FEELing (key)Pressed: Implicit Touch Pressure Bests Brain Activity in Modelling Emotion Dynamics in the Space Between Stressed and Relaxed

X. Laura Cang, Rubia R. Guerra, Bereket Guta, Paul Bucci, Laura Rodgers, Hailey Mah, Qianqian Feng, Anushka Agrawal & Karon E. MacLean
University of British Columbia, Vancouver, Canada

***Abstract*—In-body lived emotional experiences can be complex, with time-varying and dissonant emotions evolving simultaneously; devices responding in real-time to estimate personal human emotion should evolve accordingly. Models assuming generalized emotions exist as discrete states fail to operationalize valuable information inherent in the dynamic and individualistic nature of human emotions. Our multi-resolution emotion self-reporting procedure allows the construction of emotion labels along the Stressed-Relaxed scale, differentiating not only what the emotions are, but how they are transitioning – *e.g.,* "hopeful but getting stressed" *vs.* "hopeful and starting to relax". We trained participant-dependent hierarchical models of contextualized individual experience to compare emotion classification by modality (brain activity and keypress force from a physical keyboard), then benchmarked classification performance at F1-scores=[0.44, 0.82] (chance $F1 = 0.22$, $\sigma = 0.01$) and examined high-performing features. Notably, when classifying emotion evolution in the context of an experience that realistically varies in stress, pressure-based features from keypress force proved to be the more informative modality, and more convenient when considering intrusiveness and ease of collection and processing. Finally, we present our FEEL (Force, EEG and Emotion-Labelled) dataset, a collection of brain activity and keypress force data, labelled with self-reported emotion collected during tense videogame play (N=16) and open-sourced for community exploration.**

***Index Terms*—Affective Touch, Dynamic Emotion Classification, Emotion Labelling Methods, Keypress Force, Brain Activity**

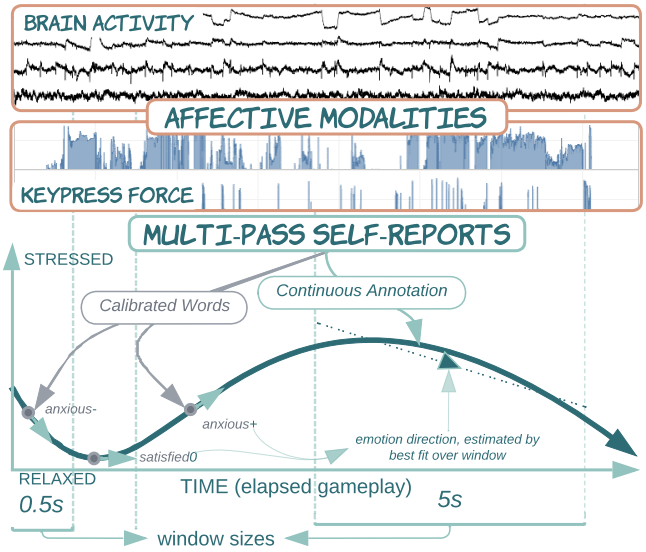

Figure 1: An emotion experience trajectory estimated by emotion transition. We built models on two modalities: brain activity (EEG) and keypress force (FSR), distinguishing intensifying(+), stable(0), or resolving(-) stress, at 0.5s and 5s windows.

I. Introduction

If emotionally reactive machines could interpret the transitional nature or direction of their inherently emotional human users, responses could be designed to be contextually appropriate. Due to variations in human emotion expression and personal preferences of a desired response, such machines will likely need to be customized and tuned to the individual. In particular, a system must be able to recognize user-specific emotion *transition* through some identifiable parameter, such as intensity or polarity. For instance, when a custom emotion-aware game system estimates a user's "anxiety" levels as low, it could ramp intensity up to a personal "frustration" threshold, to avoid game burnout.

Natural (unmediated) interpersonal emotion communication relies on many nonverbal cues: we interpret emotion expressions from others through eye contact, vocal inflections, body language and touch behaviour [1]. Using machines to recognize social touch unlocks the significant emotional content encoded in physical contact [2]–[4].

To model spontaneously evolving emotion in the vicinity of a participant-defined Stressed-Relaxed scale, we collected participant biosignal data while they played Playdead's *Inside* [5], an emotionally evocative videogame. We followed the multipass data labelling protocol described in [6], recording brain activity using electroencephalography (EEG) and keypress force via a Force Sensitive Resistor (FSR)-embedded keyboard. Both have been shown to encode emotion [7]–[9] and are reasonable to collect during videogame play. While we considered other well-studied emotion-encoding biosignals (namely electrodermal activity, pulse oximetry, and electrocardiography), sensors that were worn on fingers or otherwise generated electrical interference with the sensitive EEG system proved unsuitable for this study.

In this paper, we present our FEEL dataset (collected under a separately peer-reviewed protocol [6]) and use it to ask: **How well can we classify emotion *transitions* or directions using keypress force *vs.* brain activity collected during**

an emotionally evocative video gameplay? Specifically, we demonstrate a personalized-to-participant emotion interpretation paradigm, then assess model performance, efficacy and practicality of classifying emotions as they are in flux, by comparing two distinct implicit and highly personalized expressive modalities which play out at different timescales (brain activity and keypress force). To further inform model design, particularly with respect to modality-specific frequency characteristics, we provide an evidence-based reference scale for window size selection. We contribute:

1) The FEEL dataset, collected using a multipass labelling protocol featuring co-designed scales for annotating emotion self-report on keypress force and brain activity data.
2) An empirical demonstration of personalized emotion *transition* classification that distinguishes between emotion transition labels across a Stressed-Relaxed scale (*e.g.,* cautious +, 0, or - as "feeling cautious and getting more stressed" *vs.* "cautious and stable" *vs.* "cautious but relaxing" respectively).
3) Evidence that hierarchical classification of emotion evolution along a Stressed-Relaxed dimension using touch pressure features performs nearly twice as well as continuous brain activity.

II. Background

Machine interpretation of spontaneous emotion requires models built on ecologically valid emotion data. From choice of expressive modality to data labelling, we ground our data collection and modelling choices in existing literature.

Affect-Encoding Modalities: Although the biological mechanisms through which emotion modulates touch are still unclear [10], touch is a concrete, perceivable and expressive act [11] and a promising modality for both inferring and influencing emotion experiences [12]. Relative to other channels commonly used in emotion research – EEG, brain imaging, heart-rate, facial configurations, body posture, speech [13], [14] – touch can be easier to harness, less intrusive to collect, and gives the participant more immediate agency in terms of behaviour compared to biological signals.

Affective touch classification has largely been based on observation and evaluation of toucher behaviour when they are prompted to reflect on a past experience [15], or to act in an emotional context [2]. While interpersonal touch pressure has been shown to communicate currently felt affect [4], [15], investigating keypress force or pressure for evidence of emotion "leakage" in the absence of communicative intent is relatively new [8]. Using pressure-sensitive keyboards, emotion has been classified using typing pressure with up to 93% correspondence to self report (chance 17%) [9], with [8] finding a positive correlation between stress and typing force. Now, we explore how keypress force may communicate emotional transitions between Stressed and Relaxed on pressure-sensitive keys.

Changes in electrical potential in brain activity or electroencephalography (EEG) [16] for emotion classification is dominated by Event-Related Potentials (ERPs). However, as ERP time windows are typically constructed within 100-750ms after an event [17], [18], the ERP fails to capture emotion *evolution*, where change occurs over the course of minutes and hours [19]. Recently, 2D differential entropy-based features capturing spatial relationships and Convolutional Neural Networks (CNNs) can classify 1s data instances over emotional experiences (positive, negative, neutral) lasting 4 minutes at an accuracy of 97.10% (chance 33%) [20].

Here, we build on machine classification of emotion transition using multiscale self-reports on brain activity and keypress force during video gameplay – a dynamic emotion experience.

Emotion Self-Report: Time-varying emotion expression can be attributed to complex neurological and physiological regulation mechanisms [21], appraisal effects [22], cognition and contextual factors [23], [24]. To simplify in-lab research, computational emotion modelling often relies on emotions being represented as a point in an emotion plane along easy-to-read scales with dimensions of arousal, valence, and/or dominance [15], [25]. While these models are convenient, in real use we need to address emotion evolution over time. However, commonly used labels on the arousal-valence circumplex model [26], PANAS [27], or SAM [28] (among others) quickly become intractable for sampling at the rate of change for emotion (ranging from a few seconds to several hours [29]).

Emotion self-report with any measurement scheme raises generalizability concerns. Our understandings of the instrument scale are highly subjective [25], [30] and influenced by life experiences and personal history [31]. Any set of ground-truth labels for self-reported emotion are likely similarly personalized: *e.g.,* one person's *anger* scale may be unrecognizable by another, or even by themselves at another time. In an evolving emotion experience, recognizing a particular user's near-future emotional expression can improve the temporal and situational appropriateness of a machine response.

Emotion Modelling with Multiple Reporting Passes: With time and reflection, emotional assessment of an experience may be dramatically different from initial evocation [22], [24]. Emotions may be most intense while directly in an experience [19], [32], but articulation can only occur after some time to assess and consider the appropriate language [33]. [34] suggests the ideal window of time for emotion-naming may be shortly after an experience, to give time for processing [32] but before memory degrades [35].

Computational emotion models often rely on a single pass of emotion that is self-reported [13], [14], [36] or observed and labelled by judges. To our knowledge, our study is the first to triangulate multiple self-report methods for more reliable observation of emotion evolution.

We demonstrate the use of our FEEL dataset for exploring classification models of incidental touch pressure as a modality that captures implicit emotion expression, comparing performance to models of the more intrusive, but more studied, brain activity signals.

III. Dataset Description

The FEEL collection protocol [6] was a significant investment requiring ∼400 researcher hours: each 2-hour session required a team of 4 researchers, with 2 hrs of setup, calib-

ration and breakdown time, plus earlier piloting. As a quality assurance measure, we reviewed protocol adherence during data collection and signal quality for all 23 participants. Given our plans to publish this dataset, we used a very high standard for data quality and consistency, setting aside a participant's entire record where at any point during the session there was any suspicion of excessive noise in EEG data, equipment malfunction, synchronization mishap or possible recording errors. This left us with 16 publishable records (7 omitted due to any combination of the above set of minor issues).

The FEEL dataset consists of comma separated value (.csv) files organized by participant. Video data is excluded for participant privacy. Analyses start with this 5.4GB dataset, available at https://www.cs.ubc.ca/labs/spin/FEEL_dataset *[to be posted upon acceptance]*.

Data Capture and Preparation: As part of recruitment, participants completed a questionnaire adapted from the Trait Meta Mood Scale (TMMS) [37]. Based on these results, we invited only those scoring with high emotion clarity and low emotion suppression based on their responses.

Of the N=16 participants, 8 are female and 8 male; 8 between 19-24 and the other 8 between 25-34 years of age. All played videogames regularly from a few hours a month up to 4 hours daily, nearly all of whom report 1-6+ hours per week; none had played *Inside*. All were compensated $30 for the 2-hr data collection session.

Data collection was conducted in four steps [6]:

1) *Initial Gameplay* generated streams of participant brain activity (EEG) and keypress force from an FSR-embedded keyboard timestamped from the first keystroke, indicating the start of gameplay.
2) In *Word Scale Calibration*, participants placedpre-selected emotion words relative to one another on a Stressed-Relaxed emotion scale.
3) In the first self-report cycle (*Calibrated Interview*), participants then reviewed and annotated the gameplay video with their calibrated word sets.
4) Finally, in the second self-report cycle (*Continuous Annotation*) they used a 1D joystick (position sampled at 256Hz) to annotate the video.

We timestamped data streams with corresponding frames from the Initial Gameplay video, where participant gameplay averaged 13:24 minutes (min 8:25, max 21:37, SD 3:53).

Brain Activity Data Stream (EEG): Participants were instructed to minimize conscious movement; researchers noted sessions with excessive motion to check for unusable EEG data (deciding to omit it if so).

We captured brain activity data using EGI's EEG 400 system[1], sampled at 1kHz, with a band pass filter of 1-50Hz applied in post-processing. We followed standard practice in removing high frequency jitter and 60Hz mains noise [38] while retaining α, β, θ, and γ frequency bands (associated with emotion processing [39]). We did not downsample because (a) we were able to efficiently capture important dynamics using spectral-domain features and (b) we are still exploring which frequency components are important.

We checked classification performance over a number of data cleaning procedures using MNE-Python tools[2], including artifact removal and baseline correction by the entire gameplay duration, and by adjacent windows. We also tried applying Independent Component Analysis (ICA) to address eye blinks and removing channel segments with exceptionally high noise levels. These procedures yielded no significant classification improvement or a marginal performance decline over 30 training and testing iterations. So, we report results and publish the dataset with minimal pre-processing[3], largely leaving EEG "alone" as recommended by [40]. We used this data version in the classification models reported in this paper.

Keypress Force Data Stream (KFP): We embedded force-sensitive resistors (FSRs) on game-specific control keys (four direction keys and ALT) on a standard keyboard. Force ranged from 0 (no contact) to 1023 units ($\sim$1kg)[4]. We downsampled FSR data from 52Hz to match videogame framerate at 30Hz.

Timeline with Calibrated Words (TwCW): The Timeline was created from collection sequence Steps 2 and 3.

Word Calibration: Following gameplay, players calibrated a Stressed-Relaxed emotion scale, contextualizing scale-points with memories of their recent gameplay experience and marking 13 pre-selected emotionally "Calibrated Words": *Cautious, Satisfied, Hopeful, Frustrated, Anxious, Nervous, Threatened, Resigned, Alert, Accomplished, Fearful, Dread*, and *Curious*. Participants were also allowed to write-in up to two additional words. This individualized calibration step contextualizes how each person perceives and uses these words with respect to the Stressed/Relaxed dimension, improving participant-researcher grounding on language usage [6], [25].

TwCW Construction: Players reviewed their gameplay video, annotating (calibrated) emotion words at timepoints associated with strong emotion. To construct the TwCW, we associated each interview annotation with the calibration value for that word, at the annotated gameplay timestamp.

Continuous Annotation Stream (CA): In the second gameplay review, the CA is generated from a non-biased joystick (holds last position rather than returning to centre) tracing an emotion time series, where the resulting curve is a proxy for a participant's true emotion trajectory between Relaxed and Stressed over the timeline of the gameplay experience. Joystick position readings were matched with video frame rate of 30Hz to ensure alignment with video playback. We smoothed analog jitter in the joystick data with a simple moving average filter, then normalized range to [0:1].

[1]EGI EEG system details: https://www.egi.com/research-division/eeg-systems/geodesic-eeg-systems. Model 400 features a 64-channel Routine Hydrocel geodesic sensor net, proprietary NetStation data collection and visualization software.

[2]MNE tutorials available at https://mne.tools/stable/index.html

[3]Included processing ensures labelling format consistency and time alignment across data streams. The FEEL dataset is published unfiltered with no artifact, segment, nor baseline correction. Any processing prior to classification is described in Section IV - Methods.

[4]As defined by the FSR specifications available commercially at https://www.robotshop.com/en/force-sensing-resistor-fsr.html.

Figure 1 highlights the data collected during the study: player-specific gameplay streams (EEG and FSR), emotion word calibrations, and the TwCW and CA – two time-series of emotion self-report annotated on the same dimensional plane of the Stressed-Relaxed scale over the gameplay timeline.

IV. Methods

To demonstrate personalized emotion transition classification using our FEEL dataset, we created participant-specific hierarchical multi-label models to leverage several benefits, in the context of multi-label classification tasks featuring multiple label streams – here, emotion words and quantitative stress measures. By incorporating a hierarchical structure into the model, we capture the complex and dependent relationships that may exist between labels, thereby improving classification accuracy [41], [42]. This approach is also more flexible in handling different types of label streams, and comprehensive in its view of the individual's emotional state in both brain activity and keypress force.

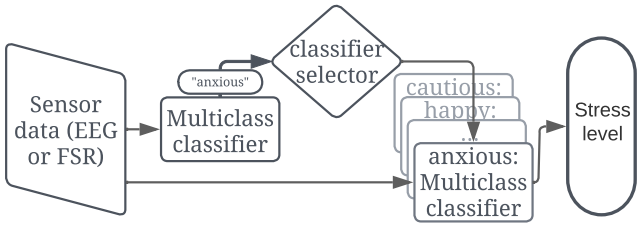

Figure 2: Visual representation of our hierarchical approach. We use a local multi-class classifier per parent node: first predicting the Calibrated Word, then training models for each emotion word subset, and outputting three possible directions relative to Stressed.

A. Data Instances: Labels and Window Lengths

We aligned FSR, EEG, and emotion self-report time series, dividing streams into non-overlapping, equal-duration windows. We analyzed window lengths of 0.5, 1, 2 and 5s, spanning ERP window range [17], [18] up to perceived emotion duration of "a few seconds" [19], [43]; 1s and 2s windows match other emotion-related classification studies [15], [44], [45]. Results from intermediate lengths followed the trend set by the extreme values, so we report only 0.5 and 5s for brevity.

A single data instance consists of features and labelled emotion class calculated from data within one window. Across all participant sessions, we collected an average (over all participants) of 1435.13 data instances for 0.5s windows (σ=405.51) and 142.63 instances for 5s windows (σ=40.75).

To implement hierarchical multi-label outputs, we used the Python package HiClass [46] with algorithms implemented in scikit-learn [47], XGBoost [48], and the Pytorch framework [49]. In a 2-stage approach, we first trained participant-specific multi-class models that output the emotion words from the TwCW, then trained a classifier from the CA by each Calibrated Word, outputting binned direction values (slope of best-fit-line as in [6]).

The resulting label set across 16 participants consisted of approximately 11 ($\mu = 10.94$, $\sigma = 1.91$) distinct calibrated emotion *word* labels, out of a possible 15 calibrated emotion words (13 provided and 2 write-ins per participant, Table I).

For each word, there are three possibilities regarding transition direction; *e.g., Nervous* could be $nervous$+, $nervous$0, and $nervous$-, representing being nervous but with intensification along the *Stressed* scale, stable stress, and resolving stress respectively. When looking at transition directions for each word used by each participant, we found that all three possibilities appear for most words, except in cases where an emotion word is mentioned only once or twice (such that it could not be associated with three distinct directions). Observed distributions were $[\mu, \sigma]$: $[2.76, 0.53]_{5s}$ and $[2.96, 0.22]_{0.5s}$).

Figure 2 exemplifies the hierarchical process with two streams of self-reported emotion. Where window boundaries do not coincide with a logged data point, we imputed with the previous data point, turning our time-series into a higher-resolution stepped signal. We resolved windows containing multiple labels by using mode for the Calibrated Words and the slope of the best fit line in the continuous annotation.

Table I: Full list of Calibrated Words used by at least one Participant in their TwCW.

Calibrated Word	Number of Participants	Calibrated Word	Number of Participants
Anxious	15	Confused*	11
Frustrated	15	Curious	11
Dread	14	Resigned	10
Indifferent*	14	Threatened	8
Satisfied	14	Annoyance*	5
Hopeful	13	Resolve*	4
Accomplished	12	Excited*	3
Alert	12	Clueless*	1
Cautious	12	Triumph*	1

Participants used 11 of the 13 provided words (none spoke of feeling *Fearful* nor *Nervous* during the interview stage so both are omitted). Starred * words are participant-generated write-ins.

B. Force Sensitive Resistor (FSR) Data

FEEL's keypress force (FSR) data exhibited an average of <1 distinct keystrokes per window, contraindicating deep learning models. We extracted features from keystroke activity, frequency, and statistical analysis, generated data instances aligned with brain activity (0.5s and 5s), and performed model selection with classical machine learning models.

Data Preparation: To mitigate FSR signal noise while maintaining the overall shape of a keystroke, we applied an Exponentially Weighted Moving Average (EWMA) [50] with smoothing factor $\alpha = 0.5$. We aggregated game keypress activity from the original game-control keys (denoted A0-A4 in the dataset) into two additional channels as '**composite keys**', computing over all keys the force sum (A5) and maximum (A6), resulting in a total of 7 keypress channels.

Frequency and statistical features: Based on previous studies of emotion expression of social touch pressure [15], we calculated a set of descriptive statistics for each window of pressure data – minimum, maximum, variance, mean, area under the curve, and sum of absolute differences. From the same windows, we calculated the most prominent frequency

(amplitude and frequency bin), amplitude variance, amplitude mean, and peak count for frequency-domain features [15].

Keystroke features: Since participants activated keys based on gameplay rather than typing, certain features of keystroke dynamics – such as travel time between keys – are less relevant here. We therefore calculated touch features highlighting fluctuations in force and duration in both time and frequency domains [15], [51]. We also borrowed parameters related to the Attack Decay Sustain Release (ADSR) envelope [52], commonly employed in synthesizers to describe piano keyboard output. For each keystroke in a window, we calculated: keystroke duration (in ms), peak count, amplitude of maximum peak, time from keystroke start to maximum peak, time from maximum peak to key release, force variance, average force, and area under the keypress curve. Parameters are aggregated by taking the mean over each data window.

For the purposes of multi-modal window alignment and the simulation of real-time application of emotion classification on keypress force, we used uniform data windowing. However, we note that distortion may occur where keystrokes cross window boundaries.

C. EEG Data

We calculated Differential Entropy (DE) for the 5 frequency bands demonstrating activity during emotion expression [20], [53]: $\boldsymbol{\delta}$ (1-4Hz), $\boldsymbol{\theta}$ (4-7Hz), $\boldsymbol{\alpha}$ (8-12Hz), $\boldsymbol{\beta}$ (12-30Hz) and $\boldsymbol{\gamma}$ (30-50Hz). For each band, we calculated the difference between channel pairs to create a 2D Asymmetrical Map (AsMap) feature [20]. The resulting feature is an image with size 64×64 and a depth of 5 frequency bands.

D. Classification Model Implementation

To compare EEG- *vs.* FSR-based models classifying emotion transition, we ran 30 iterations of 5-fold cross-validation (training and validation sets randomized every iteration). Figure 3 summarizes the overall experimental pipeline.

FSR: We performed grid search cross-validation (CV) ($k = 5$) to select the best-fit model by participant among seven machine learning models. Due to the sparseness of the FSR data (low sampling rate with some keys pressed in only a few brief instances), we elect to compare performance across Extra Trees, Random Forest, AdaBoost, Gradient Boosting, XGBoost, Logistic Regression, and SVM [47], [48], [54], options that are more amenable to the size and scale of this data than deep learning models. Given the high dimensionality of our feature set ($d = 82$ features per participant), we selected features by employing a zero variance threshold to remove all constant-valued features and use recursive feature elimination (RFECV) [55] with CV ($k = 5$). We report mean test scores over the 16 participants after 30 trials using the best-fit model for each, with a 70/30 training-test split ratio.

EEG: We used a CNN model with a 2D feature set to take advantage of the automated learning demonstrated by deep-learning models. In the interests of balancing model complexity with overfit risk [56], we implemented the structure proposed by Ahmed *et al.* (2022) [20] – a 2-layer CNN using 3x3 kernels and 2 Max Pooling layers – for affect classification, adjusting the input size to $5 \times 64 \times 64$ to account for the size of our features. We created train and test sets using a 50/50 split ratio. Figure 4 summarizes the CNN architecture.

We performed grid search CV ($k = 5$) on the train set to tune the number of epochs (5, 10, 20), the batch size (128, 256, 512) and the learning rate (10^{-3}, 10^{-4}, 10^{-5}) to select the best participant-specific hyperparameters for our model. Larger epoch sizes (≥ 100) were omitted from the search space since similar training performances were observed, while being resource intensive. Once we obtained the parameters that maximized the macro-hierarchical F1-scores [46], we trained the participant-specific model 30 times on the full training set, each time using the unseen test set to calculate performance metrics. We report mean test scores from the 30 runs.

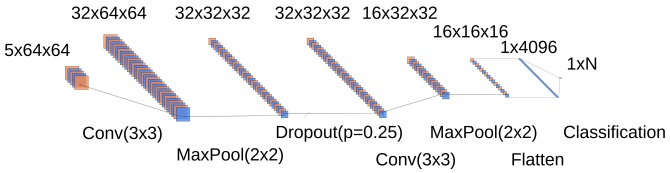

Figure 4: Structure of the EEG CNN model for classification where each convolution layer uses a 3×3 kernel (of depth 32 and 16 respectively) followed by a ReLU activation function. The inputs to the model are the $5 \times 64 \times 64$ AsMap features [20], while the output is the class output ($N = 3$).

V. Classification Performance by Modality

We analyzed macro hierarchical F1-scores [46] by model and window size (Table II) finding that classification performance monotonically increases with window size. For brevity, we report in depth on 0.5s and 5s windows. With two modalities and two window sizes, our data does not pass Levene's test for equality of variances (F(3,1916)=51.0, $p < 0.001$), so we report results using a two-way aligned rank transform analysis of variance (ART ANOVA), implemented with R's ARTool [57]. All reported effects are statistically significant at $p \leq 0.001$. The main effects of affective modality (M) and window size (W), and interaction effect (W/M) yield F ratios of $F_M(1, 1916) = 5283.98$ ($\eta_p^2 = 0.733$), $F_W(1, 1916) = 56.88$ ($\eta_p^2 = 0.028$), and $F_{W/M}(1, 1916) = 285.26$ ($\eta_p^2 = 0.130$).

Table II: Hierarchical classification scores for each (W)indow / (M)odality where the best combination is **5s-FSR**. All W/M models exceed chance by ∼2-4x.

W/M	F1-Score	Precision	Recall
5s EEG	0.415 ± 0.110	0.415 ± 0.109	0.422 ± 0.123
0.5s EEG	0.494 ± 0.070	0.544 ± 0.118	0.469 ± 0.090
0.5s FSR	0.686 ± 0.039	0.681 ± 0.036	0.682 ± 0.037
5s FSR	**0.823 ± 0.012**	0.827 ± 0.013	0.825 ± 0.013
0.5s chance	0.215 ± 0.010	0.215 ± 0.010	0.215 ± 0.010
5s chance	0.216 ± 0.009	0.216 ± 0.009	0.216 ± 0.009

Scores are calculated over 480 hierarchical metrics (16 participants × 30 trials, average macro hierarchical F1 taken over all classes).

We ran post-hoc tests using a Holm correction to further investigate the individual mean differences in Table II (significance at $p_{\text{Holm}} \leq 0.001$ unless indicated). Results show

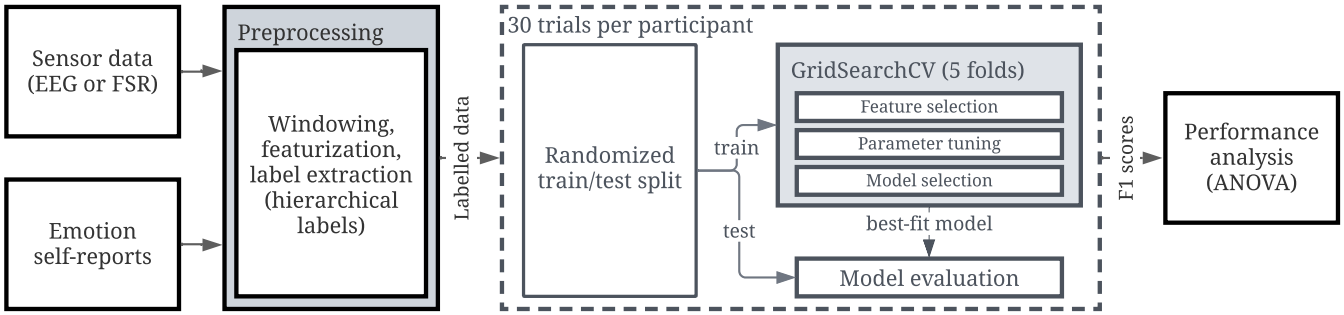

Figure 3: Pipeline for model selection and evaluation. We performed grid search CV ($k = 5$) on the training set to tune hyperparameters and select best-fit models for FSR data. The models were then evaluated on an unseen test set to calculate performance metrics. We repeated this process 30 times per participant, and report mean test scores across the 30 runs and 16 participants.

that (1) mean F1-score was significantly greater for FSR-based models than EEG-based models; (2) mean F1-score increased with window size, with 5s windows performing strongest across modalities; and (3) FSR at 5s windows performed best overall. Additionally, we found that the chosen CNN parameters for batch size and epochs tend to differ by participant, while the learning rate remained stable. For 0.5s, the optimal parameters by participant were seen for batch sizes of μ=160.0, σ=96.0 and training epochs of μ=13.8, σ=6.5; for 5s, μ=248.0, σ=139.0 batch size and μ=11.6, σ=6.1 epochs. In all cases, loss curves stabilize by 15 epochs, suggesting diminishing returns in classification performance with additional training epochs.

VI. FSR Feature Analysis

For insight on how features inform classification, we ran RFECV on the feature set of both FSR models (0.5s and 5s windows), and grouped selected features by type – pressure-based (direct measures of keypress force), time-based (measures of duration), and frequency-based (FFT-based features). We analyzed model performance using F1-score for feature group. Figure 5 summarizes the top performing feature groups.

Our data for both models (0.5s and 5s) again does not pass Levene's test for equality of variances ($F_{0.5s}(2,39356)=831.74$, $p_{0.5s} < 0.001$; $F_{5s}(2,39356)=906.32$, $p_{5s} < 0.001$) with three feature groups, so we report F1-scores after two one-way ART ANOVA for each window size. Main effects of both tests are statistically significant at $p \leq 0.001$ significance, yielding F ratios of $F_{0.5s}(2, 39356) = 1049.3$ ($\eta_p^2 = 0.051$) and $F_{5s}(2, 39356) = 761.97$ ($\eta_p^2 = 0.037$), respectively.

To investigate individual mean differences, we ran post-hoc tests using a Holm correction significance at $p_{\text{Holm}} < 0.001$ unless otherwise indicated. The mean F1-score was significantly greater for models that rely on **pressure-based features, for both window sizes**, followed by time-based features.

VII. Discussion and Future Work

Here, we reflect on our research question, and how our findings can inform the use of touch pressure data in modelling dynamic emotions and contribute to the development of emotionally responsive devices.

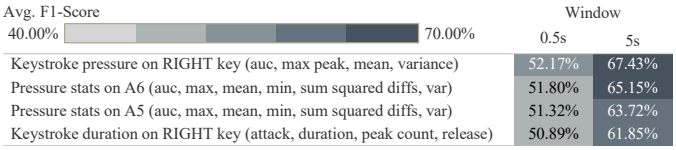

Figure 5: Relative feature performance by window size. Darker cells indicate frequent selection of better-performing features. The RIGHT directional key is used to advance the character – and game storyline – through the side-scrolling game. A5 corresponds to the sum of the pressure across all keys, while A6 corresponds to the max force over all keys.

A. Real-Time Predictors of Dynamic Emotion

Longer Time Windows Favour Keypress Force: For personalized classification models of evolving Stress built on participants screened for high emotion clarity, FSR models perform better than those built on continuous EEG for both window sizes we analyzed. Individualistic emotion evolution inherent in real life events, particularly when reflecting or reacting to memory retrieval, may require more than 0.5s [29], [58]. We posit that longer windows will better capture lower-frequency information and thus benefit manual keyboard interactions for models of keypress force, but may blur the picture of higher-frequency brain activity features [38], [44].

Manual touch pressure encodes valuable emotion content: Our feature extraction techniques were informed by analyses of a variety of affective touch interactions: keystroke dynamics in typing behaviour [9], pressure and location features from social touch [15], and ADSR features from sounds produced from a music keyboard [52]. Feature evaluation reveals that of the 20 most important features from all three domains, 16 are pressure- or force-related. Increases in typing force were previously known to correlate with higher stress experiences [8], and machine-mediated social touch [51] has been differentiated by variations in pressure. Now, we have evidence that Stressed-scale emotion expression can also be captured implicitly through keypress force using an easily modified videogame keyboard. We continue to investigate other contexts and emotion scales where we subconsciously express emotion via touch pressure, leaving dimensional examination of dynamic emotion evolution and touch pattern correlates to

future work. In the meantime, we posit that the information available by tracking pressure in devices where interactions feature manual affective touch outweigh the cost of adding this functionality.

The case for emotion transitions – timing matters: When modelling human emotions, we may consider how the emotion space changes over time: when we feel sad, it may be easier to get angry than calm, despite these emotions being separated by comparable Euclidean distances on the Affect Grid [26]. An emotion experience can feel more like a trajectory over a constantly changing landscape than a point [25]. After studying the evolution of *stress*, we infer that predicting direction of an emotion trajectory may be particularly important when delivering interventions for emotion regulation. For example, strategies may differ for the *onset* of anger *vs.* after rage has *cooled* [32].

B. Building Effective Models for Dynamic Emotion Prediction

Potential confounds: First, we point out that there are a number of potentially confounding factors, including (but not limited to): participant interest in, and proclivity for, this video game genre; fluctuations in skin conductivity; extraneous motion; and cognition in action planning; personal experiences of Stress-Relative emotions; individual differences in the ability to express, appraise, and resolve emotions. We minimized these limitations through participant screening, personalized word calibration, multipass data labelling for richer experience capture, and individualized emotion classification models. However, they may still have influenced the reported classification performance.

Modality capture: The collection of this dataset was time-intensive and effortful, in large part due to setup and calibration of the EEG data collection system. Given EEG signal sensitivity to surrounding conditions as well as collection effort and intrusiveness, the comparable-to-better classification performance of FSR signals for emotions adjacent to the Stressed-Relaxed scale means that under certain conditions – *e.g.,* slower evolution as for Stress, emotion reflection tasks requiring appraisal or memory retrieval [58], low compute and/or time resources, or prioritization of personalized over general models – we are hereby able to recommend reliance on, or the addition of, keypress force or other manual touch data for emotion interaction.

Labelling effort: Collecting multipass emotion self-reports affords rich triangulation of a numerical emotion rating onto personalized emotion scales. But it also incurs a time cost: altogether, personalized calibration, emotion elicitation, interview, and continuous annotation take 3 to 4 times as long as the emotion elicitation task alone. Where tasks run long, multipass reviewing procedures require careful consideration to ensure annotation can occur contemporaneously without interfering with the natural evolution of the emotional experience.

Emotion elicitation and affect scale: Calibrating how users placed emotion words on a Relaxed-Stressed scale allowed us to simultaneously pool data and personalize models. While participants had personalized understandings of the measurement scale, they all engaged in the same emotion elicitation experience (a horror video game). For personalized models to work "in the wild", they must be built on participant-defined emotion experiences that evolve longitudinally and spontaneously. Human emotional experience is ever-evolving; so also must be the calibrated scales, training data, and accompanying models across multiple named emotions and touch interaction patterns. Future work examines how longitudinal calibration can trace evolution of emotion models over multiple data collection sessions.

Context Matters in Personalized Emotion Models: A deployed model could face a wide range of priorities. Naturalness of a responsive agent may value minimal latency over accuracy. In other situations, some scenarios may be more important to capture accurately ('something's wrong') than others ('everything's fine'). Machine learning accuracy metrics are useful for comparing performance, but for contextually effective machine responses, new metrics may be necessary to reflect the nuances of the overall experience.

VIII. Conclusion

We present the FEEL dataset, the first of its kind: affective multimodal data (brain activity and keypress force estimated by EEG and FSR) collected during an emotional videogame experience and labelled using a multipass emotion self-report described by [6] – resulting in multi-timescale, and personally calibrated emotion labels rooted on the Stressed-Relaxed scale. This paper describes the dataset and the specifics of its collection, and demonstrates participant-dependent machine learning classification performance differentiating emotions in **transition** – *e.g.,* whether one's *stress* is growing or resolving, benchmarked here at $F1 = 0.82$ at the best case (chance $F1 = 0.22$, $\sigma = 0.01$). We invite the community to explore other computational strategies and advance the exploration into dynamic emotion classification.

Comparing classification performance over factors of window size, feature set, and modality, we find that, overall:

1) Window sizes influence recognition behaviour for both brain activity and touch pressure, the choice of which depends on intended observation (longer windows are better able to capture slower changes but shorter windows can capture high frequency activity)
2) Feature evaluation of the FSR feature set reveals that pressure features used in machine-mediated social touch rank highest in terms of selection frequency.

From these findings, we propose that emotion interaction systems should (1) consider window size in labelling; and (2) improve emotion recognition opportunities by incorporating pressure sensors where manual human touch is enacted.

Acknowledgments

Thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC) for funding this work. Human user research was conducted under UBC Ethics #H15-02611.

REFERENCES

[1] B. App *et al.*, "Nonverbal channel use in communication of emotion: how may depend on why.," *Emotion*, vol. 11, no. 3, p. 603, 2011.

[2] S. Yohanan and K. E. MacLean, "The role of affective touch in human-robot interaction: Human intent and expectations in touching the haptic creature," *Int'l J of Social Robotics*, vol. 4, no. 2, pp. 163–180, 2012.

[3] D. Silvera-Tawil, D. Rye, and M. Velonaki, "Interpretation of social touch on an artificial arm covered with an eit-based sensitive skin," *Intl J of Social Robotics*, vol. 6, no. 4, pp. 489–505, 2014.

[4] M. J. Hertenstein *et al.*, "Touch communicates distinct emotions.," *Emotion*, vol. 6, no. 3, p. 528, 2006.

[5] Playdead-Denmark, "Playdead's inside." https://playdead.com/games/inside/, 2022. Accessed: 2022-08-21.

[6] X. L. Cang *et al.*, "Choose or fuse: Enriching data views with multi-label emotion dynamics," in *IEEE 10th Int'l Conf on Affective Computing & Intelligent Interaction (ACII)*, 2022.

[7] S. Alarcao and M. Fonseca, "Emotion recognition using eeg signals: A survey," *IEEE Trans on Affective Computing*, vol. 10, no. 3, 2017.

[8] J. Hernandez, P. Paredes, A. Roseway, and M. Czerwinski, "Under pressure: sensing stress of computer users," in *Proc of the SIGCHI Conf on Human factors in computing systems*, pp. 51–60, 2014.

[9] H.-R. Lv, Z.-L. Lin, W.-J. Yin, and J. Dong, "Emotion recognition based on pressure sensor keyboards," in *2008 IEEE Int'l Conf on multimedia and expo*, pp. 1089–1092, IEEE, 2008.

[10] M. J. Hertenstein *et al.*, "The communicative functions of touch in humans, nonhuman primates, and rats: a review and synthesis of the empirical research," *Genetic, social, & general psychology monographs*, vol. 132(1), 2006.

[11] T. Kinnunen and M. Kolehmainen, "Touch and affect: Analysing the archive of touch biographies," *Body & Society*, vol. 25(1), 2019.

[12] M. Leng, Y. Zhao, and Z. Wang, "Comparative efficacy of non-pharmacological interventions on agitation in people with dementia: A systematic review and bayesian network meta-analysis," *Int'l J of Nursing Studies*, vol. 102, p. 103489, 2020.

[13] T. Thanapattheerakul *et al.*, "Emotion in a century: A review of emotion recognition," in *Int'l Conf on advances in information technology*, 2018.

[14] N. J. Shoumy *et al.*, "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals," *J of Network and Computer Applications*, vol. 149, p. 102447, 2020.

[15] X. L. Cang *et al.*, "Discerning affect from touch and gaze during interaction with a robot pet," *IEEE Trans on Affective Computing*, vol. Early Access, no. 01, pp. 1–1, 2021.

[16] J. S. Kumar and P. Bhuvaneswari, "Analysis of electroencephalography (eeg) signals and its categorization–a study," *Procedia engineering*, vol. 38, pp. 2525–2536, 2012.

[17] G. A. Rousselet, "Does filtering preclude us from studying erp time-courses?," *Frontiers in psychology*, vol. 3, p. 131, 2012.

[18] C. Deveney and D. Pizzagalli, "The cognitive consequences of emotion regulation: an erp investigation," *Psychophysiology*, vol. 45(3), 2008.

[19] P. Verduyn, P. Delaveau, J.-Y. Rotgé, P. Fossati, and I. Van Mechelen, "Determinants of emotion duration and underlying psychological and neural mechanisms," *Emotion Review*, vol. 7, no. 4, pp. 330–335, 2015.

[20] M. Z. I. Ahmed, N. Sinha, S. Phadikar, and E. Ghaderpour, "Automated Feature Extraction on AsMap for Emotion Classification Using EEG," *Sensors*, vol. 22, p. 2346, Mar. 2022.

[21] J. J. Gross, "Antecedent-and response-focused emotion regulation: divergent consequences for experience, expression, and physiology.," *J of personality and social psychology*, vol. 74, no. 1, p. 224, 1998.

[22] A. Moors *et al.*, "Appraisal theories of emotion: State of the art and future development," *Emotion Review*, vol. 5, no. 2, pp. 119–124, 2013.

[23] B. Mesquita, L. F. Barrett, and E. R. Smith, *The mind in context*. Guilford Press, 2010.

[24] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. Cambridge university press, 1990.

[25] P. Bucci *et al.*, "Real emotions don't stand still: Toward ecologically viable representation of affective interaction," in *IEEE Int'l Conf on Affective Computing & Intelligent Interaction (ACII)*, pp. 1–7, 2019.

[26] J. A. Russell, "A circumplex model of affect.," *J of Personality and Social Psychology*, vol. 39, no. 6, p. 1161, 1980.

[27] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales.," *J Personality & Social Psychology*, vol. 54, no. 6, p. 1063, 1988.

[28] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *J of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[29] P. Verduyn, "Emotion duration," in *Affect Dynamics*, pp. 3–18, Springer, 2021.

[30] M. F. Jung, "Affective grounding in human-robot interaction," in *12th Int'l Conf on Human-Robot Interaction (HRI*, pp. 263–273, IEEE, 2017.

[31] L. F. Barrett, B. Mesquita, K. N. Ochsner, and J. J. Gross, "The experience of emotion," *Annu. Rev. Psychol.*, vol. 58, pp. 373–403, 2007.

[32] G. Sheppes and J. J. Gross, "Is timing everything? temporal considerations in emotion regulation," *Personality and Social Psychology Review*, vol. 15, no. 4, pp. 319–331, 2011.

[33] J. B. Torre and M. D. Lieberman, "Putting feelings into words: Affect labeling as implicit emotion regulation," *Emotion Review*, vol. 10, no. 2, pp. 116–124, 2018.

[34] B. Dudzik and J. Broekens, "A valid self-report is never late, nor is it early: On considering the "right" temporal distance for assessing emotional experience.," in *2nd Momentary Emotion Elicitation & Capture Workshop at CHI*, 2021.

[35] T. Ritchie, J. J. Skowronski, J. Hartnett, B. Wells, and W. R. Walker, "The fading affect bias in the context of emotion activation level, mood, and personal theories of emotion change," *Memory*, vol. 17, no. 4, pp. 428–444, 2009.

[36] P. V. Rouast, M. T. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Trans on Affective Computing*, vol. 12, no. 2, pp. 524–543, 2019.

[37] P. Salovey, J. D. Mayer, S. L. Goldman, C. Turvey, and T. P. Palfai, "Emotional attention, clarity, and repair: exploring emotional intelligence using the trait meta-mood scale.," *Emotion, disclosure, & health*, pp. 125–154, 1995.

[38] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Trans on autonomous mental development*, vol. 7, no. 3, 2015.

[39] K. Alarabi Aljribi, "A comparative analysis of frequency bands in eeg based emotion recognition system," in *The 7th Int'l Conf on Engineering & MIS 2021*, 2021.

[40] A. Delorme, "Eeg is better left alone," *Scientific reports*, vol. 13, no. 1, p. 2372, 2023.

[41] C. N. Silla and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Mining and Knowledge Discovery*, vol. 22, pp. 31–72, 2011.

[42] E. Tieppo, R. R. d. Santos, J. P. Barddal, and J. C. Nievola, "Hierarchical classification of data streams: a systematic literature review," *Artificial Intelligence Review*, pp. 1–40, 2022.

[43] P. Verduyn, I. Van Mechelen, and F. Tuerlinckx, "The relation between event processing and the duration of emotional experience.," *Emotion*, vol. 11, no. 1, p. 20, 2011.

[44] X.-W. Wang, D. Nie, and B.-L. Lu, "Emotional state classification from EEG data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94–106, Apr. 2014.

[45] M. Li and B.-L. Lu, "Emotion classification based on gamma-band EEG," in *2009 Annual Int'l Conf of the IEEE Engineering in Medicine and Biology Society*, pp. 1223–1226, Sept. 2009. ISSN: 1558-4615.

[46] F. M. Miranda, N. Köehnecke, and B. Y. Renard, "Hiclass: a python library for local hierarchical classification compatible with scikit-learn," *arXiv preprint arXiv:2112.06560*, 2021.

[47] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[48] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc of the 22nd acm sigkdd Int'l Conf on knowledge discovery and data mining*, pp. 785–794, 2016.

[49] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019.

[50] J. S. Hunter, "The exponentially weighted moving average," *J of quality technology*, vol. 18, no. 4, pp. 203–210, 1986.

[51] M. M. Jung *et al.*, "Touching the void–introducing cost: corpus of social touch," in *Proc of the 16th Int'l Conf on Multimodal Interaction*, pp. 120–127, 2014.

[52] K. Jensen, "Envelope model of isolated musical sounds," in *Proc of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99)*, vol. 12, Citeseer, 1999.

[53] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for eeg-based emotion classification," in *2013 6th Int'l IEEE/EMBS Conf on Neural Engineering (NER)*, pp. 81–84, IEEE, 2013.

[54] M. M. Jung, "Towards social touch intelligence: developing a robust system for automatic touch recognition," in *Proc Intl Conf on Multimodal Interaction*, pp. 344–348, 2014.

[55] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1, pp. 389–422, 2002.

[56] J. Chen, P. Zhang, Z. Mao, Y. Huang, D. Jiang, and Y. Zhang, "Accurate eeg-based emotion recognition on combined features using deep convolutional neural networks," *IEEE Access*, vol. 7, pp. 44317–44328, 2019.

[57] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proc of the SIGCHI Conf on human factors in computing systems*, pp. 143–146, 2011.

[58] T. Staudigl, S. Hanslmayr, and K.-H. T. Bäuml, "Theta oscillations reflect the dynamics of interference in episodic memory retrieval," *Journal of Neuroscience*, vol. 30, no. 34, pp. 11356–11362, 2010.