioning the strength of the link between expressing and experiencing affect and on the influence of social, environmental, and cultural factors on affective states.[5] The major inconsistency between the promising published results given the immense complexity of the problem leads us to ask whether affect detection from wearables in the real-wild is fact or fantasy or somewhere inbetween?

In 2017, the U.S. Intelligent Advanced Research Project Agency (IARPA) provided a unique opportunity to address this question. IARPA's Multimodal Objective Sensing to Assess Individuals with Context (MOSAIC) program aimed to "to develop and validate unobtrusive, passive, and persistent sensor-based methods to assess stable and dynamic psychological, cognitive, and physiological aspects of an individual."[a] In addition to accuracy, which is the main performance measure used in the field, MOSAIC emphasized robustness (estimates/predictions had to be provided even with noisy/missing data) and generalizability (modeling approaches had to be user-independent and reflect real-world experiences).

How did the affect detection systems fare when put to this rigorous test? The short answer—not very well—indeed, none of the three teams came close to meeting the target metrics. Even more concerning, the results were null (zero) in several cases. In an attempt to reconcile these sobering results with the aforementioned past successes, we, who were performers on two separate teams, reflect on our experiences by asking: what worked, what went wrong, why, and where do we go from here?

[a]p.6 of the request for proposals available at https://www.iarpa.gov/research-programs/mosaic and https://osf.io/ax6yg/.

OVERVIEW OF MOSAIC

MOSAIC was structured such that participating teams collected their own data using different suites of sensors and modeling approaches. However, the evaluation methods and metrics were standardized and conducted by an independent testing and evaluation team.

Key Aspects of the MOSAIC Challenge

Participants and Context

Because IARPA focuses on the intelligence community, the participants had to be employed in occupations that resemble the demands on the intelligence workforce. For this reason, relying on convenience samples (i.e., students/faculty) was not permitted; this component itself reflects a major deviation from past affect detection studies. Further, data collection had to occur as participants engaged in their normal, everyday routines, which in prepandemic times entailed commuting into the office (though remote work was also permitted) and work-related travel.

Constructs and Ground Truth

The constructs to be measured included physical health, mental health and well-being, intelligence, personality, and job performance. Here, we focus on the measurement of four affect-related constructs: positive and negative affect, stress, and anxiety. All "ground truth" measures consisted of validated self-report questionnaires. Each construct was assessed as both a stable "trait" once at the start of the study and also as a contextually varying "state" once per day (including weekends) for at least two months. Daily measurement frequency varied by construct, but affect was measured once per day at some predetermined time (e.g., at either 8am, 12pm, or 4pm local time) using ecological momentary assessments or EMAs (i.e., participants received a text to complete a 3–5 min survey). Positive and negative affect were measured with the 60-item PANAS-X (trait) and 10-item PANAS-Short (state) measure. Anxiety was measured with the 20-item STAI (trait) and with single omnibus item (state). Stress was only measured as a state, also with a single omnibus item (i.e., "Overall,

how would you rate your current level of stress?"). Details on scoring and assessment are discussed in MITRE Corporation's report.[6]

Modeling Constraints and Assessing Performance

Ground-truth data from a subset (20%–40%) of participants was withheld (i.e., blinded) from the teams either pseudorandomly or from a separate cohort. Teams could train and internally validate their models on the nonblinded data. They submitted predictions on the blinded data, which were used to assess performance.

Teams could use any modeling approach but could only rely on automatically sensed information to generate predictions. Even demographic information could not be used unless it was automatically detected and location coordinates were not permitted. These criteria, along with the blinding, were established to assess generalizability to new (unseen) participants. A prediction was scored if there was a corresponding ground-truth measure irrespective of whether any sensor data were recorded. This is an important component of robustness. Further, all code and data were independently verified by the testing and evaluation team.

Scoring focused on predicting between-individual differences (trait measures) and within-individual differences (state/daily measures). For trait measures, the target metric was a correlation of 0.5 or higher between sensor-based predicted and the self-reported ground-truth score. Scoring for state measures was a bit more involved,[6] but essentially the target was an R^2 (proportion of variance explained) of 0.25, which corresponds to a correlation of 0.5. The 0.5 metric corresponds to a Cohen's d of about 1.2 sigma (a "large" effect) or an area under the curve (AUC) of 0.8.

Overview of Team Tesserae and the Tracking Individual Performance With Sensors (TILES) Team

We discuss the methods and results of two of the three participating teams called Tesserae[b] and TILES.[c] Both were multidisciplinary, multiorganizational teams encompassing more than 30 individuals each.

Team Tesserae

Key ideas of team Tesserae were to: 1) collect a large, geographically diverse dataset over an entire year to improve generalizability and understand seasonal effects; 2) jointly model physiology, behavior, social interactions, and context by leveraging sensors that people already use; and 3) develop novel computational approaches to robustly integrate heterogeneous data streams. Accordingly, the Tesserae team collected longitudinal, year-long data from 757 information workers (e.g., engineers, consultants, managers) from five cohorts distributed across the United States. The sensors included a wearable fitness tracker (Garmin Vivosmart 3.0 to collect physical activity, sleep, and heart rate), a smartphone application (to collect communication metadata [not content]), Bluetooth beacons (to track relative location), and social media (Facebook posts). Modeling approaches ranged from top-down methods (i.e., theoretically driven features (e.g., time spent commuting) and standard machine learning [Random Forests]) to more bottom-up approaches including higher order networks and sequential deep learning. An ensemble approach, where models were trained/optimized on individual modalities (and combinations thereof) and selectively deployed based on available sensor data, was used to address missing data; see Robles-Granda et al.'s work[7] for details.

Team TILES

The key aims of the TILES project were to: 1) collect data from a working demographic, which experiences high levels of stress, fatigue, and burnout; 2) jointly model physiology, behavior, social interactions, and context using commercially available and unobtrusive sensing technologies; and 3) develop novel multimodal modeling techniques for uncovering the main factors contributing to daily changes in well-being. The TILES team gathered 10 weeks of sensor data from 212 hospital workers (e.g., nurses, technicians, therapists) working in different units (e.g., intensive care, step-down). The suite of passive and wearable sensors included a wrist-worn fitness tracker (Fitbit Charge 2 to gather physical activity, sleep, and heart rate), a fitness garment (OMSignal shirt for high-fidelity heart rate, breathing rate, body movement), a portable vocal audio tracker (Unihertz Jelly Pro phone to capture personal speech patterns [not content]), Bluetooth hubs and beacons (to track relative location, ambient temperature, humidity, light), and a smartphone application (to collect social media). A range of modeling approaches were investigated, including top-down traditional and deep machine learning as well as bottom-up motif analysis, signal-aware sequential data imputation, and low-level featurization; see Yan et al.'s work[8] for details.

[b][Online]. Available: https://tesserae.nd.edu/
[c][Online]. Available: https://tiles-data.isi.edu/, https://sail.usc.edu/tiles/

TABLE 1. Results from the MOSAIC challenge on blinded set.

	Trait (Criteria: r of 0.5)		State (Criteria: R^2 of 0.25)	
Construct	**Tesserae**	**TILES**	**Tesserae**	**TILES**
Positive affect	.16	$< .01$	$< .01$	.01
Negative affect	$< .01$	.14	$< .01$	$< .01$
Anxiety	.14	.13	$< .01$	$< .01$
Stress	–	–	.01	.03

Summary of Results

The results (see Table 1) support three main conclusions. First, neither team met the program metrics for either trait (r of 0.5) or state (R^2 of 0.25) affect detection. Second, performance for trait detection was higher than that for state detection. Third, state detection accuracies were essentially zero with the exception of stress, where automated methods explained 1%–3% of the variance in self-reported daily stress. Further, the accuracies reported in Table 1 were representative of the other nonaffective constructs. Specifically, Tesserae achieved a mean r of 0.14 (SD $=$ 0.12) across 14 traits and a mean R^2 of 1% (SD $=$ 3%) for 17 states. Equivalent results for TILES were a mean of. 10 (SD $=$ 0.17) for traits and a mean R^2 of 1% (SD $=$ 4%) for states.

POSTMORTEM: SIX CHALLENGES

Whereas it was unsurprising that none of the teams met the aspirational program metrics, the null results for affective state detection were especially concerning. The following challenges arose in response to a request from the government (IARPA) to opine as to why the program metrics were not achieved. Here, we focus on affective state detection since this is the primary focus of the community and where results were the lowest.

Challenge 1: The Mythical Experience–Expression Link

Affect detection has historically been rooted on a myth that there are exist robust and generalizable mappings between affective expression (e.g., a big smile) and experience (e.g., feeling happy). Instead, research indicates that the expression–experience link is weak and modulated by numerous factors (e.g., context, culture, individual traits).[5] Thus outside of carefully controlled, homogeneous, lab studies, a more realistic expectation is that the link is "above-chance probabilistic"—i.e., better than guessing.[9] Unfortunately, this myth appears to be persistent and is now embodied in commercial products, with occasional caveats in the fine print.[d]

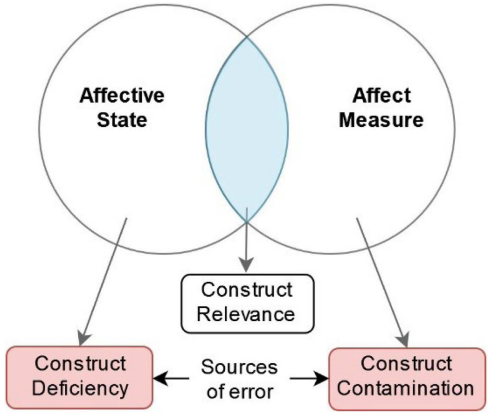

FIGURE 1. Sources of error in affective ground truth.

Challenge 2: Deficiency in Ground-Truth Measurement

MOSAIC, like many other real-world affect detection studies, used self-reports as the sole ground-truth measure. Whereas self-reports are often eschewed as being subjective, and consequently not reliable and valid, this is a major misconception, since the field of psychometrics has demonstrated that despite being subjected to several biases, self-reports can yield reliable and valid data. The issue is that a singular measurement instrument (e.g., a self-report) is inadequate to measure a complex construct. As noted in Figure 1, construct deficiency occurs when a measure only targets a subset of the construct (e.g., self-reports cannot access subconscious information), whereas construct contamination occurs when a measure targets irrelevant information (e.g., self-reports can be subject to social desirability bias). The obvious solution is to incorporate a diversity of methods (e.g., observer/informant reports/annotations, biomarkers, such as cortisol) to maximize capture of construct relevant variance. However, this is difficult to implement

[d]Amazon Rekognition documentation states: "Note that a prediction of an emotional expression is based on the physical appearance of a person's face only. It is not indicative of a person's internal emotional state, and Rekognition should not be used to make such a determination" https://docs.aws.amazon.com/rekognition/latest/dg/what-is.html(retrieved 8/23/22). Affectiva makes no such caveat, instead maintaining that "Emotion recognition is completed in iMotions using Affectiva, which uses the collection of certain action units to provide information about which emotion is being displayed." https://imotions.com/blog/facial-action-coding-system/ (retrieved 8/23/22).

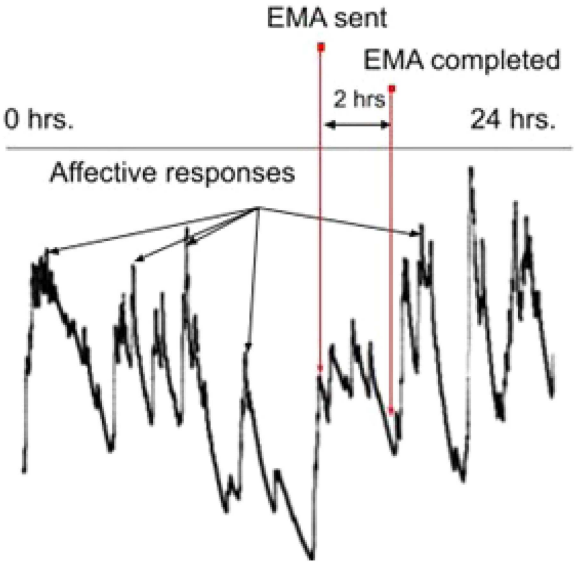

FIGURE 2. Temporal granularity and temporal misalignment between affective responses and measurement of ground truth via ecologically momentary assessments (EMA).

in the wild especially for large-scale longitudinal studies as in the MOSAIC program.

Challenge 3: Temporal Granularity and Temporal Misalignment of Ground Truth

Affective computing methods require fine-grained ground-truth annotations to be precisely aligned to the sensed signals, which is common for studies in the lab or quasi-wild, where measurement frequencies range from milliseconds (frame-level annotation) to a few minutes. However, ground-truth sampling rates in MOSAIC were coarse grained (1/day) compared to the sensors (about 1/sec), so as to not be disruptive to busy individuals where achieving a modicum of EMA compliance was a major concern. A related problem is that of misalignment in that there can be significant delays between experiencing an emotional episode and reporting it (e.g., participants are handling a stressful event rather than responding to the EMA). For example, there was a median 2-hour delay between the onset of an EMA and the subsequent response in the Tesserae data. In general, there is the challenge of achieving precise temporal alignment between the sensor data, onset of the affective events, and collection of ground truth for studies in the real-wild (see Figure 2).

Challenge 4: Low-Intensity Affective Responses

High-intensity affective responses can be elicited in the lab, for example, by inducing pain from heat or stress via public speaking. In contrast, affective responses in nonclinical samples largely consist of low-intensity baseline affect (e.g., neutral, mild relaxation, or mild anxiety), which are occasionally punctuated by strong emotional responses to events/triggers. For example, participants in the Tesserae study reported considerable stress (i.e., 4 or 5 on a 1–5 scale) only 5% of the time. This yielded limited training samples of high-intensity responses and complicated machine learning due to the class imbalance problem.

Challenge 5: Low-Fidelity Commodity Sensors

Scalable long-term affect detection entails using commodity sensors that people already use, a key principle of Tesserae and somewhat of TILES, which did include a higher fidelity physiological sensor (OMSignal shirt) but was still subjected to considerable motion artefacts.[10] The convenience of ambulatory monitoring with wearable sensors incurs a tradeoff with respect to the fidelity of the sensors, such as which components of physiology can be sensed, sampling rate, susceptibility to motion artefacts, and other factors. To this point, a recent survey[11] of 18 studies comparing wearable photoplethysmography (PPG) with gold-standard electrocardiography (ECG) to measure heart rate variability (HRV) found excellent alignment between the signals at rest, but progressively declining correlations as activity increased. Thus, a major caveat is that the sensors were of lower fidelity than the research-grade sensors used in the lab.

Challenge 6: Heterogeneity of Contexts

Context can be broadly defined as the physical and social environment surrounding a measurement. Unlike studies in the lab or quasi-wild, data collection unfolded across a range of heterogeneous contexts. Indeed, the MOSA(IC) program aimed for assessments of "individuals with context." However, ground-truth data on a person in their real-world context was only collected once a day, making it difficult to determine how to integrate context into the models. Thus, like much of affect detection research, both the Tesserae and TILES teams adopted context-general approaches (i.e., a single model was trained across all contexts), which may be a fatal design decision if the affect expression–experience link is context dependent as most emotion theories would suggest.[5,9] This might also explain why results were higher for trait assessment, which aggregates across contexts than the context-dependent state assessments.

IMPLICATIONS: THREE CONJECTURES

We discuss implications of the aforementioned challenges as three conjectures—opinions based on incomplete information.

Conjecture 1. Affect Detection in the Lab and "quasi" Wild Might Be a Different Problem Than Detection in the "real" Wild

Two salient aspects of current affect detection research make it particularly alien from affect detection in the real-wild. First, current work occurs in homogeneous contexts—i.e., it exists within a particular configuration of time, space, and environment (e.g., undergraduates silently viewing 20 minutes of videos designed to elicit sadness in a lab). This is irrespective of whether affect is acted, experimentally elicited, or occurs naturally, and it also applies when experimental control is relaxed as in the quasi-wild. For example, detecting affect from the diverse videos of automotive reviews (i.e, the MuSE-CaR dataset[12]) reflects an expanded but still homogeneous context.[e] Second, most current affect detection approaches require the underlying signals (video, audio, physiology, etc.) to be aligned with fine-grained, temporally precise annotations (i.e., ground-truth affect; e.g., annotating each frame in a video or collecting self-reports of affect every 15 secs).

Conversely, affect detection in the real-wild must operate across heterogeneous contexts, which include multiple activities (work, rest, leisure, housework), locations (home, office, etc), social interactions (along, peers, friends, family, etc.), and timescales with unique rhythms (e.g., diurnal cycles, seasonal effects). It must also handle coarse-grained, misaligned annotations because it is implausible to expect people to self-report affect every few minutes or to have observers provide fine-grained annotations without resorting to mass surveillance. Thus, research in the real-wild must contend with challenges posed by heterogeneous contexts and temporal granularity/misalignment of annotations and signals, two features, which do not pose major complications in the lab or quasi-wild.

Conjecture 2. Published Results on Wearable Sensing in the Wild Might Not Reflect Robust, Generalizable Performance

Why were the current results completely at odds with studies reporting impressive accuracies for affect detection from wearables in the wild (60%–99%[2,3,4])? One possibility might be factors specific to the MOSAIC program, such as the target populations, the infrequent (1/day) and exclusive use of self-reports to measure the ground truth, and the specific survey instruments themselves. Further, the expedited timescale of 17–20 months from inception, data collection, modeling, to evaluation might not have promoted a creative, discovery-oriented approach, instead requiring teams to rapidly adapt and apply existing affect detection methods, which might have been ill suited for the real-wild (see Conjecture 1). We must also acknowledge that our teams might not have been sufficiently skilled, and other teams would have been more successful (although similar results were obtained by a third team[f]).

Alternatively, there might be a cause to question the veracity of the impressive accuracies reported in published studies on affect detection from wearables in the wild. As we have recently argued,[13] there is a tendency to simplify the problem to optimize accuracy at the expense of robustness and generalizability. Table S1[g] tabulates a set of design decisions from published studies that pose threats to robustness and generalizability. Briefly, these include: 1) quantizing continuously measured affect into discrete low versus high categories, while disregarding the more difficult medium category; 2) avoiding the data imbalance problem by balancing class labels (including testing data); 3) discarding missing data by only generating estimates (predictions) for cases with high sensing fidelity; 4) overfitting due to a lack of strict person-level independence in training and testing sets; 5) reporting accuracy metrics, which do not adjust for baseline performance (i.e., when there is class imbalance) or not considering counterfactual comparison models (e.g., shuffling labels); and 6) adopting arbitrary criteria for several decisions including cutoffs used for quantization, number of folds, treatment of missing data, and so on. To be clear, we are not implying any nefarious intent, as we have also published studies that are susceptible to these threats. Instead, we suggest that the field values/rewards the false idol of accuracy at the expense of robustness and generalizability.

Conjecture 3. Expectations for High Accuracy in Low Signal to Noise Conditions Might Be Implausible in the Real-Wild

It is worth considering why the field expects high or even moderate accuracies in the real-wild. We argue that these expectations arise from a tendency to overly extrapolate findings from so called "biomarkers" (measurable indicators) of mental states and from conflating

[e]The approach adopted by some commercial vendors of scraping the web for large volumes of data on affect expressions (e.g., Google images) does involve heterogeneous contexts, but it is questionable as to whether it actually entails affect detection because there is no evidence that actual emotions are involved (annotating happiness from smiling faces does not mean the person is happy—see footnote 4).

[f]We cannot disclose specifics due to confidentiality requirements since IARPA does not make the results public (personal communication 01/12/2022).

[g]Supplementary materials available at https://osf.io/ax6yg/.

a statistically significant effect with the size of the effect. To illustrate, consider the strength of the biologically plausible and empirically supported inverse relationship between HRV and stress.[14] A meta-analysis[14] of 43 studies comparing differences in HRV for individuals diagnosed with posttraumatic stress disorder and healthy controls at baseline (i.e., without a stressor) revealed effects (i.e., |Hedges' g|) ranging from .23 to .66 (depending on HRV measure). The average $|g|$ of .43 corresponds to an R^2 of 4.4%, which can be considered a medium-sized effect (i.e., Cohen's d around 0.43). The vast majority of studies investigating this relationship occurred in controlled lab conditions (e.g., 77% in Schneider and Schwerdtfeger's work[14]) using research-grade sensing (ECG) while restricting movement. These study constraints increased the signal to noise ratio (SNR), yet effects were still moderate (i.e., HRV explains $<$ 5% of the variance).

Conversely, many factors inherent to ambulatory (*in situ*) real-world studies on the HRV-stress relationship diminish SNR, (see review in Martinez et al.'s work[15]), such as the use of commodity sensing (e.g., PPG), which have lower accuracy (see Challenge 5), unrestrained movement, lack of clearly defined/measurable stressors, and lower intensity responses, which cannot be precisely aligned with the onset of the stressor (see Challenges 3 and 4). Thus, expectations of accuracy must be calibrated with respect to the SNR ratio, with lab studies involving biomarkers providing upper bounds. To this point, regression models predicting self-reported stress from several HRV measures in the Tesserae study yielded an R^2 of about 1%,[15] a small effect (Cohen's d of 0.2) and only 25% of the above average meta-analytic effect of 4.4%. Similar results have been reported for facial expressions,[5] where data are lacking on other bimarkers, such as speech and body movements.

WAY FORWARD: FIVE SUGGESTIONS

We end with some suggestions for the way forward.

Suggestion 1. Embrace the Potential of Wearable Sensors

At the risk of throwing the proverbial baby out with the bathwater, we emphasize that wearable sensors are a game changer because they enable the study of human behavior *in situ*. Although we have argued that these sensors have yet to demonstrate their potential for affect detection in the real-wild, the challenges are not exclusive to the sensors themselves, but are more systematic of the complexity of the problem. Beyond affect detection, wearable sensors can provide insights into human behavior and experiences as it unfolds in the real-world, finally enabling an escape from the confines of the lab.

TABLE 2. Two-dimensional framework for affect detection studies.

	Environmental Realism		
Contextual Variability	*Lab*	*Quasi-wild*	*Real-world*
Homogeneous Contexts	e.g., eliciting stress in the lab	e.g., annotating stress in videos of public speaking events	e.g., naturalistic stress while taking standardized tests
Heterogeneous Contexts	e.g., eliciting stress via multiple stimuli (memory recall, social interactions, etc.)	e.g., annotating stress from videos of stressful events (public speaking, sporting events, etc)	e.g., naturalistic stress during unconstrained daily activities (home, work, commute, sport, etc)

Suggestion 2. Focus on Heterogeneous Contexts

Researchers tend to focus on the lab-to-real-world continuum but overlook whether the underlying context is homogeneous or heterogeneous. Table 2 provides a 2×3 framework to integrate both dimensions. Whereas there are an abundance of lab studies and some quasi-wild studies in homogeneous contexts, but heterogeneous contexts are rarely considered. We suggest that affect detection research in heterogeneous contexts, but in controlled settings of the lab or quasi-wild, might provide stepping stones toward affect detection in the real-wild (i.e., heterogeneous contexts in the real-world).

Suggestion 3: Recognize That We can not Simply "Deep Learn" a Solution

Modern affect detection systems have harnessed the power of deep learning with some success (e.g., Majumder et al.'s work[16]). Although there is usually insufficient data for end-to-end training, fine tuning pretrained models is a promising approach. However, the major successes of deep learning in object recognition and language understanding might not be replicated for affect detection, which focuses on ill-defined conceptual entities (feelings and emotions) rather than well-defined physical attributes (e.g., object detection and speech recognition). Thus, in addition to improvements in deep learning methods, we need complementary advances in how data are

collected and annotated and an increased scientific understanding of emotion expression and experience to achieve breakthrough results on affect detection in the real-wild.

Suggestion 4: Leverage Alternate Methods to Collect Ground Truth

There is a need for alternate approaches to collect ground truth in cases where affective responses are muted and there are a limited number of opportunities for self-reports via EMAs. Potential strategies include triggering EMAs based on the sensed signals (e.g., when heart rate is elevated), stratifying EMAs based on automatically sensed context, scheduling EMAs to align with specific affect-elicitation events, and adopting human-in-the-loop machine learning techniques (e.g., active learning). EMAs can also be complemented by alternate methods, such as day reconstruction, where participants use structured questionnaires to reconstruct activities and experiences of the previous day.

Suggestion 5: Adopt a Multidimensional Value and Reward System

In addition to the current overemphasis on accuracy, the field should also consider robustness and generalizability in its value and reward structures, and given that deep learning methods are increasingly "black-box," explainability and bias/fairness should also be important considerations. A binary categorization (i.e., low versus high) of these five factors yields 32 combinations, and a given affect detection system can be evaluated with respect to this multidimensional space. Researchers can also develop validity arguments—systematic evidence-based arguments on the validity of an assessment tool for a particular context. This is an important first step to change the conversation from how accurate? to how valid for what purpose?

CONCLUDING REMARKS

The MOSAIC program provided the inspiration to think big and the means to do so. It resulted in integration of large multiorganizational, multidisciplinary research teams, collection of massive longitudinal, in-the-wild datasets from working professionals (which are available for research purposes—see footnotes[b,c]), new ways to integrate multimodal sensing streams, and numerous scientific findings about human behavior, cognition, emotion, and social interactions in real-world contexts. By these standards, the program was a resounding success, despite none of the teams achieving the deliberately challenging program metrics. So, where do we go from here? Whereas it is tempting to disregard the poor affective state detection results by adopting the position that the problem of assessing self-reported affect from sensor streams was ill-defined to begin with, there is also an opportunity for reflection. More broadly, do the lack luster results reported here, mounting critiques of affect detection from affect scientists,[5] poor performance of commercial systems for nonposed expressions,[17] and philosophical debates on the feasibility and ethics of affect detection[18] suggest a looming crisis for the field?

A parallel can be drawn to the replication crisis in the psychological sciences, where more than 60% of high-impact studies failed to replicate.[19] Although there are several debates as to the extent of the crisis, the general consensus was that there was a problem with the status quo. This resulted confronting several methodological shortcomings and adopting reforms aimed at developing a more rigorous science (e.g., the new statistics.[20] But methodological reforms can only go so far—values and reward structures need to be re-examined. In the psychological sciences, statistical significance (i.e., detecting an effect) at the expense of robustness and generalizability were rewarded because null findings were mostly unpublishable. Similarly, affect detection values accuracy and technical novelty, at the expense of robustness, generalizability, explainability, and bias/fairness. As the field of affective computing turns 30 years old, it might also benefit from reflection and reformation so as to come closer to realizing its awesome promise and potential.

ACKNOWLEDGMENTS

This research was supported in part by the Office of the Intelligence Advanced Research Projects Activity (IARPA; under Grant 2017-17042800007), and in part by the National Science Foundation (NSF; under Grant SES 2030599 and Grant SES 1928612). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of IARPA, NSF, or the U.S. Government.

REFERENCES

1. S. K. D'Mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, 2015, pp. 1–36.
2. P. Schmidt et al., "Wearable-based affect recognition–A review," *Sensors*, vol. 19, no. 19, 2019, Art. no. 4079.

3. Y. S. Can, B. Arnrich, and C. Ersoy, "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey," *J. Biomed. Inform.*, vol. 92, 2019, Art. no. 103139.
4. N. Long et al., "A scoping review on monitoring mental health using smart wearable devices," *Math. Biosci. Eng.*, vol. 19, no. 8, pp. 7899–7919, 2022.
5. L. F. Barrett et al., "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychol. Sci. Public Int.*, vol. 20, no. 1, pp. 1–68, 2019.
6. MITRE Corporation, "Multimodal objective sensing to assess individuals with context (MOSAIC): Testing and evaluation procedures guide phase 1," Tech. Rep., 2017.
7. P. Robles-Granda et al., "Jointly predicting job performance, personality, cognitive ability, affect, and well-being," *IEEE Comput. Intell. Mag.*, vol. 16, no. 2, pp. 46–61, May 2021.
8. S. Yan, H. Hosseinmardi, H. -T. Kao, S. Narayanan, K. Lerman, and E. Ferrara, "Estimating individualized daily self-reported affect with wearable sensors," in *Proc. IEEE Int. Conf. Healthcare Inform.*, 2019, pp. 1–9.
9. S. D'Mello, A. Kappas, and J. Gratch, "The affective computing approach to affect measurement," *Emotion Rev.*, vol. 10, no. 2, pp. 174–183, 2018.
10. K. Mundnich et al., "TILES-2018, A longitudinal physiologic and behavioral data set of hospital workers," *Sci. Data*, vol. 7, no. 1, pp. 1–26, 2020.
11. K. Georgiou et al. "Can wearable devices accurately measure heart rate variability? A. systematic review," *Folia Medica*, vol. 60, no. 1, pp. 7–20, 2018.
12. L. Stappen, A. Baird, E. Cambria, and B. W. Schuller, "Sentiment analysis and topic recognition in video transcriptions," *IEEE Intell. Syst.*, vol. 36, no. 2, pp. 88–95, Mar./Apr. 2021.
13. B. M. Booth, H. Vrzakova, S. M. Mattingly, G. J. Martinez, L. Faust, and S. K. D'Mello, "Toward robust stress prediction in the age of wearables: Modeling perceived stress in a longitudinal study with information workers," *IEEE Trans. Affect. Comput.*, 2022, to be published, doi: 10.1109/TAFFC.2022.3188006.
14. M. Schneider and A. Schwerdtfeger, "Autonomic dysfunction in posttraumatic stress disorder indexed by heart rate variability: A meta-analysis," *Psychol. Med.*, vol. 50, no. 12, pp. 1937–1948, 2020.
15. G. J. Martinez et al., "Alignment between heart rate variability from fitness trackers and perceived stress: Perspectives from a large-scale in situ longitudinal study of information workers," *JMIR Hum. Factors*, vol. 9, no. 3, 2022, Art. no. e33754.
16. N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intell. Syst.*, vol. 32, no. 2, pp. 74–79, Mar./Apr. 2017.
17. D. Dupré et al., "A performance comparison of eight commercially available automatic classifiers for facial affect recognition," *PLoS One*, vol. 15, no. 4, 2020, Art. no. e0231968.
18. K. Crawford et al., "AI now report," AI Now Institute, New York, NY, USA, 2019.
19. P. E. Shrout et al., "Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis," *Annu. Rev. Psychol.*, vol. 69, no. 1, pp. 487–510, 2018.
20. G. Cumming, "The new statistics: Why and how," *Psychol. Sci.*, vol. 25, no. 1, pp. 7–29, 2014.

SIDNEY K. D'MELLO is a professor in the Institute of Cognitive Science and Department of Computer Science, University of Colorado Boulder. His research focuses on applying multimodal machine learning to investigate the interplay between the cognitive and affective states of individuals and teams engaged in real-world activities. He is a corresponding author of this article. Contact him at sidney.dmello@colorado.edu.

BRANDON M. BOOTH is a research scientist with the Institute of Cognitive Science, University of Colorado Boulder. He has a diverse industry background researching, publishing, and developing video games, serious games, robots, computer vision, human–computer interaction systems, and geospatiotemporal visualizers. His research focuses on using multimodal machine learning and mental state measurement techniques to model human perception, behavior, and experiences and developing algorithms to reduce the impact of inadvertent human biases and errors. Contact him at brandon.booth@colorado.edu.