Sub-Population Specific Models of Couples' Conflict

KRIT GUPTA, ADITYA GUJRAL, and THEODORA CHASPARI, Texas A&M University
ADELA C. TIMMONS, Florida International University
SOHYUN HAN, YEHSONG KIM, SARAH BARRETT, STASSJA SICHKO, and
GAYLA MARGOLIN, University of Southern California

Interpersonal conflict between couples is a significant source of stress with long-lasting effects on partners' physical and psychological health. Motivated by findings in psychological science, we study how couples with distinct relationship functioning characteristics experience conflict in real life. We propose sub-population specific machine learning models using hierarchical and adaptive learning frameworks to automatically detect interpersonal conflict through the ambulatory monitoring of couples' physiological signals, audio samples, and linguistic indices. Results indicate that the proposed models outperform a general model learned for the entire population and separate models independently trained on each sub-population, providing a foundation toward personalized health applications.

CCS Concepts: • **Computing methodologies** → **Supervised learning**; • **Theory of computation** → *Unsupervised learning and clustering*;

Additional Key Words and Phrases: Interpersonal conflict, sub-population specific models, multi-task learning, feedforward neural network, ambulatory monitoring

ACM Reference format:
Krit Gupta, Aditya Gujral, Theodora Chaspari, Adela C. Timmons, Sohyun Han, Yehsong Kim, Sarah Barrett, Stassja Sichko, and Gayla Margolin. 2020. Sub-Population Specific Models of Couples' Conflict. *ACM Trans. Internet Technol.* 20, 2, Article 9 (March 2020), 20 pages.
https://doi.org/10.1145/3372045

1 INTRODUCTION

Interpersonal conflict is known for its deleterious effects on personal and professional functioning. Conflict between co-workers can result in prolonged fatigue and poor general health and can negatively affect productivity [20, 32, 60]. Conflict in romantic relationships serves as a risk factor for partners' psychological and physical health problems and can result in decreased relationship satisfaction, as well as emotional and physical withdrawal [22, 56]. Understanding the causes, antecedents, and sequelae of interpersonal conflicts can therefore help to promote healthier and more fulfilling romantic relationships.

This work was partially supported by the National Science Foundation (NSF BCS-1627272).
Authors' addresses: K. Gupta, A. Gujral, and T. Chaspari, Texas A&M University, 710 Ross St., College Station, TX, 77843-0001; emails: {kritgupta, aditya.gujral, chaspari}@tamu.edu; A. C. Timmons, Florida International University, 1250 SW 108th Ave., Miami, FL, 33199-2516; email: atimmons@fiu.edu; S. Han, Y. Kim, S. Barrett, S. Sichko, and G. Margolin, University of Southern California, 3620 S. McClintock, Los Angeles, CA, 90089-1061; emails: {sohyunha, yehsongk, skbarret}@usc.edu, ssichko@gmail.com, margolin@usc.edu.

1533-5399/2020/03-ART9 $15.00
https://doi.org/10.1145/3372045

Recent advances in ambulatory technologies now allow the continuous monitoring of human behavior in real-life settings. Smart sensing devices can collect data around the clock with little energy expenditure, and yield high volumes of multimodal recordings. Such data can afford researchers valuable insights into a person's life by monitoring how various behaviors and feelings are elicited and manifested across time and under realistic conditions [24, 48, 63]. By monitoring and detecting the onset and evolution of psychological events of interest, researchers could intervene in real time and real life as these events develop or even before they occur [61].

In this era of highly voluminous and highly variable data, emerging advances in computational science can help translate raw sensor values obtained from ambulatory devices into behavioral markers related to health and well-being [26, 45]. A large body of work has focused on the development of signal processing and machine learning techniques for analyzing collected data and relating them to outcomes of interest. Applications of such approaches span various topics of interest within mental health, including treatment of depression, social anxiety, and bipolar disorders, with current results indicating the feasibility of automated data analysis techniques for detecting events of interest in real life [35, 49]. Despite these advances, several methodological limitations regarding the data labeling, integration of context, and the generalizability of automated systems have arisen from these studies [23, 61].

The large inter-individual variability across people is a major hurdle for current automated systems, interfering with adequate generalization to unseen test samples. Current approaches typically assume a uniform group of individuals when modeling human behavior [38]. However, human behavior is unique to each person, who expresses and experiences the same stressor event or emotional stimulus in different ways [7, 44]. These inter-individual differences are amplified in real life, since the unstructured nature of such settings can increase the variability and complexity of human expression. Indicatively, Levenson et al. found that distressed couples' interactions are more strongly linked to emotional reactivity compared to their non-distressed counterparts [41]. According to Campbell and Simpson, anxiously attached individuals perceive conflict in a more intense way compared to securely attached partners and depict high levels of distress [10]. Research on adolescents further suggests that insecurely attached individuals show high negative affect in interpersonal conflicts compared to securely attached partners [19]. In the light of these findings from psychological science, general machine learning models, which assume the same association between the input features and the outcome of interest for all people, are not always effective for quantifying human conflict, thereby indicating the need for alternative models that can better address these challenges by taking inter-individual differences into account.

Personalized computational models of conflict are one possible method to address inter-individual variability. Personalized models can achieve high predictive ability, since they are fine-tuned to each individual separately [4]. However, this approach has several limitations. Individualized machine learning models assume the presence of labeled data for each person, which might not always be the case. Therefore, they are not generalizable to unseen individuals. Moreover, low-frequency behaviors, such as conflict, impose additional constraints; obtaining both positive and negative samples for each individual is time-consuming and resource-intensive. An alternative method for modeling inter-individual variability is to cluster people into sub-populations based on individual characteristics relevant to the outcome of interest. Sub-population refers to a group of individuals with common characteristics, while sub-population specific machine learning systems refer to models that learn the most relevant features of the sub-population of interest and make the final decision based on the sub-population to which a test sample belongs [6, 38]. In this way, sub-population specific models can more reliably represent the outcome of interest compared to general machine learning models, while also being more generalizable compared to fully personalized models.

The current article focuses on building sub-population specific machine learning models for detecting interpersonal conflict between couples in real life using ambulatory speech and physiological recordings collected from wearable and mobile devices. We propose a hierarchical and an adaptive framework to learn the common characteristics across populations and utilize the population specific information to cater to each specific population separately. The hierarchical model is implemented using a feedforward neural network (FNN), trained in a multi-task learning (MTL) framework. According to this framework, the first layers of the FNN are shared across all individuals, while the latter sub-population specific layers are learned for each cluster. Adaptive learning is implemented with an FNN, initially trained on the entire dataset. The last layers of the FNN are separately fine-tuned for each sub-population, resulting in one final FNN per sub-population. The proposed hierarchical and adaptive learning approaches are compared to a general machine learning model trained on all participants. We test whether the proposed models outperform the aforementioned baseline to evaluate the benefits of integrating sub-population specific information for detecting interpersonal conflict. Quantitative analysis regarding the most conflict-relevant features for each sub-population is further provided and discussed in relation to findings from psychological studies.

2 PRIOR WORK

Sub-population specific models have been previously proposed to detect human outcomes of interest and are generally divided into signal-based and machine learning approaches. Signal-based approaches compare signals obtained during a baseline state to the ones collected during the events of interest. The amount of divergence from baseline is used as a probability of occurrence of the target phenomenon [30]. De Santos et al. proposed the extraction of signal trajectories across various tasks to quantify person-dependent deviations [21], while Zeevi et al. suggested augmenting the signal-based feature space with person-dependent characteristics [69]. Despite the intuitive and cost-effective nature of signal-based approaches, it is not always possible to collect baseline data for each individual. Also, data labels are not considered during the learning process; therefore, the outcome of interest might not always be reliable.

The following four major approaches have been proposed for the design of sub-population specific machine learning models: (a) models independently learning data from a group of participants (separate); (b) models whose initial components are catering to the entire population, while the later ones become specific to the sub-population specific information (hierarchical); (c) models involving a two-stage process, where the model parameters are initially estimated using all data samples and are then refined using sub-population specific information (adaptive); and (d) models utilizing only the subset of data that is most relevant to the sample of interest in order to make a final decision (ensemble).

According to the separate models approach, a pre-determined criterion, such as demographics or medical history, is used to divide the original population into different clusters, which are then used to separately train one model for each sub-population. For example, Koldijk et al. divided the original set of participants into two separate groups based on their body movements and facial expressions in order to detect stress in work environments [38]. Bertsimas et al. and Kallus further proposed the use of separate learners jointly trained based on a target optimization criterion [6, 34]. Research to date suggests that models trained on separate sub-population clusters tend to outperform models learned based on the entire population [6, 34, 38]. However, separate models might not be adequately generalizable, since they do not incorporate population data from the entire sample in the final decision-making process.

Hierarchical models are comprised of several layers of hierarchy. Information from the entire population is represented in the first layers of these models, while sub-population specific

knowledge is integrated in the later levels. To implement this model, a hierarchical MTL framework implemented with FNNs has been proposed for the detection of stress, mood, and happiness in real-life scenarios, as well as for the detection of interpersonal conflict between couples [29, 31, 62]. Jaques et al. and Taylor et al. used individual characteristics as clustering criteria, while Gujral et al. employed individual and relationship-specific criteria to cluster the original population.

Adaptive models leverage transfer learning techniques and are achieved via fine-tuning of general models. According to these, general models are being initially trained on the entire dataset and are refined using data samples from the sub-population of interest [17, 43, 54]. Although adaptive models are conceptually similar to hierarchical models, they tend to be less resource-intensive to train because only part of these models is re-adjusted based on each sub-population.

Ensemble methods make decisions for the outcome of interest using the sub-population of participants most relevant to the test sample. Previous studies have used this approach to detect social and physical activity indicating the superiority of ensemble methods compared to general and separate models [33, 39]. Despite their intuitive nature and encouraging results, ensemble approaches can underperform with a small number of data samples, given that the final decision is made from limited number of participants, compromising the generalizability of the system. Due to the limited number of data samples and highly unbalanced classes in our problem of interest, ensemble approaches will not be considered in this article.

Previous studies have further attempted to detect interpersonal conflict in real-life settings. In Refs [36] and [52], the authors used audio signals and video as an input feature to detect conflicts in political debates. Speech interruptions have been used as a reliable marker of conflict during group discussions [12, 28]. Speech signals from body-worn audio sensors have been employed to detect conflict in police-public interactions [40]. In contrast to the aforementioned studies, where interpersonal conflict is well-defined, conflict between romantic partners can be a complex and sometimes subtle event, affected by a variety of interpersonal and psychological factors. In our previous work, we have attempted to use general machine learning models and hierarchical approaches to detect couples' interpersonal conflict [29, 63].

The current article provides the following contributions to existing studies: (1) assuming that interpersonal conflict is expressed in various ways depending on partners' relationship functioning characteristics, the current article introduces the use of sub-population specific models for the automatic detection of conflict in real life; (2) statistical analysis provides insights into the most discriminative features with respect to the outcome of interest for each sub-population; (3) sub-population specific models implemented with hierarchical and adaptive approaches are evaluated and compared against each other and against general machine learning models learned on the entire population; (4) in addition to acoustic and linguistic indices, the current study further leverages physiological data, which provides access to the generative processes in the human body and can be indicative of the emotional arousal present during interpersonal conflict.

3 DATA DESCRIPTION

Our data come from the University of Southern California (USC) Couple Mobile Sensing Project [3] and include 87 couples aged between 18–25 years old. Participant recruitment focused on young adult couples to investigate how adverse experiences in childhood and adolescence relate to romantic relationships in young adulthood, a unique developmental stage during which dating partners begin to take on a more central role [25]. Additionally, relationships during young adulthood predict functioning in future marital relationships [11, 51], suggesting that identifying relationship patterns during young adulthood may have long-term implications. Although participants depicted a limited age range, they were ethnically/racially diverse, and in different stages of their academic/professional life and romantic relationship. Specifically, 27.0% of participants identified

as Caucasian, 25.9% Hispanic/Latino, 16.7% African American, 12.6% Asian, 13.2% multiracial, and 4.6% other. Approximately half of the participants were part-time or full-time students (51.1%), while the majority were employed at least part time (77%). Couples had been dating for 29.2 (±24.2) months on average and 43.7% of couples were cohabitating.

Prior to the beginning of data collection, participants were asked to complete the Quality of Marriage Index (QMI) [50] and Experiences in Close Relationships-Revised (ECR-R) [59] questionnaire. The QMI includes six questions capturing each partner's satisfaction in various areas of the relationship. The first five questions are related to partners' specific emotions of happiness, stability, and strength elicited by the relationship and will be referred to as "QMI-1." The last item of the questionnaire captures general relationship satisfaction and will be referred to as "QMI-2." The ECR-R questionnaire includes 36 items, scored separately to provide indices of individuals' anxiety and avoidance toward their partner. Relationship anxiety refers to feelings of fear and worry regarding the partner's love and support, while avoidance is related to sharing private thoughts and feelings, as well as to feelings of closeness in the relationship.

Participants in the study were lent smartphone and wearable devices collecting data from 9:00 a.m. till midnight for one day. During the data collection procedure, ecological momentary assessments (EMA) were administered hourly through a Nexus 5 smartphone and assessed couples' mood and quality of interactions (MQI). The detailed items of the EMA are listed in the Appendix. The smartphone device also continuously recorded GPS coordinates, as well as 3-minute audio samples for every 12 minutes. The Actiwave sensor [1] was placed on participants' chest to obtain an electrocardiogram (ECG) signal with a $32Hz$ sampling frequency. The Q sensor [55] recorded electodermal activity (EDA), wrist acceleration, and body temperature from participants' non-dominant wrist with a sampling frequency of $8Hz$.

Conflict labels were provided by each partner on an hourly basis through the EMAs. The final conflict label per couple for each hour was obtained if any of the two partners reported conflict during that time. The data include 117 conflict and 1,126 (90.5%) non-conflict samples.

4 METHODOLOGY

Human behavior is inherently complex and diverse, providing fundamental challenges to general machine learning models, which assume a homogeneous group of individuals and learn a common feature representation related to the outcome of interest for all participants. Conversely, due to data sparsity issues, it is not always feasible to obtain highly personalized models. The proposed work will examine how interpersonal conflict is manifested differently in couples with various levels of relationship satisfaction and attachment characteristics, and how emerging machine learning models can take this information into account to achieve more reliable decisions compared to general machine learning models. Section 4.1 will describe the acoustic, linguistic, physiological, and contextual indices of the feature space. Section 4.2 will provide the clustering criteria and methodology. Section 4.3 will outline the most discriminative features for each sub-population through statistical analysis. Section 4.4 will describe the proposed sub-population specific machine learning systems implemented in a hierarchical and adaptive framework. Finally, Sections 4.5 and 4.6 will provide the details on the experimental setting, including baselines to the proposed sub-population specific models, and the evaluation metrics used in this article.

4.1 Feature Extraction

Five different types of features, including acoustic, linguistic, physiological, contextual, and self-reported MQI were extracted on an hourly basis. A summary of the features is provided in Table 1.

Audio signals were pre-processed to identify the speech segments for each partner. Fundamental frequency and loudness were computed over 30msec frames. Mean, median, maximum, minimum,

standard deviation, and range of the aforementioned acoustic measures were then calculated over each hour, resulting in a total of 12 acoustic features per partner.

Manual transcriptions of the audio signals were employed to derive language features, which were computed using the Linguistic Inquiry and Word Count (LIWC) software [53]. LIWC calculates the degree to which various categories of words, related to different emotions, thinking styles, social concerns, and parts of speech, are used in a text. It compares each word in the text against a pre-defined dictionary. The dictionary identifies which words are associated with which psychologically-relevant categories. For the purposes of this research, we used the built-in LIWC 2015 dictionary. Language features for each partner included 39 measures of psychological constructs, such as positivity, negativity, anxiety, insight, swearing, and personal concerns; 24 linguistic indices, such as word count and personal pronouns; 7 features of personal concern, such as home- and work-related words; and 3 paralinguistic markers, such as fillers. These 66 features were computed for each partner, resulting in 132 language measures in total. A detailed enumeration of the language features is provided in Table 8.

EDA was pre-processed through a low-pass filter of 16 samples to remove high-frequency noise. Movement artifacts were automatically detected by fitting a predetermined knowledge-driven structure to the original EDA signal [16]. SCR detection was automatically performed through the LedaLab toolbox [5]. Mean skin conductance level, as well as the number, frequency, and amplitude of skin conductance responses were extracted for the EDA features. Two different thresholds of 0.01 and $0.02\mu S$ were used to quantify the EDA response. EDA synchrony was further employed in order to quantify the co-activation of the sympathetic nervous system between the two partners, a construct highly relevant to relationship connectedness and satisfaction [64]. The Sparse EDA Synchrony Measure (SESM) was computed as the similarity of EDA signals between the two partners using joint sparse representation techniques [15]. The objective of this approach was to jointly model two EDA signals as the linear combination of a set of common atoms selected from an EDA-specific dictionary. The dictionary includes 4,340 parametric atoms that yielded from different combinations of the recovery time a, rise time b, time scale s, and time shift t_0 parameters in the Bateman function $g_{Bateman}(t) = (e^{-a(st-t_0)} - e^{-b(st-t_0)})u(t-t_0)$ ($u(t)$ is the step function), used to simulate the steep rise and slow recovery of SCRs. Joint decomposition of the two signals based on the aforementioned dictionary was performed using the orthogonal matching pursuit (OMP) algorithm of analysis windows of 5 and 15 min. duration. Various numbers of selected atoms N were used for each window ($N = 5, 10, 15$ for 5 mins; $N = 15, 20, 25, 30$ for 15 mins) in order to model multiple resolutions in the EDA representation. SESM was computed as the inverse of the joint representation of the EDA signals from both partners based on the commonly selected atoms. Intuitively, if the EDA signals are similar to each other, the common atoms can reliably capture their structure, resulting in low representation error, and therefore high synchrony quantified through the SESM. An asynchronous version of the SESM was further computed in order to obtain an estimate of directionality between the EDA signals of the two partners. According to the asynchronous SESM, the atoms selected from the sparse decomposition of one's EDA signal from one partner were used to reconstruct the EDA signal of the second partner. Asynchronous SESM was computed as the reconstruction error of the second partner. A detailed description of the SESM indices and their validation can be found in Refs [14] and [15]. The BioSig toolbox was used to detect artifacts in the ECG signal using a histogram-based approach and detect the ECG beats [65]. Automatically detected artifacts were visually inspected and revised by human annotators. Time- and frequency-based ECG features were extracted, including average beats per minute, average R-R interval, as well as the very-low, low, and high frequency component (0–0.04 Hz, 0.04–0.15 Hz, and 0.15–0.4 Hz, respectively). In addition to mean body temperature, activity count was computed as the $l2$-norm of the 3-axis acceleration signals. These resulted in 48 physiological features

Table 1. Language Features for Conflict Classification

Category	Extracted features
Linguistic indices	number of total words, words longer than six letters, words in LIWC dictionary, function words, pronouns, personal pronouns, "I, We, You, He/She, They" pronouns, impersonal pronouns, articles, verbs, auxiliary verbs, past/present/future tense verbs, adverbs, prepositions, conjunctions, negations, quantifiers, numbers
Psychological constructs	social processes (family, friends, humans), affective processes (positive/negative emotion, anxiety, anger, sadness), cognitive processes (causation, discrepancy, tentative, certainty, inhibition, inclusive, exclusive), perceptual processes (see, hear, feel), biological processes (body, health, sexual, ingestion), relativity (motion, space, time), personal concerns (work, achievement, leisure, home, money, religion, death)
Paralinguistic indices	assent, non-fluencies, fillers

per partner, a detailed description of which is provided in Table 2. Similar physiological features have been employed in previous research studying affect and emotions using wearable devices [37, 47].

Contextual indices were further obtained from the EMA reports in order to integrate context to the signal-based measures. Contextual features included the hourly consumption of caffeine, alcohol, tobacco, and other drugs, as well as the duration of exercise, interaction with others, and driving within an hour. These data resulted in seven measures per person. Self-reported measures collected each hour from the EMAs for each partner included stress, happiness, sadness, nervousness, and anger.

Mean substitution was done for the missing features in the data, while the entire record was removed if the conflict label was not available.

4.2 Sub-Population Specific Clustering

The criteria used for sub-population clustering are motivated by previous studies in psychology indicating that conflict is experienced and expressed differently by insecure and anxiously attached partners compared to their counter-peers, as well as by partners with different levels of relationship satisfaction [8, 10, 19, 46]. Campbell and Simpson in Ref. [10] found that the perception of more anxiously attached individuals about conflicts with their partners was higher. They also found that these individuals had greater feelings of distress, which can be potentially depicted through partners' physiological, acoustic, and linguistic indices. Research on adolescents further suggests that individuals with an insecure attachment style depict more negative affect during inter-personal conflict compared to their counter-peers [19]. These findings from psychological studies suggest that individuals with distinct types of relationship functioning might depict distinct bio-behavioral patterns, as reflected by physiological, acoustic, and linguistic metrics, when experiencing and expressing conflict.

Taking this into account, clustering criteria included the two relationship satisfaction dimensions of the QMI questionnaire (QMI-1 and QMI-2), as well as the avoidance and anxiety measures from ECR-R (Section 3). These measures were used as the input of a K-means algorithm, which was applied to obtain the sub-population clusters. From a total of 87 couples, we performed

Table 2. Physiological Features for Conflict Classification

Signal	Extracted features
Electrodermal activity	Skin conductance level, skin conductance response (SCR) frequency (SCR thresholds = 0.01, 0.02 muS), # SCRs (SCR thresholds = 0.01, 0.02 muS), mean SCR amplitude (SCR thresholds = 0.01, 0.02 muS), symmetric sparse EDA synchrony measure (SESM) (analysis window = 5 min, #atoms = 5, 10, 15), symmetric SESM (analysis window = 15 min, #atoms = 15, 20, 25, 30), asymmetric SESM (analysis window = 5 min, #atoms = 5, 10, 15), asymmetric SESM (analysis window = 15 min, #atoms = 15, 20, 25, 30)
Electrocardiogram	Interbeat interval, mean heart rate (HR), standard deviation of HR, min/max HR, rate variability (HRV), HRV triangular index, mean R-R interval, standard deviation of R-R intervals, root mean square of successive R-R interval differences, number of adjacent NN intervals more than 50 ms apart, percentage of adjacent NN intervals more than 50 ms apart, triangular interpolation of normal-to-normal intervals, HRV peak frequency at very low frequency (VLF: 0–0.04 Hz), low frequency (LF: 0.04–0.15 Hz), and high frequency (HF: 0.15–0.4 Hz), HRV absolute power at VLF/LF/HF, HRV relative power at VLF/LF/HF, ratio between LF and HF power, total power, Shannon entropy
3-Axes Acceleration	$l2$-norm of 3-axes acceleration
Body temperature	mean body temperature

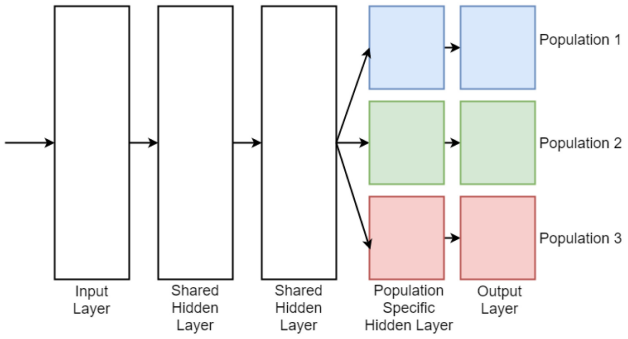
sub-population clustering based on the 123 participants for which all attachment and relationship satisfaction scores were available. Three clusters provided the best empirical tradeoff between the number of total samples and the resolution of each cluster.

4.3 Identifying Sub-population Specific Features of Conflict through Statistical Analysis

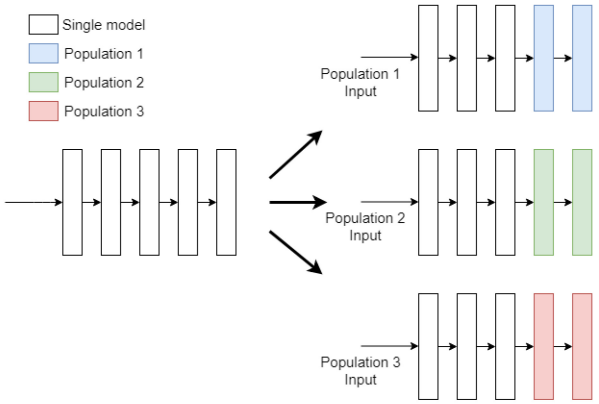
Hypothesis testing was performed to identify the most indicative features of the conflict outcome for each sub-population. An independent samples t-test was used to compare physiological, acoustic, and linguistic metrics obtained from conflict samples compared to the corresponding metrics from the non-conflict samples for each sub-population. Our working hypothesis is that different types of features will be most useful in detecting conflict for each sub-population (i.e., cluster of individuals).

4.4 Sub-population Specific Models of Conflict

Following evidence from previous literature (Section 3), and in an effort to obtain systems that conceptually make most sense for the data of interest, sub-population specific models of conflict have been designed using hierarchical and adaptive learning. Hierarchical models include multi-level representations, where the first levels include information for the entire population and the last levels capture information specific to each sub-population. Hierarchical models were implemented with an MTL FNN, whose first hidden layers are shared among all individuals and last output layers are split for each sub-population (Section 4.4.1). Adaptive approaches initialize a



(a) Hierarchical learning implemented with a multi-task feedforward neural network



(b) Adaptive learning implemented with a feedforward neural network whose last layers are fine-tuned for each sub-population

Fig. 1. Hierarchical and adaptive learning for sub-population specific machine learning models.

model using data from the entire population and refine the same model for each sub-population separately. Adaptive sub-population specific models are implemented by fine-tuning a FNN to each sub-population separately. This results in a final number of FNNs equal to the number of total sub-populations (Section 4.4.2). The main difference between hierarchical and adaptive models lies in the fact that knowledge is simultaneously learned for all sub-populations in the hierarchical models, while weights of the FNN are separately adapted for each sub-population in the adaptive approach.

4.4.1 Hierarchical Sub-Population Specific Models. MTL is inspired by human learning, where knowledge from one task is applied to obtain knowledge from another related task. MTL is used to learn signal-based representations common among all samples and refined for each sub-population. This method has been implemented with an FNN, whose initial layers capture the common knowledge for the entire population and whose later layers cater explicitly to the different populations. The proposed model has one input layer, three hidden layers, and one output layer. The number of layers was empirically determined to make sure that the parameters of the network were learned based on the total number of training samples. The first two hidden layers are shared among all individuals, while the last hidden and the output layer of the network are kept specific to each sub-population (Figure 1(a)). Hence, the model is jointly learning both the

Table 3. Set of Original Values for Hyper-Parameter Tuning

Hyper-parameter	Values
# neurons in hidden layers	60, 80, 120
dropout	0, 0.2, 0.3
optimization algorithm	adam, sgd, rmsprop
class weights	{non-conflict:1, conflict:15}, {non-conflict:1, conflict:25}

Nested leave-one-couple-out cross-validation is performed to identify the best combination based on the validation data using a grid search.

features inherent to the entire population, as well as the ones that are particularly relevant to each sub-population.

4.4.2 Adaptive Sub-Population Specific Models. The proposed adaptive learning approach included an FNN initially trained on all data samples and fine-tuned for each sub-population separately. Since the data samples per sub-population might not be enough to fine-tune all layers of the FNN, only the parameters of the last two layers of the FNN were adapted based on each sub-population. This resulted in three separate FNNs, one per sub-population. The FNN is initially learned based on all participants. Subsequently, the parameters of first two layers are kept frozen, while the parameters of the last two layers are fine-tuned for the sub-population of interest (Figure 1(b)). The main difference between the MTL and FNN fine-tuning frameworks lies in the data samples upon which the loss function is optimized. The loss function of the MTL includes all data samples, while the loss function of each of the fine-tuned FNNs includes the samples that belong to the corresponding sub-population.

4.5 Experimental Setting

The modular structure of neural networks allows for flexible representations of the input space. For this reason, neural networks with fully-connected layers were used as a basis for the proposed MTL and fine-tuning approaches, implementing the hierarchical and adaptive sub-population specific models, respectively. In order to ensure fair comparison of the proposed models, our first baseline consists of a fully-connected FNN trained on all couples without any adaptation, referred to as "Single". The FNN has the same number of layers (i.e., five) as the MTL and FNN fine-tuning models (Figure 1) to further make the systems as comparable as possible. A decision tree and a K-Nearest Neighbor were further tested as baselines. The results from these classifiers were similar to the single FNN; therefore, for the sake of brevity, they will not be further discussed in the paper. Our second baseline will include separate 5-layer FNNs trained for each sub-population independently, referred to as "Separate".

Hyper-parameter tuning for both the MTL, FNN fine-tuning, and baseline approaches was performed using a couple-independent 5-fold nested cross-validation [13]; samples from the same couple were not included in the training, validation, or test sets during the same fold. The outer fold of the nested cross-validation included the data from the test set, based on which the final classification metrics were reported. Similar to Calefato et al. [9], the inner fold of the nested cross-validation selected a sample of 20% of the data in a stratified manner as a validation set to tune the hyper-parameters of the MTL and FNN networks. The test set did not overlap at any point with the training and validation sets. Tuning was performed using a grid search with an initial set of hyper-parameters as shown in Table 3. Both MTL and FNN fine-tuning were implemented using the Keras toolbox [2]. Back-propagation was performed with a batch size of 32.

Table 4. Results of Hypothesis Testing for Identifying Significant Differences between the Presence and Absence of Conflict for Each Sub-Population with Respect to Skin Conductance Response (SCR) Frequency, R-R Interval, and Fundamental Frequency (F0)

Measure	Partner	Cluster 1	Cluster 2	Cluster 3
SCR Frequency	Male	t(73.8) = 0.21, p = 0.82	t(28.5) = 0.2, p = 0.82	**t(324.8) = 2.1, p = 0.03**
	Female	t(66.6) = −0.6, p = 0.49	**t(133.9) = 2.1, p = 0.03**	t(62.18) = 0.29, p = 0.77
R-R Interval	Male	t(80.5) = 1.4, p = 0.16	**t(135) = 4.1, p = 0.0**	**t(54.3) = 0.1, p = 0.98**
	Female	t(74.7) = 0.6, p = 0.52	**t(135) = 3.6, p = 0.0**	t(61.3) = 0.7, p = 0.45
F0	Male	**t(77.3) = 2.5, p = 0.0**	**t(21.7) = 3.3, p = 0.0**	**t(47.9) = −2.4, p = 0.02**
	Female	**t(75.2) = 3.6, p = 0.0**	t(15.9) = 0.6, p = 0.53	t(52.7) = −0.7, p = 0.47

Bold fonts indicate significant differences.

4.6 Evaluation

The aim of classification experiments was to detect the presence or absence conflict per hour based on the physiological, acoustic, linguistic, and contextual features. Because of the highly unbalanced nature of the dataset (90.5% non-conflict), the proposed systems are evaluated in terms of both weighted and the unweighted precision, recall, and F1-scores. Unweighted scores remove the distribution bias by computing each metric for the conflict and non-conflict class and then providing the mean. Weighted metrics are computed using all samples without taking into account the number of samples per class.

5 RESULTS

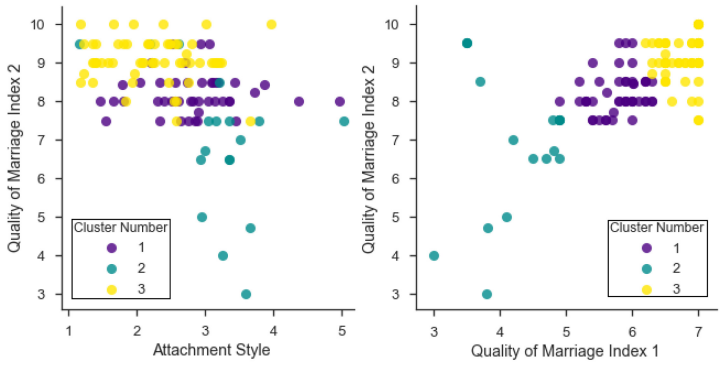
In this section, we first discuss the sub-populations resulting from the clustering algorithm (Section 5.1). We then examine the most discriminative features per sub-population with respect to the outcome of conflict (Section 5.2). Finally, we present the conflict classification results with the proposed hierarchical and adaptive frameworks of sub-population specific models, as well as their comparison with general machine learning models trained for the entire population (Section 5.3).

5.1 Clusters

Sub-population clustering resulted in 45 participants for cluster 1, 19 participants for cluster 2, and 59 participants for cluster 3. To gain insight regarding the types of individuals included in each cluster, we visualized the corresponding clusters in 2-D plots using the relationship satisfaction and attachment scores (Figure 2). Relationship satisfaction appeared to be the most separable sub-population clustering criterion (Figure 2(b) and (c)), while relationship attachment characteristics appeared to provide less distinct clusters (Figure 2(a)). Based on the aforementioned visual representations, we can intuitively understand that three main clusters appear in the data: an anxiously attached and avoidant group of individuals with low relationship satisfaction (Cluster 3), a securely attached and non-avoidant cluster of partners with high relationship satisfaction (Cluster 2), and a group of individuals lying in the middle (Cluster 1).

5.2 Statistical Analysis

Hypothesis testing indicated that different types of features were the most indicative of conflict in each cluster (Table 4). EDA measures, such as skin conductance response frequency, obtained from the male partners, appeared to be most indicative of conflict for securely attached individuals highly satisfied in their relationships (Cluster 3), while R-R interval obtained from the female partners discriminated between the presence or absence of conflict for anxiously and avoidantly attached partners with low relationship satisfaction (Cluster 2). In contrast, acoustic features, such



(a) QMI-2 v/s Attachment Style (b) QMI-2 v/s QMI-1

(c) Attachment Style v/s QMI-1

Fig. 2. Visualization of sub-population clusters using different criteria.

as speech loudness and fundamental frequency from both male and female partners, were most discriminative for the group lying between those two clusters (Cluster 1). These findings suggest that there are distinct associations between input features and the outcome of interest for the three sub-populations of our data.

5.3 Classification Results

Conflict classification results are provided from a single FNN trained on the entire population, as well as the proposed sub-population specific machine learning approaches using hierarchical and adaptive learning (Table 5). In the majority of cases, both types of sub-population specific models outperform the general FNN trained on the entire population (Single) and the three separate FNNs trained on each sub-population (Separate), indicating the importance of incorporating sub-population specific information into the machine learning models. Adaptive learning implemented with FNN fine-tuning slightly outperformed the multi-task FNN. This might be because the optimization criterion of the fine-tuned FNN only includes samples from one sub-population, potentially resulting in a smoother loss function with less local optima compared to the one of the MTL.

The different physiological, acoustic, and linguistic modalities resulted in various levels of performance, with the self-reported MQI features outperforming all separate signal-based indices (Table 6). However, when all signal-based indices were combined (Table 5), they yielded higher

Table 5. Conflict Classification Using a Single Feedforward Neural Network (FNN) Trained on the Entire Population and Sub-Population Specific Machine Learning Models Based on Hierarchical and Adaptive Learning, Implemented with a Multitask FNN and FNN Fine-Tuning, Respectively

Model	Class	Precision	Recall	F1
Single	Conflict	0.13	0.67	0.22
	Non-Conflict	0.94	0.54	0.69
	Weighted Average	0.86	0.55	0.64
	Unweighted Average	0.53	0.60	0.45
Separate	Conflict	0.13	0.57	0.22
	Non-Conflict	0.93	0.61	0.74
	Weighted Average	0.86	0.61	0.69
	Unweighted Average	0.53	0.59	0.48
Hierarchical	Conflict	0.15	0.48	0.22
	Non-Conflict	0.93	0.71	0.8
	Weighted Average	0.86	0.69	0.75
	Unweighted Average	0.54	0.60	0.51
Adaptive	Conflict	0.21	0.76	0.34
	Non-Conflict	0.96	0.68	0.80
	Weighted Average	0.89	0.70	0.76
	Unweighted Average	0.59	0.73	0.57

Input features include acoustic, linguistic, physiological, and contextual indices.

Table 6. Conflict Classification Using Self-Reported Mood and Quality of Interaction Measures, Acoustic, Linguistic, and Physiological Indices with the Adaptive Sub-Population Specific Machine Learning Models

Modality	Class	Weighted F1	Unweighted F1
Self-reported	Single	0.61	0.43
	Hierarchical	0.66	0.46
	Adaptive	0.61	0.42
Acoustic	Single	0.47	0.35
	Hierarchical	0.61	0.43
	Adaptive	0.37	0.28
Language	Single	0.59	0.41
	Hierarchical	0.58	0.41
	Adaptive	0.08	0.12
Physiological	Single	0.42	0.32
	Hierarchical	0.52	0.35
	Adaptive	0.35	0.27

accuracy compared to the self-reported measures, indicating the feasibility of employing passive sensing modalities in real-life situations for detecting events of interest. As expected from the statistical analysis (Section 5.2), acoustic indices appeared to be the most discriminative modality, yielding higher performance compared to linguistic and physiological measures.

Table 7. Results of Conflict Classification per Cluster Using the Adaptive Sub-Population Specific Models Implemented with Feedforward Neural Network Fine-tuning

Model	Class	Precision	Recall	F-1
Cluster 1	Conflict	0.19	0.78	0.31
	Non-Conflict	0.95	0.55	0.7
	Weighted Average	0.859	0.577	0.65
	Unweighted Average	0.57	0.665	0.505
Cluster 2	Conflict	0.23	0.71	0.35
	Non-Conflict	0.97	0.78	0.86
	Weighted Average	0.906	0.77	0.816
	Unweighted Average	0.6	0.745	0.605
Cluster 3	Conflict	0.23	0.81	0.36
	Non-Conflict	0.98	0.76	0.86
	Weighted Average	0.92	0.763	0.82
	Unweighted Average	0.605	0.785	0.61

We further observed performance differences across the three sub-populations (Table 7). Conflict classification for Clusters 2 and 3 generally yielded better results than did Cluster 1. This might be due to the fact that Clusters 2 and 3 represented the groups of couples with extreme levels of relationship satisfaction (low or high) and attachment characteristics (secure or insecure attachment, high or low anxious attachment; Figure 2). Conversely, Cluster 1 included the individuals in the "gray-area", for which the aforementioned relationship-based characteristics are located in the middle of the distribution. This might indicate that individuals in Cluster 1 have highly variable conflict-relevant patterns, which the corresponding machine learning models cannot adequately learn. Increasing the resolution of Cluster 1 by further splitting the corresponding samples into smaller groups might result in more discriminative sub-populations and might benefit the final system performance.

6 DISCUSSION

We proposed sub-population specific machine learning models for detecting couples' interpersonal conflict in real life. In accordance with previous work [29, 38, 62], our results indicate that different features are predictive of conflict for different sub-populations. Despite their limited age range, the recruited participants in this study depicted significant variability in terms of their socio-demographic (e.g., race, ethnicity, employment status) and relationship-based characteristics (e.g., current length, cohabitation status, satisfaction, attachment). While our demographics and survey data were typical of young adults in this age range, we cannot be certain that our algorithms would translate to different populations. Further work with participants recruited based on a different set of criteria will need to validate the effectiveness of existing methodologies.

Our result indicate that the clustering of individuals in different sub-populations allows the machine learning models to learn the most discriminative features per sub-population and benefits the performance of the final system. These findings corroborate results in psychological science, which report that conflict is experienced differently for couples with different relationship functioning characteristics [10, 19, 41]. Levenson and Gottman found that the level of relationship satisfaction moderates the amount of physiological reactivity between couples during their interactions [41], while Campbell and Simpson indicated that anxiously attached individuals perceive conflicts

differently compared to their counter-peers [10]. Our results indicate that EDA and ECG measures were informative of conflict for individuals with very low or very high relationship satisfaction (Clusters 2 and 3, respectively) but not for individuals in the middle of the spectrum (Table 4). Similarly, conflict classification yielded higher accuracy for these extreme sub-populations compared to the middle sub-population (Table 7). These results suggest that signal-based patterns of couples with very high and very low relationship functioning reflect conflict in a more pronounced way compared to the general population.

Results further indicate that by incorporating relationship-specific information, we can augment the performance of predictive models. This possibly reflects the ability of sub-population specific approaches to better address the inherent inter-individual variability of human behavior compared to general machine learning models. Our best results for conflict detection yielded F1-score of 0.76 and unweighted recall of 0.70 using the adaptive-based sub-population specific models and all the signal-based data (i.e., physiology, speech, language). Although the experimental frameworks are not the same, these results are equivalent to similar approaches for classifying the presence or absence of stress with 0.68 accuracy [62], as well as pain intensity with 0.40 accuracy in a 5-class problem [33]. Jaques et al. utilized MTL for predicting the next-day happiness and stress, yielding an accuracy of 0.60 for happiness and 0.63 for stress [31]. Kim et al. used a 3-way SVM classifier on a two-label problem (low-conflict vs. high-conflict) and obtained an average recall of 0.71 and an average F-measure of 0.71 on taking both the conversational and prosodic features in the case of political debates [36]. Taking into account that detecting conflict between couples is an inherently difficult task, since it involves the modeling of the interplay between various complex, psychological, interpersonal processes, the proposed multi-modal system achieved comparable and sometimes slightly better performance to previous studies, indicating the feasibility of sub-population specific machine learning for modeling subtle facets of human behavior.

The clustering criterion is an important factor for building sub-populations from a set of data. We followed a knowledge-driven approach by taking into account findings from psychological science indicating that individuals with different relationship satisfaction and attachment characteristics perceive interpersonal conflict differently [10, 19]. While the majority of previous work has successfully used such predetermined clustering criteria [6, 39, 62], there have also been studies that have split participants based on their signal-based measures (e.g., mean physiological levels) [33]. In future work, we plan to explore joint optimization approaches to learn the optimal cluster configuration for a given outcome of interest and examine whether this data-driven configuration aligns with knowledge-driven approaches.

Sub-population clustering was performed using a K-Means classifier. Despite the intuitiveness and effectiveness of this approach, more sophisticated clustering mechanisms, such as hierarchical dendogram approaches [67], might be able to recover highly variable sub-populations with better resolution and potentially stratify individuals in the middle of the spectrum. Using ensemble learning that randomizes the various groups of features and clustering criteria [39] might also be a useful approach to this problem. We further observed that while the proposed system was able to reliably detect conflict for couples belonging to the "extreme" clusters (i.e., couples with low and high relationship satisfaction), conflict detection for the intermediate cluster (i.e., couples with medium relationship satisfaction and anxiety) was not as accurate. A potential reason behind this could be that couples in the "gray-zone" might depict higher variability in the way they express and experience conflict compared to the couples in the extreme clusters. This ambiguity might prevent the model parameters for the couples in the "gray-zone" from being adequately learned. A way to address this limitation would be to perform soft clustering, according to which each couple is assigned to a probability of belonging to a given cluster, implemented through mixture of experts methodologies [27, 62]. In this way, the parameters of each cluster will be learned using

all samples, each weighted with an importance proportional to the strength of its belonging to a given cluster, potentially yielding more robust representations.

The data was collected in an ambulatory setting, which inevitably increased the amount of noise in the acoustic and physiological signals. Pre-processing techniques related to high-frequency noise elimination, movement artifact detection, and voice activity detection (Section 4.1) were applied to reduce the inherent noise in the data. Inspection by human annotators was further performed to ensure that the automated techniques have reliably removed the noisy parts of the signals (e.g., motion artifacts, high-frequency noise) and have retained the meaninfgul ones (e.g., skin conductance responses, QRS intervals). Although it is not guaranteed that noise is fully eliminated, our results indicate that the features extracted from the denoised signals can still provide meaningful patterns related to the outcome of interest. Signal denoising and pre-processing comprises a fundamental and necessary step for any ambulatory monitoring application, especially when real-life interventions are of interest. Researchers are advised to understand the inner mechanics behind signal denoising techniques and make sure that the resulting denoised signals retaing meaninfgul information regarding the outcome of interest.

Human transcribers converted the automatically detected speech segments of the audio signals into text. This might impose a constraint on level of automation of the proposed model. Previous studies, however, suggest that linguistic features derived based on automatic speech recognition systems can yield comparable performance for detecting human-related outcomes to the same features extracted based on human transcriptions [66, 68]. We expect that automatic speech transcription systems will continue to improve in the near future [58], supporting the feasibility of a fully automated approach for leveraging language features from speech in real-life applications.

Results from this work indicate the feasibility of detecting behaviors of clinical interest in real-life for enhancing mental and emotional well-being. Despite the encouraging results, a variety of steps need to occur before the broad adoption of such technologies in real-life applications. It is of the utmost importance to rigorously test such algorithms in real-life situations in order to get a better understanding their performance is various populations and under different conditions. Baseline data from each user might be potentially valuable to increase the reliability and precision of such algorithms. Technical considerations in terms of internet connectivity, storage capacity, and on-device computational power for data analysis need to be taken into account to ensure that the proposed applications can be accessible from diverse pool of individuals, such as people with low socio-economic status, elderly adults, or individuals residing in remote geographical locations [42, 57]. Tuning the sensitivity and specificity of the proposed algorithms and designing user-friendly human-computer interfaces for feedback provision needs to be performed in collaboration with researchers from Psychological Science and Human-Computer Interaction, in order to ensure that potential users can unobtrusively and meaningfully engage with such systems [18].

7 CONCLUSIONS

In this article, we proposed the integration of sub-population specific information into machine learning systems for accurately detecting couples' conflict in real life. We used two different learning approaches: the first relied on hierarchical learning with a MTL FNN, while the second relied on adaptive learning using FNN fine-tuning. We compared the aforementioned approaches to general machine learning models through a dataset containing couples' interactions in real life. Different types of features were found to be more discriminative for various sub-populations, reflecting the high levels of inter-individual variability observed in our population. The proposed sub-population specific approaches outperformed general machine learning models and separate models trained for each sub-population for detecting conflict. Our results further suggest that conflict is more easily detected for couples with extreme levels of relationship satisfaction and attachment style.

The same does not hold for couples in the intermediate "gray-zones," indicating the presence of more complex interactions between physiological and acoustic variables for detecting conflict in these couples. Findings from this study can inform the development of machine learning systems for detecting events of interest relevant to health and well-being in ambulatory settings, setting the foundation for developing in-the-moment interventions in real life.

APPENDIX

This appendix includes a detailed information on the items of the EMA questionnaire, administered every hour to each partner (Section 3).

Table 8. Items of the Ecological Momentary Assessment (EMA) Questionnaire

Item	Choice of answers
What is your ID number?	N/A
Which partner are you?	a. Partner 1; b. Partner 2
How stressed were you in the last hour?	0–100 (not at all–extremely)
What was the source of stress? Please check all that apply.	a. The romantic partner; b. Another person; c. Work or school; d. Other events/news; e. Not applicable. I answered 0 to the last question.
Did you consume any of the following in the last hour? Please check all that apply.	a. Coffee, tea, or energy drinks; b. Alcohol; c. Tobacco; d. Other drugs; e. None of the above
Did you engage in any physical activity in the last hour?	a. Not at all; b. Low Intensity; c. Moderate intensity; d. High intensity
In the last hour, how happy were you?	0–100 (not at all–extremely)
In the last hour, how sad were you?	0–100 (not at all–extremely)
In the last hour, how nervous were you?	0–100 (not at all–extremely)
In the last hour, how angry were you?	0–100 (not at all–extremely)
In the last hour, how close or connected did you feel toward your romantic partner?	0–100 (not at all–extremely)
In the last hour, how irritated or annoyed did you feel toward your romantic partner?	0–100 (not at all–extremely)
Did you express this irritation to your romantic partner (speaking, texting, etc.)?	a. Yes; b. No; c. Not applicable. I answered 0 to the last question.
In the last hour, have you had any contact with your romantic partner via text or phone?	a. Yes; b. No
If you disabled the audio, please enable it now if you are able.	a. Okay

REFERENCES

[1] [n.d.]. Actiwave Cardio. Retrieved February 24, 2019 from http://www.camntech.com/products/actiwave-cardio/.

[2] [n.d.]. Keras. Retrieved February 26, 2019 from https://keras.io/.

[3] [n.d.]. The USC Couple Mobile Sensing Project. Retrieved February 24, 2019 from http://homedata.github.io/.

[4] Jonathan Aigrain, Séverine Dubuisson, Marcin Detyniecki, and Mohamed Chetouani. 2015. Person-specific behavioural features for automatic stress detection. In *Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 3. IEEE, 1–6.

[5] M. Benedek and C. Kaernbach. 2010. Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology* 47, 4 (2010), 647–658.

[6] Dimitris Bertsimas, Nathan Kallus, Alexander M. Weinstein, and Ying Daisy Zhuo. 2017. Personalized diabetes management using electronic medical records. *Diabetes Care* 40, 2 (2017), 210–217.

[7] Richard Brislin. 1993. *Understanding Culture's Influence on Behavior*. Harcourt Brace Jovanovich.

[8] Bethany Butzer and Lorne Campbell. 2008. Adult attachment, sexual satisfaction, and relationship satisfaction: A study of married couples. *Personal Relationships* 15, 1 (2008), 141–154.

[9] Fabio Calefato, Filippo Lanubile, Federico Maiorano, and Nicole Novielli. 2018. Sentiment polarity detection for software development. *Empirical Software Engineering* 23, 3 (2018), 1352–1382.

[10] Lorne Campbell, Jeffry A. Simpson, Jennifer Boldry, and Deborah A. Kashy. 2005. Perceptions of conflict and support in romantic relationships: The role of attachment anxiety. *Journal of Personality and Social Psychology* 88, 3 (2005), 510.

[11] Deborah M. Capaldi, Joann Wu Shortt, and Lynn Crosby. 2003. Physical and psychological aggression in at-risk young couples: Stability and change in young adulthood. *Merrill-Palmer Quarterly (1982-)* (2003), 1–27.

[12] Marie-José Caraty and Claude Montacié. 2015. Detecting speech interruptions for automatic conflict detection. In *Conflict and Multimodal Communication*. Springer, 377–401.

[13] Gavin C. Cawley and Nicola L. C. Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* 11, Jul (2010), 2079–2107.

[14] Theodora Chaspari, Brian Baucom, Adela C. Timmons, Andreas Tsiartas, Larissa Borofsky Del Piero, Katherine J. W. Baucom, Panayiotis Georgiou, Gayla Margolin, and Shrikanth S. Narayanan. 2015. Quantifying EDA synchrony through joint sparse representation: A case-study of couples' interactions. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 817–821.

[15] Theodora Chaspari, Adela C Timmons, Brian R. Baucom, Laura Perrone, Katherine J. W. Baucom, Panayiotis Georgiou, Gayla Margolin, and Shrikanth S. Narayanan. 2017. Exploring sparse representation measures of physiological synchrony for romantic couples. In *Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction (ACII'17)*. IEEE, 267–272.

[16] T. Chaspari, A. Tsiartas, L. I. Stein, S. A. Cermak, and S. S. Narayanan. 2015. Sparse representation of electrodermal activity with knowledge-driven dictionaries. *IEEE Transactions on Biomedical Engineering* 62, 3 (2015), 960–971.

[17] Lei A. Clifton, David A. Clifton, Marco A. F. Pimentel, Peter J. Watkinson, Lionel Tarassenko, et al. 2013. Gaussian processes for personalized e-health monitoring with wearable sensors. *IEEE Transactions on Biomedical Engineering* 60, 1 (2013), 193–197.

[18] Jonathan S. Comer, Kristina Conroy, and Adela C. Timmons. 2019. Ensuring wearable devices don't wear out their welcome: Cautions for the mental health care road ahead. *Clinical Psychology: Science and Practice* 16 (2019), e12297. DOI: 10.1111/cpsp.12297

[19] Gary Creasey and Matthew Hesson-McInnis. 2001. Affective responses, cognitive appraisals, and conflict tactics in late adolescent romantic relationships: Associations with attachment orientations. *Journal of Counseling Psychology* 48, 1 (2001), 85.

[20] L. De Raeve, N. W. H. Jansen, P. A. Van den Brandt, R. Vasse, and I. J. Kant. 2009. Interpersonal conflicts at work as a predictor of self-reported health outcomes and occupational mobility. *Occupational and Environmental Medicine* 66, 1 (2009), 16–22.

[21] Alberto De Santos, Carmen Sánchez-Avila, Javier Guerra-Casanova, and Gonzalo Bailador-Del Pozo. 2011. Real-time stress detection by means of physiological signals. *Recent Application in Biometrics* 58 (2011), 4857–65. DOI: 10.5772/18246

[22] Morton Deutsch. 2003. Cooperation and conflict: A personal perspective on the history of the social psychological study of conflict resolution. *International Handbook of Organizational Teamwork and Cooperative Working* (2003), 9–43.

[23] Tara Donker, Katherine Petrie, Judy Proudfoot, Janine Clarke, Mary-Rose Birch, and Helen Christensen. 2013. Smartphones for smarter delivery of mental health programs: A systematic review. *Journal of Medical Internet Research* 15, 11 (2013).

[24] Jochen Fahrenberg, Michael Myrtek, Kurt Pawlik, and Meinrad Perrez. 2007. Ambulatory assessment–Monitoring behavior in daily life settings: A behavioral-scientific challenge for psychology. *European Journal of Psychological Assessment* 23, 4 (2007), 206.

[25] R. Chris Fraley and Keith E. Davis. 1997. Attachment formation and transfer in young adults' close friendships and romantic relationships. *Personal Relationships* 4, 2 (1997), 131–144.

[26] Enrique Garcia-Ceja, Michael Riegler, Tine Nordgreen, Petter Jakobsen, Ketil J. Oedegaard, and Jim Tørresen. 2018. Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing* 51 (2018), 1–26.

[27] Ankit Goyal, Naveen Kumar, Tanaya Guha, and Shrikanth S. Narayanan. 2016. A multimodal mixture-of-experts model for dynamic emotion prediction in movies. In *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2822–2826.

[28] Félix Grezes, Justin Richards, and Andrew Rosenberg. 2013. Let me finish: Automatic conflict detection using speaker overlap. In *Interspeech*. 200–204.

[29] Aditya Gujral, Theodora Chaspari, Adela C. Timmons, Yehsong Kim, Sarah Barrett, and Gayla Margolin. 2018. Population-specific detection of couples' interpersonal conflict using multi-task learning. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 229–233.

[30] Jennifer Healey, Rosalind W. Picard, et al. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems* 6, 2 (2005), 156–166.

[31] Natasha Jaques, Sara Taylor, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2016. Multi-task learning for predicting health, stress, and happiness. In *NIPS Workshop on Machine Learning for Healthcare*.

[32] Karen A. Jehn. 1995. A multimethod examination of the benefits and detriments of intragroup conflict. *Administrative Science Quarterly* (1995), 256–282.

[33] Markus Kächele, Patrick Thiam, Mohammadreza Amirian, Friedhelm Schwenker, and Günther Palm. 2016. Methods for person-centered continuous pain intensity assessment from bio-physiological channels. *IEEE Journal of Selected Topics in Signal Processing* 10, 5 (2016), 854–864.

[34] Nathan Kallus. 2017. Recursive partitioning for personalization using observational data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR.org, 1789–1798.

[35] Isabel L. Kampmann, Paul M. G. Emmelkamp, and Nexhmedin Morina. 2016. Meta-analysis of technology-assisted interventions for social anxiety disorder. *Journal of Anxiety Disorders* 42 (2016), 71–84.

[36] Samuel Kim, Fabio Valente, and Alessandro Vinciarelli. 2012. Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'12)*. IEEE, 5089–5092.

[37] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. DEAP: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing* 3, 1 (2011), 18–31.

[38] Saskia Koldijk, Mark A. Neerincx, and Wessel Kraaij. 2018. Detecting work stress in offices by combining unobtrusive sensors. *IEEE Transactions on Affective Computing* 9, 2 (2018), 227–239.

[39] Nicholas D. Lane, Ye Xu, Hong Lu, Shaohan Hu, Tanzeem Choudhury, Andrew T. Campbell, and Feng Zhao. 2014. Community similarity networks. *Personal and Ubiquitous Computing* 18, 2 (2014), 355–368.

[40] Alistair Letcher, Jelena Trišović, Collin Cademartori, Xi Chen, and Jason Xu. 2018. Automatic conflict detection in police body-worn audio. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2636–2640.

[41] Robert W. Levenson and John M. Gottman. 1983. Marital interaction: Physiological linkage and affective exchange. *Journal of Personality and Social Psychology* 45, 3 (1983), 587.

[42] James A. Levine. 2017. The application of wearable technologies to improve healthcare in the worldâs poorest people. *Technology and Investment* 8, 02 (2017), 83.

[43] Dianbo Liu, Fengjiao Peng, Andrew Shea, Rosalind Picard, et al. 2017. DeepFaceLIFT: Interpretable personalized models for automatic estimation of self-reported pain. *arXiv preprint arXiv:1708.04670* (2017).

[44] Elliot Mishler. 1979. Meaning in context: Is there any other kind? *Harvard Educational Review* 49, 1 (1979), 1–19.

[45] David C. Mohr, Mi Zhang, and Stephen M. Schueller. 2017. Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology* 13 (2017), 23–47.

[46] Josianne Mondor, Pierre McDuff, Yvan Lussier, and John Wright. 2011. Couples in therapy: Actor-partner analyses of the relationships between adult romantic attachment and marital satisfaction. *The American Journal of Family Therapy* 39, 2 (2011), 112–123.

[47] Sebastian C. Müller and Thomas Fritz. 2015. Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress. In *Proceedings of the 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, Vol. 1. IEEE, 688–699.

[48] Michael Myrtek, Eveline Aschenbrenner, Georg Brügner, et al. 2005. Emotions in everyday life: An ambulatory monitoring study with female students. *Biological Psychology* 68, 3 (2005), 237–255.
[49] Jennifer Nicholas, Mark Erik Larsen, Judith Proudfoot, and Helen Christensen. 2015. Mobile apps for bipolar disorder: A systematic review of features and content quality. *Journal of Medical Internet Research* 17, 8 (2015).
[50] Robert Norton. 1983. Measuring marital quality: A critical look at the dependent variable. *Journal of Marriage and the Family* (1983), 141–151.
[51] K. Daniel O'Leary, Julian Barling, Ileana Arias, Alan Rosenbaum, Jean Malone, and Andrea Tyree. 1989. Prevalence and stability of physical aggression between spouses: A longitudinal analysis. *Journal of Consulting and Clinical Psychology* 57, 2 (1989), 263.
[52] Yannis Panagakis, Stefanos Zafeiriou, and Maja Pantic. 2014. Audiovisual conflict detection in political debates. In *Workshop at the European Conference on Computer Vision*. Springer, 306–314.
[53] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The Development and Psychometric Properties of LIWC2015*. Technical Report.
[54] Kelly Peterson, Ognjen Rudovic, Ricardo Guerrero, and Rosalind W. Picard. 2017. Personalized Gaussian processes for future prediction of Alzheimer's disease progression. In *Proceedings of the ML4H:Machine Learning for Health, 31st Conference on Neural Information Processing Systems*.
[55] Ming-Zher Poh, Nicholas C. Swenson, and Rosalind W. Picard. 2010. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE Transactions on Biomedical Engineering* 57, 5 (2010), 1243–1252.
[56] David S. Riggs, K. Daniel O'Leary, and F. Curtis Breslin. 1990. Multiple correlates of physical aggression in dating couples. *Journal of Interpersonal Violence* 5, 1 (1990), 61–73.
[57] William A. Satariano, Andrew E. Scharlach, and David Lindeman. 2014. Aging, place, and technology: Toward improving access and wellness in older populations. *Journal of Aging and Health* 26, 8 (2014), 1373–1389.
[58] Odette Scharenborg. 2007. Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication* 49, 5 (2007), 336–347.
[59] Jeffry A. Simpson, W. Steven Rholes, and Dede Phillips. 1996. Conflict in close relationships: An attachment perspective. *Journal of Personality and Social Psychology* 71, 5 (1996), 899.
[60] Jagdip Singh. 2000. Performance productivity and quality of frontline employees in service organizations. *Journal of Marketing* 64, 2 (2000), 15–34.
[61] Donna Spruijt-Metz and Wendy Nilsen. 2014. Dynamic models of behavior for just-in-time adaptive interventions. *IEEE Pervasive Computing* 13, 3 (2014), 13–17.
[62] Sara Ann Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2017. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing* (2017). DOI: 10.1109/TAFFC.2017.2784832
[63] Adela C. Timmons, Theodora Chaspari, Sohyun C. Han, Laura Perrone, Shrikanth S. Narayanan, and Gayla Margolin. 2017. Using multimodal wearable technology to detect conflict among couples. *Computer* 3 (2017), 50–59.
[64] Adela C. Timmons, Gayla Margolin, and Darby E. Saxbe. 2015. Physiological linkage in couples and its implications for individual and interpersonal functioning: A literature review. *Journal of Family Psychology* 29, 5 (2015), 720.
[65] Carmen Vidaurre, Tilmann H. Sander, and Alois Schlögl. 2011. BioSig: The free and open source software library for biomedical signal processing. *Computational Intelligence and Neuroscience* (2011), 935364. DOI: 10.1155/2011/935364
[66] Sarah Weusthoff, Garren Gaut, Mark Steyvers, David C. Atkins, Kurt Hahlweg, Jasara Hogan, Tanja Zimmermann, Melanie S. Fischer, Donald H. Baucom, Panayiotis Georgiou, et al. 2018. The language of interpersonal interaction: An interdisciplinary approach to assessing and processing vocal and speech data. *Machine Learning* 7 (2018), 1.
[67] Leland Wilkinson and Michael Friendly. 2009. The history of the cluster heat map. *The American Statistician* 63, 2 (2009), 179–184.
[68] Bo Xiao, Zac E. Imel, Panayiotis G. Georgiou, David C. Atkins, and Shrikanth S. Narayanan. 2015. "Rate My Therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PloS One* 10, 12 (2015), e0143055.
[69] David Zeevi, Tal Korem, Niv Zmora, David Israeli, Daphna Rothschild, Adina Weinberger, Orly Ben-Yacov, Dar Lador, Tali Avnit-Sagi, Maya Lotan-Pompan, et al. 2015. Personalized nutrition by prediction of glycemic responses. *Cell* 163, 5 (2015), 1079–1094.

Received April 2019; revised September 2019; accepted November 2019