Interspeech 2024
1-5 September 2024, Kos, Greece

Predicting Acute Pain Levels Implicitly from Vocal Features

Jennifer Williams[1], *Eike Schneiders*[2], *Henry Card*[1], *Tina Seabrooke*[1], *Beatrice Pakenham-Walsh*[1], *Tayyaba Azim*[1], *Lucy Valls-Reed*[1], *Ganesh Vigneswaran*[1], *John Robert Bautista*[3], *Rohan Chandra*[4], *Arya Farahi*[4]

[1]University of Southampton, UK; [2]University of Nottingham, UK;
[3]University of Missouri-Columbia, USA; [4]University of Texas at Austin, USA

j.williams@soton.ac.uk

Abstract

Evaluating pain in speech represents a critical challenge in high-stakes clinical scenarios, from analgesia delivery to emergency triage. Clinicians have predominantly relied on direct verbal communication of pain which is difficult for patients with communication barriers, such as those affected by stroke, autism, and learning difficulties. Many previous efforts have focused on multimodal data which does not suit all clinical applications. Our work is the first to collect a new English speech dataset wherein we have induced acute pain in adults using a cold pressor task protocol and recorded subjects reading sentences out loud. We report pain discrimination performance as F1 scores from binary (pain vs. no pain) and three-class (mild, moderate, severe) prediction tasks, and support our results with explainable feature analysis. Our work is a step towards providing medical decision support for pain evaluation from speech to improve care across diverse and remote healthcare settings.
Index Terms: speech paralinguistics, health, medical conversations, pain assessment, speaker states

1. Introduction

The perception and management of pain, which encompasses neurological, psychological, and cognitive dimensions, poses a significant challenge in the medical field. Pain perception transcends mere sensory inputs, integrating emotional and cognitive aspects that markedly influence its intensity and character [1]. Moreover, the complexity of pain perception is exacerbated by neuroplastic changes following surgical interventions, highlighting the necessity to understand pain modulation mechanisms for efficacious management [2]. Although patients frequently possess the capability to express and explicitly communicate their pain verbally, this task becomes particularly challenging for individuals unable to do so, such as those affected by stroke [3], learning difficulties [4], autism [5], and young children [6]. For these patients, the identification of objective, non-invasive indicators of pain is crucial to ensure the provision of timely and appropriate care. Even when a patient is able to communicate with the clinician, their self-reported pain ratings reflect individual, subjective experiences of pain, making them difficult to quantify and assess reliably [7].

Historically, machine learning models have provided valuable insights into pain assessment across various contexts [8], utilising methodologies such as electromyography (EMG), electroencephalogram (EEG), skin conductance, and electrocardiography (ECG) to predict pain levels and outcomes. Recent advancements have even leveraged wearable technologies (e.g. Apple Watch) for continuous monitoring [9], facilitating early interventions in conditions like sickle cell disease. The development of personalized pain assessment offers a novel and relatively untapped opportunity. With the advent of wearable technologies, speech, replete with digital biomarkers and distinct 'Digital Biomarker Fingerprints,' can provide unique insights into individual health states [10]. This approach to healthcare, which leverages the ubiquitous nature of audio data, holds the promise of identifying personalized health characteristics through these digital fingerprints. The feasibility of this method has been demonstrated, albeit preliminarily, in recent work [11, 12, 13], which used audio analysis for assessing different types of pain in various languages.

A critical step toward trustworthy pain decision support tools requires understanding which kinds of acoustic features to exploit in order to improve the accuracy of pain assessment. This knowledge has wide application in the medical domain [14], from analgesia administration to medication titration. Healthcare professionals, through established rapport and ongoing patient interactions, are adept at discerning pain through verbal and nonverbal cues. However, recent shifts towards telephonic and video consultations have diminished the efficacy of these cues, underscoring the need for a reliable, distance-compatible decision-support tool. Our exploration of pain assessment through speech processing contributes to this enduring problem, promising to bridge the gap in patient care, particularly in distance-led (e.g. telephone/video conference) health settings.

One of the main research gaps that this paper addresses is that there are no existing English speech datasets available for acute pain assessment in adults that are labeled with clinically-relevant pain scales [15]. In this paper, we make the following contributions:

- Present a new English speech-only pain dataset from acute pain inducement and identify functional acoustic features that correlate with reported pain.
- Utilize these features with discriminative machine learning to predict pain levels in two tasks: presence/absence of pain and mild/moderate/severe pain.
- Show how different functional features contribute to pain level assessment.

2. Related Work

Previous work on pain assessment has spanned multiple disciplines and input modalities including functional near-infrared spectroscopy (fNIRS) [16], and electroencephalography (ECG) data [17]. It has also been viewed from the standpoint of 'affective computing' [18]. However, individuals express pain differently and features beyond clear affect must be considered, including speech rate, breathiness, mispronunciation, restarts, etc. Reliable detection of pain from key acoustic features remains unsolved. More recently, there was a renewed call for

10.21437/Interspeech.2024-15

research efforts to develop AI tools that support clinicians with pain assessment from bioacoustic markers in speech [19] and our work takes a step forward towards that goal.

Automated pain recognition has invited two streams of research: pain detection which treats the problem as a binary classification task, and pain intensity assessment as a multi-class classification or regression task. Another research dimension in this domain is the use of unimodal input [13, 20] versus multiple modalities (audio, video, physiological) [21]. Although speech and sound have proven very effective for analysing the expression of emotions at an early stage of human development (e.g., crying infants [22]), there remains much to be explored for adult pain assessment, especially for real-time applications. With increasing use of telehealth delivery and emergency calling, there is a growing need for speech-based research on pain assessment.

The first to report findings of bioacoustic markers of pain from the speech signal was [11]. They developed a proof-of-concept solution to detect the presence or absence of pain using speech from interviews of adult hospital patients suffering various acute and chronic conditions. While their sample size was quite small (400 speech instances), they derived two new types of features from low-level descriptors (LLDs) and reported promising results using support vector machines on a binary prediction task. Further work from [12] classified pain assessment for adults from speech using a combination of prosodic features, mel frequency cepstral coefficients (MFCCs) and deep neural network bottleneck features. Their best reported balanced accuracy was 74.2% on a binary task (mild/severe) and 54.2% on a three-class task (mild/moderate/severe).

Similar to our work in this paper, [13] collected a German speech pain dataset based on a cold pressor task to induce acute pain. However, it was an uncontrolled collection and limited only to German speech, making explainability difficult and generalisation across languages uncertain. They reported best performance on three-class pain prediction with 42.7% unweighted average recall, which is only slightly better than chance given the imbalance of pain levels in their dataset. They used acoustic features, MFCCs, and deep spectrum VGG16 features [23] on mel frequency spectrograms. Our work addresses a similar problem, but using clinically-relevant pain levels with our newly collected English speech pain dataset where acute pain was induced by a cold pressor task. Our work includes binary pain detection as well as three-class assessment with promising results using functional acoustic features. Our work also contributes new understanding towards explainability about how different acoustic features influence pain classification.

3. Data Collection

We followed a protocol common to the discipline of psychology to induce acute pain in adult human subjects using very cold water (0-4°C) in a Cold Pressor Task (CPT) [24, 25] while eliciting and recording speech[1]. Pain level ground truth was assumed, as in clinical contexts, that subjects reported pain accurately according to their experience and individual perspectives.

3.1. Recruitment and Compensation

Participants were 15 undergraduate (12 female, 3 male, mean age: 18.73, sd: 0.96) Psychology students from the United Kingdom who participated in return for partial course credit[2]. Participants were excluded from the study if they reported: high blood pressure; a heart or circulation problem; dysthymia; a cardiovascular disorder; history of Raynaud's syndrome, fainting, seizures, or frostbite; an open cut, sore, or bone fracture on or near to either hand; a neurological disorder; diabetes; epilepsy; or pregnancy. Participants were required to be aged between 18 and 35 years. Participants were provided with details of the study procedure, purpose, and their rights, including the right to withdraw from the study or to end each trial by removing their hand from the water.

3.2. Pain Inducement: The Cold Pressor Task

The CPT is a commonly used task for the induction of discomfort to mild pain [26, 27]. Participants were asked to immerse their hand into cold water (0-4°C) while performing a speaking task (Section 3.3). To provide a control condition and collect data for the no-pain condition, we included a warm water (34-37°C) condition. To minimize risk to the participants, and in accordance with the ethical guidelines, participants immersed their hand for a maximum duration of up to three minutes – or until they withdrew their hand from the water [26]. To prevent the buildup of micro climate within the water tanks, the water was circulated with a water pump (5.8 L/h).

Participants were randomly assigned to one of four groups, varying the order of which hands, left (L) or right (R), was placed into the cold (C) or warm (W) water. The four experimental groups were: (1) LC-LW-RC-RW, (2) LW-LC-RW-RC, (3) RC-RW-LC-LW, and (4) RW-RC-LW-LC. Prior to the first trial, the participant's hand temperatures were recorded using an infrared thermometer, providing a baseline to ensure that their hand temperatures could be brought back to their specific baseline following the final trial. Participants submerged their entire hand in the water, with the palm facing upwards. After every instance of cold water exposure, participants placed that same hand into a nearby warm water bath to bring their hand up to baseline temperature to minimize discomfort and to proceed with the next trial. Following the experiment, participants were debriefed and provided with a take-home debriefing sheet.

3.3. Speech Elicitation and Collection

During the CPT protocol for both warm and cold water immersion (water tubs were side-by-side to minimize movement around the room), participants read aloud sentences from a randomized selection of Harvard Sentences[3] [28] that were presented on 50" screen approximately 2 meters away. Every sixth sentence required them to say a pain statement to register their pain level on a 1-10 scale, as follows: *"On a scale from 1–10, the pain I feel right now is..."*. The order of the sentences were randomized per subject and manually advanced by the research team, to allow speakers to read the sentences at their own pace. Speech was collected from two microphones: (1) Røde Wireless PRO close-talking mic with lapel placed approximately 10" from each speaker's chin (primary), and (2) Blue Yeti desktop microphone placed approximately 20" from the speaker. All audio was collected as one utterance per wave file, as 16-bit mono PCM 16 kHz. The data collection room was approximately 10x10'. In total, from the 15 participants, we collected

[1]This study was approved by the University of Southampton ethics board, reference ERGO 80074.A1

[2]The bias towards females reflects recruitment from a Psychology undergraduate program predominately comprised of females.

[3]https://www.cs.columbia.edu/~hgs/audio/harvard.html

1,518 female utterances (avg: 126.5 sd: 27.5) and 429 male utterances (avg: 143, sd: 6.4). Due to the variability across genders and very small sample of male utterances, in our experiments (Section 5.3) we report results for female-only and mixed gender. We used data from the primary (Røde) lapel mic.

3.4. Ethics

Automatic pain discrimination has the potential to be misused outside of the medical domain (e.g., using pain levels for torture or abuse) while it can also provide decision support tools to legitimate health professionals. Our data collection and research activities aligned with a Responsible Research and Innovation (RRI) approach [4] called the AREA (Anticipate, Reflect, Engage, Act) framework [29] to guide our ethics application and to determine whether this work should be undertaken. We followed the ethical guidelines of our institution precisely and we adhere to all applicable regulations regarding data sharing.

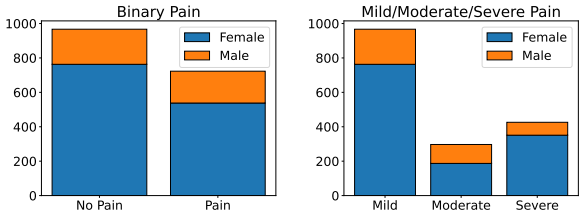

Figure 1: *Distribution of utterances labeled for pain in the binary and three-class discrimination tasks, indicating distribution of male and female utterances.*

4. Speech Data Pre-Processing

All 1,947 utterances were manually reviewed by the authors for quality control. We removed 225 utterances wherein: a speaker's words were cut off, other speakers were present in the audio, noise interference was present including noise from outside of the recording room, from the CPT equipment, the speaker bumping the microphone against the water tubs or the subject shifting in an audible manner.

The pain statements (Section 3.3) were used to label utterances by manually extrapolating the reported pain level backwards for the previous five utterances, allowing us to label every utterance with the subject's self-reported pain level. If the pain statement was not available due to our quality control procedure, then we removed the previous sentences that remained unlabeled. For pain levels wherein subjects reported two pain levels (e.g., *"between 5 and 6"* or *"8.75"*) we rounded up to the nearest integer. Likewise, if a subject reported a pain of 0, this was re-labeled as 1 since we used a pain scale of 1-10.

All recordings were trimmed using voice activity detection (VAD) to remove any leading and trailing silence using the Python *webrtcvad* toolkit with the lowest aggressiveness setting[5] and we removed 14 utterances where the duration was less than 1.5s. Our final dataset[6] contained 1,690 utterances and 76.88 minutes of data of which 1,301 utterances were from female speakers (duration avg: 2.69s, sd: 0.45) and 389 utterances (duration avg: 2.84s, sd: 0.45) were from male speakers.

5. Pain Discrimination Experiments

From Figure 1, we adjusted the reported pain levels p in our dataset to align with two discriminative tasks: presence or absence of pain ("binary task") and mild, moderate, severe pain ("three-class task"). For the binary pain labels, we considered 1-3 as *no pain* and 4-10 as *pain*. For the three-class problem, we treated 1-3 as *mild*, 4-6 as *moderate*, and 7-10 as *severe* [15].

5.1. Acoustic Features

We extracted 6,373 functional acoustic features using the Python openSMILE [30] toolkit from the ComParE2016 [31] feature set, which is known for being used in a variety of non-semantic and paralinguistic acoustic events, including previous work on pain assessment [13]. We applied standard normal scaling (mean 0, unit variance), followed by principle components analysis (PCA), and used a Scree Test to determine that the optimal number of components was 45 in both tasks .

5.2. Classifiers and Hyper-parameters

To establish the first baseline on this dataset, we explored several machine learning algorithms with hyper-parameter tuning. Using the Python Scikit-Learn toolkit [32], we selected support vector machines (SVM), which have previously shown to be useful for this task, Logistic Regression (Logit), and multilayer perceptron (MLP). To select hyper-parameters, we split our dataset into 70/30 train/test and performed gridsearch on the training set with stratified 10-fold cross-validation (stratified by label not speaker). For SVM, we explored different kernels (linear, radial basis function, and polynomial), as well as values for the regularization parameter $C = \{0.001{-}0.1, step = 0.001\}$, with $gamma = 1/n$ features. For Logit, we translated the labels into integers (0 and 1 for binary, and 0, 1, 2 for three-class) and explored values for the inverse regularization strength $C = \{0.001 - 0.1, step = 0.001\}$. For MLP, we explored different hidden layer sizes of 1-6 layers using 256 nodes per layer, and activation function set to ReLU [33]. We explored different learning rate initializations of $lr = \{0.1, 0.01, 0.001\}$ with constant, adaptive and inverse scaling, as well as $alpha = \{0.0001, 0.001, 0.01\}$, and early stopping. All other parameters were set to default values. We found the best parameters by optimizing for F1 score with micro averaging. The best hyper-parameters, feature names, and example mel frequency spectrograms are available[7].

5.3. Results

Table 1 presents performance as F1 score with micro averaging highlighting best performance on 10-fold cross validation and held-out test. We identified the best hyper-parameters from cross-validation and applied that to our test set. The SVM with polynomial kernel performs similarly to MLP in both tasks, including female-only and mixed gender utterances, with SVM slightly better on cross-validation. As a baseline, we show a random classifier (representing random guessing). We further investigated how different speakers affect classifier performance using hold-one-out speakers in a separate train/test split, and

[4] https://www.ukri.org/who-we-are/epsrc/our-policies-and-standards/framework-for-responsible-innovation/

[5] https://github.com/wiseman/py-webrtcvad/tree/master?tab=readme-ov-file

[6] Fully anonymized statistical acoustic features can be made available by contacting the lead author.

[7] https://rhoposit.github.io/interspeech2024

found that these models do not generalize well to unseen speakers, which should be explored in future work. This could be due to the size or imbalance of the dataset, and further motivates new directions toward personalized pain assessment.

Table 1: *Classifier performance on two tasks (binary and three-class), reporting normalized F1 scores, for utterances that were female-only as well as mixed male and female. Dark blue is best performance, light blue is second-best performance.*

	Binary		Three-Class	
	Female	**Mixed**	**Female**	**Mixed**
SVM				
- CV	80.0 ± 3.8	80.3 ± 1.7	69.9 ± 4.3	70.9 ± 4.6
- Test	78.0	73.8	70.6	67.3
Logit				
- CV	71.8 ± 3.4	73.8 ± 5.5	66.7 ± 3.5	66.4 ± 2.8
- Test	72.1	70.6	69.6	62.9
MLP				
- CV	79.3 ± 2.5	78.9 ± 3.0	68.3 ± 2.7	68.3 ± 4.9
- Test	78.3	72.8	70.1	68.4
Random	50.0	50.0	33.3	33.3

5.4. Feature Analysis

Explainability is an important aspect of any automated technique that may influence or contribute to clinician decision-making. Of the 45 features identified through PCA, 22 were spectral descriptors, 16 were energy descriptors, and 7 were voicing descriptors. Here, we explore how the different features contribute to classifier decisions using SHAP values [34]. We used the best-performing SVM model for each task from Section 5.3 and fit a Kernel Explainer to a stratified background model of 260 utterances for each task. We report the top 9 features for female-only speakers in the binary (Figure 2) and three-class (Figure 3) assessments. From both figures, we identify key features that contribute to classifier decisions:

1. PC1 *audspecRasta_lengthL1norm_sma_flatness*
2. PC2 *pcm_fftMag_spectralRollOff75.0_sma_de_stddev*
3. PC3 *audSpec_Rfilt_sma[1]_lpc1*
4. PC4 *audSpec_Rfilt_sma[3]_lpgain*

Feature PC1 is a feature corresponding to the relative spectral transform (RASTA) [35] applied to the auditory spectrum, which uses a band-pass filter on the energy in each frequency subband to smooth over noise variations and can simulate human audition. PC2 is a spectral roll off point, which can be linked to voice/unvoiced speech and breathiness. PC3 and PC4 represent mel frequency spectrum Perceptual Linear Prediction (PLP) cepstral coefficients and are known to be used in speech emotion recognition tasks [36]. Explanations of which features are most useful for pain assessment can contribute to developing more trustworthy tools for clinicians to adopt in practice, as we have demonstrated in this crucial first step.

6. Discussion and Future Work

We have presented a framework from which we collected and curated a new English speech dataset based on induced acute pain in adults. We identified key acoustic features that can discriminate well for pain detection and pain assessment. We showed that our SVM and MLP classifiers performed well on both tasks. Although our results are not directly comparable

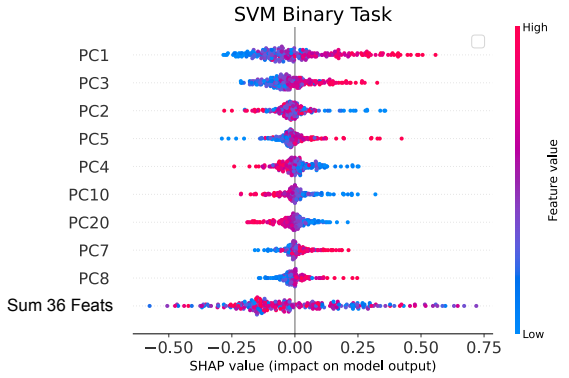

Figure 2: *Beeswarm plot showing SHAP values for the top 9 features corresponding to classification decisions on the (female-only) binary detection task, n=250.*

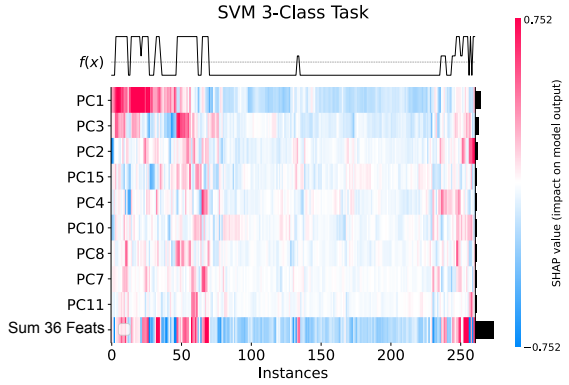

Figure 3: *Heatmap plot showing SHAP values for the top 9 features and corresponding decisions $f(x)$) on the (female-only) three-class pain assessment task, n=250.*

to previous work due to differences in underlying data, we achieved overall better performance than what has been previously reported for unimodal pain assessment from speech. The limitations of our dataset (number of speakers and gender imbalance) necessitated reporting separate results for female speakers and our classifiers did not generalize well to unseen speakers. To address this limitation, we have begun collecting a larger dataset of 50 participants with closer gender balance, which will be made publicly available for academic use. Our efforts will allow deeper exploration of automatic pain assessment, including more variety of speech features such as prosody and learning directly from spectrograms. We are interested in developing techniques that can more finely assess pain levels beyond a three-class granularity. The goal that we are working toward is to develop low-cost decision support tools for health professionals who stand to benefit from knowing that pain-related information can be found in the speech signal. Such tools have wide implications for the future of medicine, including personalized health monitoring, remote health conversations, and mitigation of bias when delivering analgesia.

7. Acknowledgements

This work was supported by the UK EPSRC through the Trustworthy Autonomous Systems (TAS) Hub (EP/V00784X/1), and

also by Good Systems, a research grand challenge at the University of Texas at Austin. The authors would like to thank Prof. Gopal Ramchurn, Prof. Joel Fischer, Prof. Sharon Strover, Dr. Liz Dowthwaite, and Dr. Anna-Maria Piskopani for their helpful feedback during this work.

8. References

[1] C. Sprenger, F. Eippert, J. Finsterbusch, U. Bingel, M. Rose, and C. Büchel, "Attention Modulates Spinal Cord Responses to Pain," *Current Biology*, vol. 22, no. 11, pp. 1019–1022, 2012.

[2] M. Granot and I. Weissman-Fogel, "The Effect of Post-Surgical Neuroplasticity on the Stability of Systemic Pain Perception: A Psychophysical Study," *European Journal of Pain*, vol. 16, no. 2, pp. 247–255, 2012.

[3] J. Nesbitt, S. Moxham, L. Williams *et al.*, "Improving pain assessment and managment in stroke patients," *BMJ Open Quality*, vol. 4, no. 1, pp. u203 375–w3105, 2015.

[4] A. K. Cook, C. A. Niven, and M. G. Downs, "Assessing the pain of people with cognitive impairment," *International Journal of Geriatric Psychiatry*, vol. 14, no. 6, pp. 421–425, 1999.

[5] J. Liu, L. L. Chen, S. Shen, J. Mao, M. Lopes, S. Liu, and X. Kong, "Challenges in the diagnosis and management of pain in individuals with autism spectrum disorder," *Review Journal of Autism and Developmental Disorders*, vol. 7, pp. 352–363, 2020.

[6] T. Sakulchit, B. Kuzeljevic, and R. D. Goldman, "Evaluation of digital face recognition technology for pain assessment in young children," *The Clinical Journal of Pain*, vol. 35, no. 1, pp. 18–22, 2019.

[7] S. C. Ahluwalia, K. F. Giannitrapani, S. K. Dobscha, R. Cromer, and K. A. Lorenz, ""Sometimes you wonder, is this really true?": Clinician assessment of patients' subjective experience of pain," *Journal of Evaluation in Clinical Practice*, vol. 26, no. 3, pp. 1048–1053, 2020.

[8] J. Lötsch and A. Ultsch, "Machine Learning in Pain Research," *Pain*, vol. 159, no. 4, pp. 623–630, 2017.

[9] R. S. Stojancic, A. Subramaniam, C. Vuong, K. Utkarsh, N. C. Golbasi, O. Fernandez, and N. Shah, "Predicting Pain in People with Sickle Cell Disease in the Day Hospital Using the Commercial Wearable Apple Watch: Feasibility Study," *JMIR Formative Research*, vol. 7, p. e45355, 2023.

[10] D. Powell, "Walk, talk, think, see and feel: harnessing the power of digital biomarkers in healthcare," *NPJ Digital Medicine*, vol. 7, no. 45, 2024.

[11] Y. Oshrat, A. Bloch, A. Lerner, A. Cohen, M. Avigal, and G. Zeilig, "Speech prosody as a biosignal for physical pain detection," in *Proc. Speech Prosody 2016*, 2016, pp. 420–424.

[12] F.-S. Tsai, Y.-M. Weng, C.-J. Ng, and C.-C. Lee, "Embedding stacked bottleneck vocal features in a LSTM architecture for automatic pain level classification during emergency triage," *ACII 2017*, pp. 313–318, 2017.

[13] Z. Ren, N. Cummins, J. Han, S. Schnieder, J. Krajewski, and B. Schuller, "Evaluation of the pain level from speech: Introducing a novel pain database and benchmarks," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[14] C. Fröjd, C. Lampic, G. Larsson, G. Birgegård, and L. v. Essén, "Patient attitudes, behaviours, and other factors considered by doctors when estimating cancer patients' anxiety and desire for information," *Scandinavian Journal of Caring Sciences*, vol. 21, pp. 523–529, 2007.

[15] J. Miró, R. de la Vega, E. Solé, M. Racine, M. P. Jensen, S. Gálan, and J. M. Engel, "Defining mild, moderate, and severe pain in young people with physical disabilities," *Disability and Rehabilitation*, vol. 39, no. 11, pp. 1131–1135, 2017.

[16] R. Fernandez Rojas, X. Huang, and K.-L. Ou, "A machine learning approach for the identification of a biomarker of human pain using fNIRS," *Scientific Reports*, vol. 9, no. 1, p. 5645, 2019.

[17] B. D. Winslow, R. Kwasinski, K. Whirlow, E. Mills, J. Hullfish, and M. Carroll, "Automatic detection of pain using machine learning," *Frontiers in Pain Research*, vol. 3, p. 1044518, 2022.

[18] Y. Liu, K. Wang, L. Wei, J. Chen, Y. Zhan, D. Tao, and Z. Chen, "Affective Computing for Healthcare: Recent Trends, Applications, Challenges, and Beyond," *arXiv preprint arXiv:2402.13589*, 2024.

[19] E. Schneiders, J. Williams, A. Farahi, T. Seabrooke, G. Vigneswaran, J. R. Bautista *et al.*, "Tame pain: Trustworthy assessment of pain from speech and audio for the empowerment of patients," in *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, ser. TAS '23, 2023.

[20] R. F. Rojas, C. Joseph, G. Bargshady, and K.-L. Ou, "Empirical comparison of deep learning models for fnirs pain decoding," *Frontiers in Neuroinformatics*, vol. 18, 2024.

[21] M. S. Salekin, G. Zamzmi, J. Hausmann, D. Goldgof, R. Kasturi, M. Kneusel *et al.*, "Multimodal neonatal procedural and postoperative pain assessment dataset," *Data in Brief*, vol. 35, p. 106796, 2021.

[22] K. Pisanski, J. Raine, and D. Reby, "Individual differences in human voice pitch are preserved from speech to screams, roars and pain cries," *Royal Society Open Science*, vol. 7, no. 2, p. 191642, 2020.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *The 3rd International Conference on Learning Representations (ICLR2015)*, 2015.

[24] L. A. Mitchell, R. A. MacDonald, and E. E. Brodie, "Temperature and the cold pressor test," *The Journal of Pain*, vol. 5, no. 4, pp. 233–237, 2004.

[25] M. H. McIntyre, 23andMe Research Team, A. Kless, P. Hein, M. Field, and J. Y. Tung, "Validity of the cold pressor test and pain sensitivity questionnaire via online self-administration," *PLOS ONE*, vol. 15, no. 4, pp. 1–16, 04 2020.

[26] S. Ghiasi, A. Greco, R. Barbieri, E. P. Scilingo, and G. Valenza, "Assessing autonomic function from electrodermal activity and heart rate variability during cold-pressor test and emotional challenge," *Scientific Reports*, vol. 10, no. 1, p. 5406, 2020.

[27] L. Schwabe, L. Haddad, and H. Schachinger, "HPA axis activation by a socially evaluated cold-pressor test," *Psychoneuroendocrinology*, vol. 33, no. 6, pp. 890–895, 2008.

[28] E. Rothauser, "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[29] B. C. Stahl, C. Aicardi, L. Brooks, P. J. Craigon, M. Cunden, Burton *et al.*, "Assessing responsible innovation training," *Journal of Responsible Technology*, p. 100063, 2023.

[30] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[31] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird *et al.*, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *INTERSPEECH*, 2016, pp. 2001–2005.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[33] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[34] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, pp. 4765–4774.

[35] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[36] D. Kamińska, T. Sapiński, and G. Anbarjafari, "Efficiency of chosen speech descriptors in relation to emotion recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2017, pp. 1–9, 2017.