Stress Inference from Abdominal Sounds using Machine Learning

Erika Bondareva[1,2], Marios Constantinides[2†], Michael S. Eggleston[3†], Ireneusz Jabłoński[4†], Cecilia Mascolo[1†], Zoran Radivojevic[2†], Sanja Šćepanović[2†*]

***Abstract*— Stress is often considered the 21st century's epidemic, affecting more than a third of the globe's population. Long-term exposure to stress has significant side effects on physical and mental health. In this work we propose a methodology for detecting stress using abdominal sounds. During the data collection step, eight participants for ten days were either exposed to a stressful (Stroop test) or a relaxing (guided meditation) stimulus. In total, we collected 104 hours of abdominal sounds using a custom wearable device in a belt form-factor. We explored the effect of various features on the binary stress classification accuracy using traditional machine learning methods. Namely, we observed the impact of using acoustic features on their own, as well as in combination with daily mood report, and hand-crafted domain-specific features. After feature extraction and reduction, by utilising a multilayer perceptron classifier model we achieved 77% accuracy in detecting abdominal sounds under stress exposure.**

***Clinical relevance*— This feasibility study confirms the link between the gastrointestinal system and stress and uncovers a novel approach for stress inference via abdominal sounds using machine learning.**

I. Introduction

Stress is an experience of anticipating or encountering adversity. In response to stress, our bodies typically react by activating the sympathetic nervous system, withdrawing the parasympathetic system, and increasing the activity of the hypothalamic-pituitary-adrenal axis [1].

Traditionally, tracking stress is achieved through subjective or objective measurements. The former include questionnaires, such as the Perceived Stress Scale (PSS) [2] or the Kessler Psychological Distress Scale (K10) [3]. While they are typically easy to administer and low-cost, they tend to suffer from subjectivity biases. To obtain objective measurements, stress biomarkers can be analysed from saliva or blood, including epinephrine and nor-epinephrine, alpha-amylase, and cortisol. As wearables are becoming more popular, they offer a cheap alternative for continuous stress monitoring in free-living settings [4].

The field of automated abdominal auscultation has only started gaining traction in the past two decades. Numerous research efforts focused on classification of various states by utilising abdominal sounds (ABS). For example, in [5] a system was devised for evaluation of gastrointestinal (GI) motility, especially for patients with diabetes mellitus. Spiegel et al. [6] explored using ABS for detection of post-operative ileus. In [7] a low-power system was devised for monitoring ABS for detection of gastric events, linking the frequency of the GI events to eating instances, and Kolle et al. [8] later looked into using ABS for meal detection, utilising SVMs. In [9] Mel-frequency cepstral coefficients (MFCCs) and wavelet entropy were used as features for a neural network for discriminating meal and no-meal GI sounds. In addition to the research efforts at detecting eating instances, a proof of concept of an irritable bowel syndrome diagnostic system was devised in [10].

A promising yet underexplored area of tracking stress is via GI activity by monitoring abdominal sounds. Stress has both short-term and long-term effects on the GI system, as it can affect gastric secretion, gut motility, mucosal permeability, etc. [11]. Building on accumulated empirical evidence, we hypothesised and tested the feasibility of detecting whether a person is stressed from abdominal sounds. In so doing, we made three sets of contributions:

- We collected a novel dataset comprised of 104 hours of abdominal sounds using a custom laminated e-stethoscope in a stretchable belt (Section II).
- Using this dataset, we extracted and compared a standard set of acoustic features versus hand-crafted features, which were used to train classifiers for stress detection (Section III).
- We used Support Vector Machine (SVM) and Multilayer Perceptron (MLP) classifiers for stress inference from the acoustic features (Section IV), with our best performing model achieving 77% accuracy.

Our primary research goal was to establish whether it is possible to infer stress from abdominal sounds, collected using a laminated e-stethoscope in a stretchable belt. To address that, we subsequently formulated three research questions as follows:

- **RQ1**: Using standard audio-based features, is it possible to infer stress from sounds collected from the abdomen?
- **RQ2**: What effect does adding features representing current mood state have on the inference accuracy?
- **RQ3**: What effect does using standard audio-based features alongside hand-crafted features have on the inference accuracy?

† Equal contribution, listed alphabetically.

*This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/L015889/1 for the Centre for Doctoral Training in Sensor Technologies and Applications, and by Nokia Bell Labs through their donation for the Centre of Mobile, Wearable Systems and Augmented Intelligence to the University of Cambridge.

[1] E. Bondareva (eb729@cam.ac.uk) and C. Mascolo are with the Dept of Computer Science and Technology at University of Cambridge, UK

[2] E. Bondareva, M. Constantinides, Z. Radivojevic, and S. Šćepanović are with Nokia Bell Labs, Cambridge, UK

[3] M. S. Eggleston is with Nokia Bell Labs, Murray Hill, NJ, USA

[4] I. Jabłoński is with Wroclaw University of Science and Technology, Poland and Nokia, Wroclaw, Poland

II. Dataset description

We collected a novel dataset comprised of 104 hours of abdominal sounds with and without stressful stimulus from 8 subjects. The data collection was approved by the ethics committee of the Department of Computer Science and Technology at the University of Cambridge.

A. Wearables for Biometric Data Collection

To record abdominal body sounds, we utilised a custom-built e-stethoscope in a stretchable belt with a high-quality microphone in a flexible tube connected to the stethoscope head. The belt had an adjustable strap with zip pockets, with a data cable connected to an audio recorder. The audio was collected in an MP3 format at $192\,\mathrm{kHz}$, which upon export was automatically resampled to $44.1\,\mathrm{kHz}$. This data recording approach was extensively tested and proved to provide a signal barely discernable from a $192\,\mathrm{kHz}$ WAV audio file, while being significantly smaller, thus easing file transfer by study participants.

To reduce the potential level of noise inadvertently collected during participants engaging in daily activities, we targeted subjects with relatively sedentary lifestyles. In particular, the participants were asked to stay seated at the desk during the data collection. While sitting, they were allowed to engage in any activity of their choosing, e.g. working.

B. Stress Stimulus

To elicit a relatively controlled psychological response from our participants, we chose two types of activities: *i)* stressful task, and *ii)* relaxing task, each approximately 10–15 minutes in duration. The participants were instructed to complete the task once they start recording the biometrics.

At random, participants were given either the stressful task or the relaxing task. The stressful task comprised of the widely used Stroop Colour Word Test [12]. This exercise was proven to reliably stimulate a stressful response, affecting the participant's cognitive load and physiological system [13], [14]. The relaxing task involved watching and following a video on YouTube with a guided meditation, from Headspace YouTube series "Guided meditation with Andy". Meditation has been shown to regulate emotion, reduce stress, and increase well-being, making it a suitable activity for inducing a relaxed state [15].

C. Data Collection Tools

Each participant was issued a unique identification number with which all the collected data was associated to ensure data anonymisation. We collected each subject's daily mood through a short questionnaire similar to [16], comprised of five questions relating their: *i)* happiness, *ii)* awakedness, *iii)* relaxedness, *iv)* sleep quality the night before, and *v)* stress levels the day before. Each question was answered on a Likert scale, from 0 ("not at all") to 5 ("very much"). These short daily questionnaires were administered via Google Forms. The daily task was set up via Pavlovia[1], which is an online platform designed for hosting psychological experiments. Every day, upon starting the experiment, each participant was asked to open a webpage set up specifically for this study. The webpage contained instructions for the daily task, and a button redirecting the participant to the relevant link on Pavlovia for the daily task.

Participants recorded data for a total duration of 2 hours. After the data collection was over, each participant was asked to upload the breakfast picture (taken to ensure adherence to the limited number of food items permitted for consumption during the study to avoid significant influence of food on GI system) and the collected ABS data to a Dropbox file request link, set up purposefully for anonymous uploads. In total, each participant was asked to collect the data for 10 days, although due to the daily data collection time being at least 2 hours, participants were allowed to collect the data on non-consecutive days, if they wished.

D. Study Protocol

Five steps were performed during data collection (Fig. 1):

1) Take a picture of the breakfast before eating it.
2) Put on the belt for collection of ABS, get comfortable at the desk at a computer. Start recording the data.
3) Open the experiment webpage and start the daily task, after which the daily questionnaire opens automatically.
4) Remain seated at the desk, while free to do anything. The total duration of the daily data collection was 2 hours.
5) Once done with data collection, upload the data.

E. Dataset Statistics

Each participant recruited for this study went through a rigorous prescreening and onboarding process. Anybody currently suffering from digestive disorders, mood disorders, or pregnant were excluded from this study in the prescreening stage. In addition, those on medication (e.g. sleeping medication or antidepressants) were also excluded from the study as medications may alter the mood or reaction time. To ensure the correct set-up of the wearables for data collection, short samples of biometric data went through a quality control process, after which the participants started the study. The quality control process involved algorithmically ensuring sufficient signal-to-noise ratio of the collected ABS signal. If the samples did not pass the quality control process, the participants were advised to increase the belt tension to ensure sufficient coupling between the skin and the sensor, which significantly improved the signal quality.

In total, we collected data from 8 participants. Each of the participants contributed 10 days worth of data, 2 hours per day. However, some files had to be discarded from the initial analysis due to inconsistencies in the recordings (e.g., missing audio samples, or missing Pavlovia performance files). In addition, all the audio samples were manually screened using an open-source software Audacity to ensure that the recordings contained abdominal sounds, and no significant noise was present. For this study, we used 52 data samples from 8 participants. Among them, 4 were female and 4 were male, with the age range 20–32 years old (μ =

[1] https://pavlovia.org/

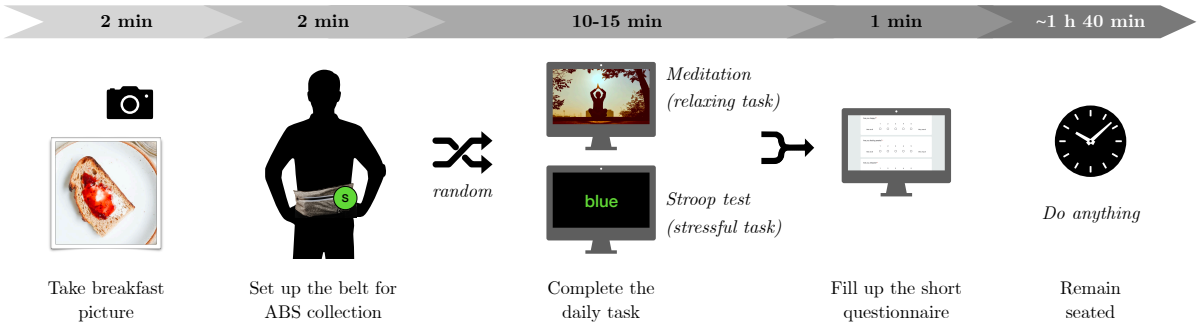

Fig. 1. The study design and the protocol steps followed by the study participants, with the associated timings. The green circle on the ABS belt shows the location of the stethoscope head, on the left lower quadrant of the abdomen.

25.13 and σ = 3.76 years). Out of these 52 samples, 29 were collected during exposure to stress, and 23 were collected with the participants meditating during the daily task.

III. Methodology

A. Data Preprocessing

We labelled the ABS data collected on the days when the participant did the Stroop test (during the stressful task) as **stress** samples, and the ABS data from the days when the participant meditated (during the relaxing task) as **no-stress**.

Each audio recording was at least 2 hours long, but due to the nature of this study, the researchers had little to no control over the participants' mood state for most of this duration. Therefore, in this work, only the first 5–30 minutes of the audio recordings were used for the classification task, when the participant is exposed to a known stimulus. To ensure that no environmental noise was inadvertently captured, each segmented window was filtered using a 4$^{\text{th}}$ order Butterworth bandpass filter (60 Hz to 3000 Hz).

B. Extracted Features

To address **RQ1**, the INTERSPEECH ComParE 2018 feature set (IS-18) [17] was extracted from the ABS recordings, yielding a vector with 6373 features. This feature set has been applied successfully for many acoustic tasks and is particularly attractive for automated classification tasks as it allows researchers to bypass creating hand-crafted features.

Given the potential limitations of IS-18 feature set to capture all the ABS components necessary for inference of stress, **RQ2** and **RQ3** were posed to evaluate the classifier performance with various feature vectors.

For **RQ2** five features were added to the IS-18 feature vector, representing numerical scores shared by our participants during the daily short questionnaire. To extract hand-crafted features for **RQ3**, we applied a denoising algorithm—a 5$^{\text{th}}$ order discrete wavelet transform (DWT) filter with a Coiflet 5 wavelet. Individual sounds likely related to gastric events were identified by using a peak detection algorithm, and a number of features was then extracted, inspired by the work of [18]. Specifically, per minute-long segments we extracted: the number of detected events, the duration μ and σ of the detected events, the amplitude μ and σ of the detected events, and the silence μ and σ between sounds duration, yielding a total of 140 additional features.

Before using the features for classification, we scaled them, and, in turn, principal component analysis was applied to retain 51 components and a 99.9% variance ratio.

C. Classification Algorithm

For stress inference, we deployed a number of traditional machine learning algorithms including that of k-nearest neighbours, decision trees, random forest, support vector machine (SVM), and multilayer perceptron (MLP). However, SVM (with a linear kernel) and MLP (100 epochs) performed better than the rest of the approaches, thus we report the performance results on these two classifiers. For performance evaluation, we used the leave-one-sample-out cross-validation method and compared the performance across four metrics: accuracy, precision, recall, and f-1 score.

IV. Results and Discussion

We ran three experiments with varied feature sets, and compared performance for two best performing classifiers (Table I). While similar performance was observed across the experiments, the slight variations offer interesting insights.

Using IS-18 features with or without the features representing the self-reported mental state of participants yielded the same performance, and our MLP model achieved the highest accuracy of 75%. Interestingly, instead of a performance improvement with the additional hand-crafted features, no changes in performance were observed for SVM, while a slight reduction was observed across all the metrics for our MLP model, except the metric of sensitivity. It appears that using GI event related features allows for better detection of stressful state, but also leads to a higher number of false positives. Higher sensitivity might be desirable for a stress detection algorithm: if this algorithm was used for stress management, untimely notification to perform a breathing exercise would cause fewer implications than missing a stressful episode. While a limited number of methods could be used given the small amount of data, this result encourages exploring neural networks and deep learning as a method both yielding better performance, but also being more sensitive to changes in the feature vectors.

We also examined the effect that the window size had on our classifiers' performance. Using MLP with IS-18 features,

TABLE I

PERFORMANCE METRICS ON STRESS INFERENCE FROM ABS, FOR THREE RQS. MLP WITH AUDIO-BASED FEATURES IDENTIFIED THE STRESSED STATE OF THE PARTICIPANT BEST, WITH A 75% ACCURACY.

	RQ1 / RQ2		**RQ3**	
	SVM	**MLP**	**SVM**	**MLP**
Accuracy	0.67	0.75	0.67	0.73
Sensitivity	0.69	0.69	0.69	0.72
Specificity	0.65	0.83	0.65	0.74
Precision	0.71	0.83	0.71	0.78
f1-score	0.70	0.75	0.70	0.75

we found that the best results are achieved by 5 and 20-minute window sizes, with the highest accuracy of 77% obtained by using a 5-minute window. Interestingly, the highest specificity is achieved by using the longest window of 30 minutes, albeit this window size also resulted in the worst sensitivity, which remained unchanged for all the other window sizes. This provides an insight into what ABS data may be the most valuable for accurate stress inference.

Our work has both theoretical and practical implications. From a theoretical standpoint, our results add empirical evidence to the growing body of literature that links the GI system to stress; stress directly affects GI motility and thus abdominal sounds. From a practical perspective, we showed that by using a sound recording belt, we uncovered a novel approach for stress inference via continuous, yet minimally obtrusive monitoring. Due to our system's ubiquitous nature, it could be used outside of medical facilities without any medical supervision, and supplement the stress monitoring capability with other ABS-related applications.

V. LIMITATIONS AND FUTURE WORK

Our work has limitations that call for future research efforts. It is worth noting that the data was collected from a limited group of participants, of a similar age and leading a similar lifestyle, thus it is unclear how the findings presented in this work would generalise to a wider population.

Stress is likely not a binary state, thus future studies could investigate a finer-grained representation of participants' stress. In addition, the physiological stress response may vary widely across participants — if a participant has a positive stress mindset [19], (s)he is likely to experience a smaller physiological impact of stress. As a result, the prediction of stress in such subjects is likely to be more challenging. However, future studies could address these limitations through personalised models wherein each participant's data is considered in isolation.

While **RQ2** focused on establishing whether current mood state affects GI sounds, the mood score did not improve our model's performance. However, this effect might be in part observed due to the scarcity of the samples, and correlation between ABS and mood might be detectable upon larger quantities of data available for analysis. In addition, future studies could well use ABS for detecting one's mental state as opposed to inferring one's experienced stress.

Based on the results obtained on the analysis of various window sizes, the highest accuracy was obtained by the shortest window, while the highest specificity was yielded by the longest window. It appears that information necessary for accurate identification of "stress-positive" cases mostly is contained in the first minutes of the exposure to stress, while most information necessary for confirming "stress-negative" cases may be most accessible in longer audio segments. Therefore, it would be imperative to explore using ML methods specific to time-series data such as recurrent neural networks (RNNs), provided that more data can be obtained.

REFERENCES

[1] S. R. Kunz-Ebrecht, V. Mohamed-Ali, P. J. Feldman, et al., "Cortisol responses to mild psychological stress are inversely associated with proinflammatory cytokines," *Brain Behav. Immun.*, vol. 17, no. 5, pp. 373–383, 2003.

[2] M. G. Nielsen, E. Ørnbøl, M. Vestergaard, et al., "The construct validity of the perceived stress scale," *J. Psychosom. Res.*, vol. 84, pp. 22–30, 2016.

[3] R. C. Kessler, G. Andrews, L. J. Colpe, et al., "Short screening scales to monitor population prevalences and trends in non-specific psychological distress," *Psychol. Med.*, vol. 32, no. 6, pp. 959–976, 2002.

[4] P. Schmidt, A. Reiss, R. Dürichen, et al., "Wearable-based affect recognition—a review," *Sensors*, vol. 19, no. 19, 2019.

[5] K. Yamaguchi, T. Yamaguchi, T. Odaka, et al., "Evaluation of gastrointestinal motility by computerized analysis of abdominal auscultation findings," *J. Gastroenterol. Hepatol.*, vol. 21, no. 3, pp. 510–514, 2006.

[6] B. M. R. Spiegel, M. Kaneshiro, M. M. Russell, et al., "Validation of an acoustic gastrointestinal surveillance biosensor for postoperative ileus," *J. Gastrointest. Surg.*, vol. 18, no. 10, pp. 1795–1803, 2014.

[7] K. A. Al Mamun, M. H. U. Habib, N. McFarlane, et al., "A low power integrated bowel sound measurement system," in *IEEE Int. Instrum. Meas. Technol. Conf. Proc.*, 2015, pp. 779–783.

[8] K. Kölle, A. L. Fougner, R. Ellingsen, et al., "Feasibility of early meal detection based on abdominal sound," *IEEE J. Transl. Eng. Health Med.*, vol. 7, pp. 1–12, 2019.

[9] T. S. Kumar, E. Søiland, Ø. Stavdahl, et al., "Pilot study of early meal onset detection from abdominal sounds," in *E-Health Bioeng. Conf. Proc.*, 2019, pp. 1–4.

[10] X. Du, G. Allwood, K. M. Webberley, et al., "Noninvasive diagnosis of irritable bowel syndrome via bowel sound features: Proof of concept," *Clin. Transl. Gastroenterol.*, vol. 10, no. 3, 2019.

[11] V. Bhatia and R. K. Tandon, "Stress and the gastrointestinal tract," *J. Gastroenterol. Hepatol.*, vol. 20, no. 3, pp. 332–339, 2005.

[12] J. R. Stroop, "Studies of interference in serial verbal reactions," *Journal of Experimental Psychology*, vol. 18, no. 6, pp. 643–662, 1935.

[13] J. H. M. Tulen, P. Moleman, H. G. van Steenis, et al., "Characterization of stress reactions to the stroop color word test," *Pharmacol. Biochem. Behav.*, vol. 32, no. 1, pp. 9–15, 1989.

[14] P. Renaud and J. Blondin, "The stress of stroop performance: physiological and emotional responses to color–word interference, task pacing, and pacing speed," *Int. J. Psychophysiol.*, vol. 27, no. 2, pp. 87–97, 1997.

[15] Y. Tang, B. K. Hölzel, and M. I. Posner, "The neuroscience of mindfulness meditation," *Nat. Rev. Neurosci.*, vol. 16, no. 4, pp. 213–225, 2015.

[16] S. Park, M. Constantinides, L. M. Aiello, et al., "Wellbeat: A framework for tracking daily well-being using smartwatches," *IEEE Internet Comput.*, vol. 24, no. 5, pp. 10–17, 2020.

[17] B. Schuller, S. Steidl, A. Batliner, et al., "The INTERSPEECH 2018 computational paralinguistics challenge: Atypical and self-assessed affect, crying and heart beats," in *Proc. INTERSPEECH*, 2018, pp. 122–126.

[18] R. Ranta, V. Louis-Dorr, C. Heinrich, et al., "Digestive activity evaluation by multichannel abdominal sounds analysis," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 6, pp. 1507–1519, 2010.

[19] A. J. Crum, P. Salovey, and S. Achor, "Rethinking stress: The role of mindsets in determining the stress response," *J. Pers. Soc. Psychol.*, vol. 104, no. 4, pp. 716–733, 2013.