Deep Breathing Phase Classification with a Social Robot for Mental Health

Kayla Matheus*
kayla.matheus@yale.edu
Yale University
New Haven, CT, USA

Ellie Mamantov*
ellie.mamantov@yale.edu
Yale University
New Haven, CT, USA

Marynel Vázquez
marynel.vazquez@yale.edu
Yale University
New Haven, CT, USA

Brian Scassellati
brian.scassellati@yale.edu
Yale University
New Haven, CT, USA

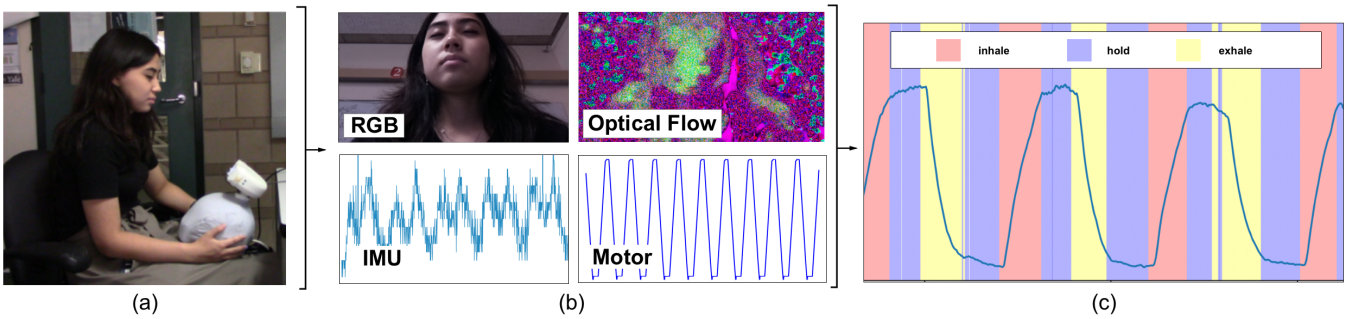

Figure 1: A visualization of deep breathing phase classification with a robot. (a) A participant deep breathing with an *Ommie* robot during data collection. (b) Visual examples of multimodal data from the robot's sensors. (c) Deep breathing phase predictions for inhales, holds, and exhales.

ABSTRACT

Social robots are in a unique position to aid mental health by supporting engagement with behavioral interventions. One such behavioral intervention is the practice of deep breathing, which has been shown to physiologically reduce symptoms of anxiety. Multiple robots have been recently developed that support deep breathing, but none yet implement a method to detect how accurately an individual is performing the practice. Detecting breathing phases (i.e., inhaling, breath holding, or exhaling) is a challenge with these robots since often the robot is being manipulated or moved by the user, or the robot itself is moving to generate haptic feedback. Accordingly, we first present OMMDB: a novel, multimodal, public dataset made up of individuals performing deep breathing with an *Ommie* robot in multiple conditions of robot ego-motion. The dataset includes RGB video, inertial sensor data, and motor encoder data, as well as ground truth breathing data from a respiration belt. Our second contribution features experimental results with a convolutional long-short term memory neural network trained using OMMDB. These results show the system's ability to be applied to the domain of deep breathing and generalize between individual users. We additionally show that our model is able to generalize across multiple types of robot ego-motion, reducing the need to train individual models for varying human-robot interaction conditions.

CCS CONCEPTS

• **Computer systems organization** → **Robotics**; • **Human-centered computing** → *Interaction devices*; • **Applied computing** → *Consumer health.*

KEYWORDS

social robotics, mental health, anxiety, human-robot interaction, deep breathing, vital signs, datasets, multimodal datasets

*Both authors contributed equally to the paper

ICMI '23, October 9–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10.
https://doi.org/10.1145/3577190.3614173

ACM Reference Format:
Kayla Matheus, Ellie Mamantov, Marynel Vázquez, and Brian Scassellati. 2023. Deep Breathing Phase Classification with a Social Robot for Mental Health. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 9–13, 2023, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3577190.3614173

1 INTRODUCTION

Interventions to support mental health are needed now more than ever given high rates of anxiety and depression in multiple populations [5, 36, 56]. One key behavioral practice for supporting mental health is deep breathing. Deep breathing is characterized as intentionally taking a sequence of extended or modified breaths in and out [21]. The practice is a common therapy used in clinical settings, however at-home compliance is a known challenge [12]. Given past success in using social robots to increase compliance for physical health behaviors [10, 23, 47], multiple research efforts are now exploring supporting deep breathing through human-robot interaction (HRI). For example, the *Ommie* robot [34] provides haptic and non-verbal audio cues to guide deep breathing, *CAKNA* [2] provides spoken guidance to do so, and a *Jibo* robot [20] verbally guides mindfulness practices that include deep breathing.

While multiple robotic systems provide instruction for deep breathing, they do not yet have a method to perceive how well a user is following these instructions. A robot that could observe a user's breathing phase (i.e., an inhale, breath hold, or exhale) would have the capacity to more deeply engage with a user. For instance, a robot could provide real-time, personalized corrections if a user is struggling with a particular breathing phase. A robot could also observe how well a user is matching expected breathing phases to evaluate their performance level. A new interaction modality could also be introduced, where a robot mimics the user's breathing as a companion rather than as a coach. These additional features offer opportunities for interaction diversity and personalization, which have been shown to help overcome novelty effects [28] and increase long-term engagement with social robots [27, 29]. Long-term use is particularly important for robots in mental health as behavioral treatments must be practiced regularly in order to provide preventative therapeutic effects [12].

Existing technologies for monitoring respiration typically rely on contact sensors, which pose challenges for HRI applications in mental health. For instance, on-body contact sensors (e.g., a chest strap or inflatable belt) introduce significant inconvenience and discomfort that could detract from robot use. Especially for a population seeking to improve their mental health, any elements that may increase stress levels or the effort required to engage with a behavioral intervention must be minimized. This includes avoiding having users feel like they are under surveillance from externally placed cameras or sensors. These challenges motivated us to create a robot-based perception system that is able to identify a user's breathing phases using non-contact sensing technologies.

Classifying deep breathing phases from a robot, in a non-contact manner, is novel and non-trivial. First, the system will have to work regularly with a number of different individuals and deep breathing variations. For example, different members of a family may use the robot at home, or various patients may be seen in a therapist's office with the robot. The system must thus be able to accommodate a variety of human appearances and deep breathing styles. Second, the system needs to support a wide variety of physical interactions from users. Robots for mental health can often be held in the lap in addition to being used on a table top (e.g., [2, 20, 34, 57]). Some of these solutions also include breathing motions from the robot itself [34, 46, 57], further influencing the robot's ego-motion.

In this paper, we investigate data-driven, non-contact models for robotic deep breathing phase classification (as outlined in Fig. 1). More specifically, we present two main contributions: (1) the Open Multimodal Deep Breathing dataset (OMMDB)[1], which addresses the lack of a publicly-available dataset for deep breathing phases with non-contact sensors; and (2) experiments with a convolutional long-short-term-memory model (LSTM) trained on OMMDB data. To collect OMMDB, we modified an *Ommie* robot [34] with a perception system consisting of an RGB camera, inertial measurement unit (IMU), and motor encoder. We utilized a respiration force belt to collect ground truth data, which was then annotated with deep breathing phases. Our model training experiments were motivated by the need to generalize across individuals and different interaction modalities with the *Ommie* robot. Results from cross-validation demonstrate the promise of using modern deep learning architectures to classify deep breathing phases in a variety of human-robot interaction conditions as well as across varied users.

2 BACKGROUND AND RELATED WORK

2.1 Therapeutic Deep Breathing

Deep breathing is characterized by extending one or more of the four breathing phases: inhaling, holding post-inhale, exhaling, and holding post-exhale. This act results in deeper expansion of the diaphragm and a slowing of the breath [21]. Deep breathing can be achieved with multiple cadences, including the popular "box breathing" [38], which extends all four phases equally. Recent research has shown how deep breathing is therapeutic. Physiologically, the act of deep breathing reduces heart rate and cortisol levels [17, 42] and calms the autonomic nervous system [22]. These changes promote emotion regulation and anxiety reduction, making deep breathing a common treatment for anxiety [21], depression [50], as well as general stressors [43]. Another benefit is that deep breathing can be done anywhere and has been shown to be effective among many different populations [35, 39]. These characteristics make the practice a powerful tool for promoting mental health and well-being.

2.2 Social Robots for Mental Health

Work in HRI is actively exploring how social robots can support the growing mental health needs of individuals. Robots such as *Paro* [46], *Haptic Creature* [48], and *Taco* [40] have helped calm individuals in stressful situations through haptic, often animal-like, interactions. Other robots focus on behavioral practices for mental health, such as mindfulness, guided imagery, and positive affirmations [2, 20, 34, 51]. A subset of these robots utilize deep breathing as a therapeutic behavioral practice. For instance, the *Ommie* robot [34] features haptic guidance where the robot's body physically expands and contracts in the cadence of deep breathing, in addition to nonverbal audio cues. Matheus et al. [34] found that deep breathing with *Ommie* provided a significant reduction in anxiety state measures. Jeong et al. [20] have explored using a *Jibo* robot to deliver positive psychology interventions, including deep breathing. Their results similarly show a significant improvement in psychological well-being from use with the robot. Also supporting deep breathing

[1]For access to the dataset, please see: http://scazlab.yale.edu/ommdb-dataset

is *CAKNA* [2], a robot that provides verbal instruction for psychological techniques such as guided imagery and deep breathing. Aziz et al. [2] found that using *CAKNA* reduced anxiety levels more so than a computer providing the same instruction.

Among the robots that support deep breathing, all are either intended to be used on, or can be used on, a table top. *Ommie* can additionally be used in the lap [34]. Other mental well-being robots such as *Haptic Creature* [48] and *Paro* [45] are intended to be used in the lap and provide robotic breathing motions, though they do not specifically instruct the user on deep breathing. This variability in robot positioning inspired us to collect a dataset and pursue a classification technique that could support use of a robot both in a lap and on a table. The capacity for social robots to have their own breathing motions additionally motivated us to collect and test on data with conditions incorporating such motions.

2.3 Non-Contact Respiration Detection

To our knowledge, there is no prior work in breathing phase classification specific to deep breathing. Instead, we draw inspiration from prior work that has occurred in two adjacent domains of non-contact respiration detection: remote vital signs detection, and breathing disorders classification. Most prior work in these domains falls into four categories of sensor types: RGB video (e.g., [30, 32]), thermal cameras (e.g., [4, 11, 18]), audio (e.g., [25, 37]), and radar (e.g., [3]). For our purposes, RGB cameras held the most promise. Thermal techniques require high resolution cameras to detect changes in nostril temperatures, which are large in physical footprint, high in cost, and can still struggle with accurate results [4, 18]. With audio, microphone arrays have been integrated in many prior robots, but in-the-wild environments like homes or doctors' offices can pose challenges due to environmental and background noise [19, 26]. Radar for respiration rate has previously required larger, custom-built systems, but new off-the-shelf units are increasingly becoming available (e.g., [1]). However, prior work has shown that radar sensors are sensitive to the distance between the robot and the user [44]. In contrast, a significant amount of prior work using RGB video for vital signs detection (e.g., [15, 49]) and breathing disorders classification (e.g., [14]) has been successful and can be applied to robotic applications.

With RGB video data, remote photoplethysmography (rPPG) [6] is a common technique for predicting respiration signals, based on estimated chest volume or peak-to-peak respiration rate. rPPG works by magnifying pixel color shifts in order to assess changes in blood flow, from which other physiological measures can be derived. A recent application using rPPG for modeling respiration signals is Microsoft Research's open-source MTTS-CAN [30], which utilizes a webcam and a multi-task temporal shift convolutional attention network. However, a limitation in using rPPG for respiration is that predictions are made based on an extrapolation from heart rate using physiological models [7, 30]. Accordingly, datasets often used with rPPG lack ground truth data for respiration (e.g., AFRL [8] and MMSE [58]). A vision-based dataset that does include ground truth respiration data is the Vision for Vitals (V4V) dataset [59]. However, the data is limited to respiration rate, which alone cannot be used for determining different breathing phases. Our work creating OMMDB addresses this gap.

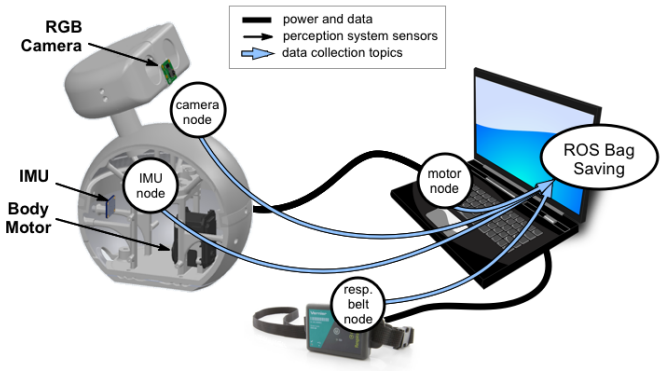

Figure 2: The data collection system including the *Ommie* robot, an auxiliary laptop, on-robot sensors, and the ground truth respiration belt.

Beyond rPPG, research has shown that pixel movement (i.e., optical flow) or changes in pixel intensity can successfully pick up on respiration rates [31–33]. Using a technique based in pixel movement is promising for deep breathing phase classification given that extensions of the inhales, holds, and exhales often leads to more physically exaggerated motions than typical breathing. Thus, our modeling technique utilizes optical flow as one of its primary inputs (Section 4.1).

3 OMMDB DATA COLLECTION

3.1 Data Collection System

In order to best support new robot features with deep breathing, we sought to collect training data with two machine learning goals in mind: generalizability across unique individuals and generalizability across different robot ego-motion conditions. We thus chose to create the Open Multimodal Deep Breathing dataset (OMMDB) with an *Ommie* robot [34], which can support deep breathing on both a table and in a lap, as well as with or without breathing motions for haptic guidance. We additionally saw the *Ommie* robot as a possible future platform for applying our deep breathing classification system into new features for skill development and personalization.

To capture information on the robot's ego-motion, we included position data from the robot's internal body motor as well as data from an inertial measurement unit (IMU) in OMMDB. We implemented a camera for perceiving how the user is deep breathing, and a respiration force belt for ground truth labeling of breathing phases. Figure 2 outlines the data collection setup, and the following sections provide further detail on each of these components.

3.1.1 Ommie Robot. Physically, the *Ommie* robot stands 11" (279.4 mm) high when in a neutral position, with its body made of a 7" (177.8 mm) diameter sphere. The top portion of *Ommie*'s body, where users most often place their hands during haptic interactions, moves vertically to provide the robot's "breathing" effect. Each breath consists of roughly 2/3" (16.93 mm) of physical displacement. *Ommie*'s head features a one degree-of-freedom motor that allows it to move its head up and down. Prior to data collection, this motor's

position can be adjusted based on the location of the users' head and shoulders with regards to the robot's field of view. The head can be still or nodding during breathing effects (the head remained still during our data collection).

Ommie contains an on-board 8GB Raspberry Pi 4B, which senses and controls the following I/O devices: two motors (one in the robot's head and one in the robot's body), a capacitive touch sensor, two TFT screens for the robot's eyes, and a speaker. The system runs the Raspberry Pi OS with ROS Noetic. For our dataset, motor position data is collected via the *Dynamixel SDK* and *Dynamixel Workbench Controllers* ROS packages.

3.1.2 Paired Laptop and Networking. In order to provide additional computational power, we utilized a System76 laptop running Ubuntu 20.02 connected to the robot's RaspberyPi via a CAT6 cable. ROS Noetic was installed on both the robot and laptop for interprocess communication and logging, with the laptop serving as the ROS master. The laptop also provided time synchronization via a *chrony* implementation of the Network Time Protocol.[2]

3.1.3 RGB Camera. A Raspberry Pi v2 Camera was placed on *Ommie*'s head as seen in Fig. 2 in order to best capture an individual user's head and shoulders. This camera features a Sony IMX219 8-megapixel sensor and was connected via a flexible camera serial interface (CSI) cable to the robot's internal Raspberry Pi. The camera was configured using the ROS package *raspinode* to capture at 30 frames per seconds at a compressed size of 640 x 360 pixels.

3.1.4 Inertial Sensor. We utilized an Adafruit 9-DOF Absolute Orientation IMU Fusion Breakout Board (BNO055) and accompanying Python libraries for capturing inertial data from the robot. We saw this information as potentially useful in determining when the robot is on a table or in the lap, as well as for stabilizing ego-motion from the camera's perspective while a robot is held in the lap. The IMU could also possibly pick up on deep breathing signals from the user's diaphragm movements when in the lap. Early testing indicated the following data streams from the IMU were the most relevant: (a) *linear acceleration* as a three-axis vector in m/s^2 and (b) *angular velocity* as a three axis vector in rad/s. We therefore capture these two data streams in the OMMDB dataset.

3.1.5 Respiration Belt. For ground truth data of individual respiratory signals, we utilized a Vernier Go Direct® Respiration Belt alongside the *godirect* Python module. This belt was placed around the chest and used force measurements (in Newtons) to indicate the physical displacement of the chest during deep breathing. The oscillatory force measurements were later used by annotators for labeling breathing phases (Section 3.2.2).

3.2 Data Collection Methodology

3.2.1 Collection Conditions. The OMMDB dataset was collected with four robot ego-motion conditions with three different deep breathing cadences each. All conditions exhibited audio chimes to guide the user through the phases of deep breathing, but only some included haptic interactions. The four robot ego-motion conditions were as follows:[3]

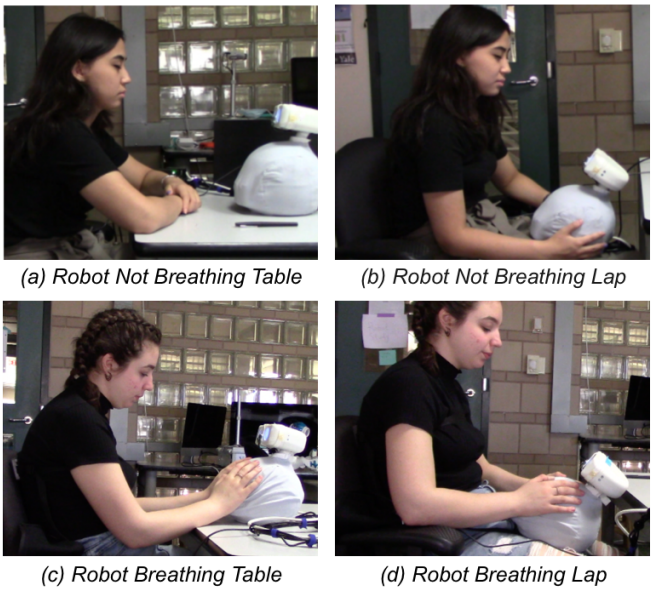

Figure 3: Participants deep breathing with an *Ommie* robot in the four collection conditions for OMMDB.

(1) *Robot Not Breathing Table* (Fig. 3a): The robot remained still and fixed on the table while the participant performed deep breathing in front of it.
(2) *Robot Not Breathing Lap* (Fig. 3b): The robot did not exhibit any breathing motions but experienced ego-motion from the user holding the robot in their lap while deep breathing.
(3) *Robot Breathing Table* (Fig. 3c): The robot's body expanded and contracted in correspondence to the deep breathing pattern, but the robot remained fixed on the table with the head unmoving. Users placed their hands on the robot to feel the robot's breathing.
(4) *Robot Breathing Lap* (Fig. 3d): The robot's body expanded and contracted in correspondence to the deep breathing pattern while held in a participant's lap. Users placed their hands on the robot to feel the robot's breathing. The robot's head did not move on its own, but the camera experienced ego-motion from the user's movements.

For each motion condition, participants performed three, 90-second breathing sessions, one in each of the following cadences:

(1) 3-2-3 pattern: the participant would inhale for three seconds, hold their breath for two seconds, exhale for three seconds, pause briefly, then repeat.
(2) 4-4-4-4 pattern: the participant would inhale for four seconds, hold for four seconds, exhale for four seconds, hold again for four seconds, then repeat.
(3) 5-3-5 pattern: the participant would inhale for five seconds, hold for three seconds, exhale for five seconds, pause briefly, then repeat.

[2]This implementation achieved time synchronization with disparities of $< .00005$ seconds.

[3]Video footage of the four conditions can be observed in the supplemental video.

These patterns were based on expert guidance from a local university psychology clinic and on the popular "box breathing" cadence [38]. The patterns were selected to provide variety in the lengths of breathing phases in the dataset while tempering the number of variations the participant needed to learn and produce.

3.2.2 Ground Truth Labeling. The respiration curves collected using the respiration belt were hand-labeled by two team members to capture phases of deep breathing. Labeling was performed using the open source graphics-based platform *Label Studio* [52]. Due to the subjectivity of the annotation process, a set of rules was generated to help labeling remain consistent. For example, the end of a top hold was defined as: "a sudden inflection that begins a steady and sustained drop in force measurements." Inter-rater reliability was assessed using percentage matching and Cohen's Kappa of discretized label values with 10% of the data. Percentage matching between the two annotators was 93% with an average Cohen's Kappa value of .89.

3.2.3 Post-Processing. All data streams were down-sampled from their original collection rates to 10 Hz.[4] Due to differences in collection rates, we synchronized sensor streams starting with the latest recorded start time for any of the sensors. From this start time, samples were drawn every .1 seconds for as long as there remained data for all sensors (roughly 900 timesteps). To enable future online processing of sensor streams, the closest data point prior to the next timestep was stored.

3.3 Dataset Population

A total of 50 individuals (38 Female, 11 Male, 1 Nonbinary) without any known history of severe respiratory disease were recruited for the data collection. The recruitment process included online platforms, word of mouth, and flyers. Given the disproportionately high rates of mental health challenges in young adults [13, 36] and a number of social robots targeting these individuals (e.g., [20, 34]), we restricted the age of participants to be 18-28. The data from 3 individuals was completely removed from the dataset due to technical problems occurring during data collection that affected all of their collected sessions. Therefore, the final dataset contains data from 47 individuals (36 Female, 10 Male, 1 Nonbinary), with an average age of $M = 22.17$ ($SD = 2.86$). We chose to retain a higher number of female-identifying individuals in our dataset given that mental health challenges disproportionately affect women [55]. The recruitment and data collection process was approved by our local Institutional Review Board.

3.4 Data Breakdown

A total of 330 breathing sessions were collected from the 47 individuals in the dataset.[5] Each session lasted 90 seconds. 50 sessions were removed due to technical issues with data collection. Therefore, there are a total of 280 sessions in the dataset, of which 65 are in the *Robot Not Breathing Table* condition, 66 are in the *Robot Not Breathing Lap* condition, 76 are in the *Robot Breathing Table* condition, and 73 are in the *Robot Breathing Lap* condition. Across the entirety of the dataset, 28% of the data points are inhales, 31% of the data points are exhales, and 41% of the data points are holds.

4 CLASSIFICATION METHODS

In this section, we study the performance of a multimodal, recurrent deep learning model for classifying deep breathing phases. Our problem consists of classifying a temporal sequence of observations of a user (as captured in OMMDB) into one of three classes: inhale, exhale, or hold. Our learning architecture utilizes optical flow, inertial, and motor positional data as inputs to a convolutional LSTM (outlined in Fig. 4). This system was designed in consideration of: (1) the inherent physical motion of the user when performing deep breathing, (2) the time-series nature of the data, and (3) the cyclical nature of the data.

Given that deep breathing consists of purposefully extended inhales, holds, and exhales, the physical motion of the user during deep breathing is often greater than with normal shallow breathing. This translates into more noticeable movements of the diaphragm, chest, and shoulders. We were therefore inspired to utilize optical flow in our work, a common technique from activity recognition [53] that analyzes the relative motion of pixels between two sequential images. Prior work with respiration rate detection has used this approach successfully [31–33]. We thus utilized this technique to capture features related to the user's breathing motion more directly than raw RGB signals.

Given the temporal nature of the data, we chose to implement a time-series sequence modeling approach. We did so in the form of a commonly-used recurrent architecture: a long short-term memory (LSTM) [16] neural network. We expected the LSTM to learn the patterns of deep breathing based on their cyclical nature.

4.1 Data Pre-Processing

All three types of data from the OMMDB dataset are processed prior to modeling, as shown in Figure 4. The 3-channel RGB image data is first converted to 2-channel optical flow data using *OpenCV*'s implementation of the dense Farnebeck algorithm [9]. Optical flow data is then z-score normalized. The IMU data (3-dimensional linear acceleration and 3-dimensional angular velocity) is also z-score normalized, and the motor position value is min-max normalized.[6]

The data is then split into smaller sequences for model training. Each deep breathing session in the dataset is approximately 90 seconds in duration. We utilized a rolling-window approach to classify a series of observations into a series of predictions, each window consisting of 1 second of data (10 frames). Each new window started 5 frames after the prior one, resulting in consecutive windows that overlapped for half of their observations.

[4]The original collection rates are as follows. RGB Camera: 30Hz, Respiration Belt: 20 Hz, Inertial Sensor: 24 Hz, Motor Position: 21 Hz.

[5]Data collection occurred in two phases. During the first phase, the two *Robot Not Breathing* conditions were collected. During the second phase, the two *Robot Breathing* conditions were collected. Some participants participated in both phases of data collection, and therefore provided up to 12 sessions worth of data to the dataset (4 conditions x 3 breathing patterns). Other participants only participated in one phase of data collection, and therefore provided up to 6 sessions worth of data to the dataset (2 conditions x 3 breathing patterns).

[6]The optical flow data was chosen to be z-score normalized standardized based on higher performance than min-max normalization. The IMU data was chosen to be z-score normalized based on the high levels of noise in the data and occasional anomalies, which would yield non-representative minimum and maximum values. The motor position values were min-max normalized, however, given that values are always between the same numerical range.

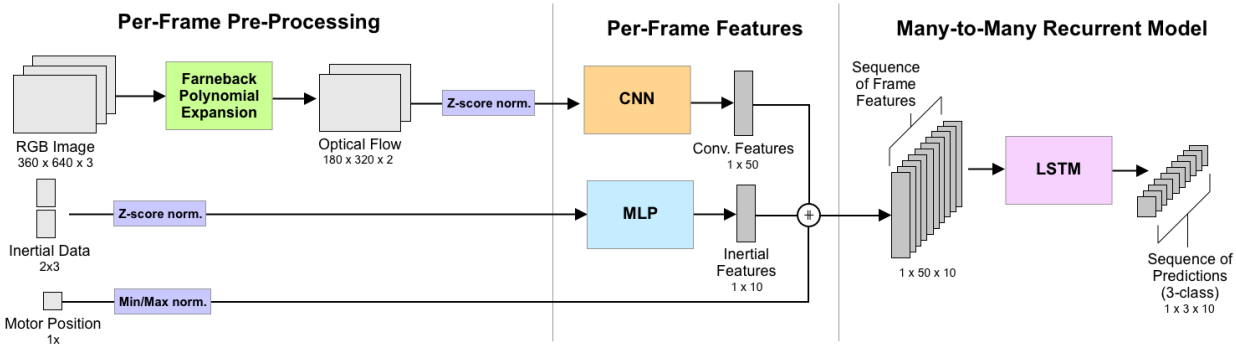

Figure 4: A visual overview of deep breathing phase classification. Each frame (i.e. time step) of data includes two-channel optical flow data, inertial data in the form of linear acceleration and angular velocity, and a single motor position value. The data is initially processed per-frame prior to being re-formed into a sequence for recurrent modeling via LSTM.

4.2 Architecture

4.2.1 Feature Extraction. Prior to sequence modeling with the LSTM, we implemented a convolutional neural network (CNN) and multi-layer perceptron (MLP) to extract latent features from the optical flow and inertial inputs respectively. The CNN consists of two ReLU-activated 2-D convolutional layers with maxpooling in between. The flattened outputs of the convolutional layers are passed through a ReLU-activated fully connected layer resulting in 50 convolutional features. The MLP expands the linear acceleration and angular velocity data into separate latent spaces, with a single ReLU-activated fully connected layer each. The two individual outputs are then concatenated and passed into another ReLU-activated fully connected layer, resulting in 10 inertial features.

4.2.2 Sequence Modeling. We implemented sequence-to-sequence training of the LSTM with hidden size 50 to predict one class per frame for 10-frame sequences of input data. Further details on the LSTM's hyperparameters are in Section 4.3.

4.3 Training Details

We utilized the *PyTorch* library to implement our architecture [41]. The training machine was equipped with four Nvidia GeForce RTX 2080 Ti GPUs, an Intel i7 3.7GHz processor, and 32 GB of RAM. Training utilized the Adam optimization algorithm [24], with early stopping implemented after 20 epochs of no improvement in validation loss. We chose model hyperparameters based on a staged search for: number of convolutional features (ranging from 10 to 500), number of inertial feature (ranging from 5 to 50), number of LSTM layers (one or two), number of LSTM features (ranging from 5 to 500), learning rate (.0001, .00001), and regularization techniques (.5 dropout; .001, .01, and .1 weight decay; and none). The search resulted in the following parameters based on the performance of validation data subsets: a single LSTM layer, a learning rate of 0.0001, and regularization using a weight decay value of 0.1.

4.4 Evaluation

We evaluated our model with two cross-validation techniques based on our generalization goals (Section 3.1). Regarding individual generalizability, we split the dataset into five folds that each contained unique participants, with each fold containing both female and non-female identifying individuals. Each fold also contained similar numbers of each of the four ego-motion conditions (i.e., *Robot Breathing* versus *Robot Not Breathing*, and robot on the *Table* versus *Lap*). Due to data cleaning, some folds had uneven numbers of sessions by condition. Five models were then trained, each with a separate fold held out for testing and a separate fold for validation.

Regarding ego-motion generalizability, we split the dataset into 4 folds. Each fold represented one of the four ego-motion conditions and had all participants represented. We withheld 20% of each fold for testing and 10% for validation. We then trained five models: four models each trained on the training subset corresponding to one of the four ego-motion conditions, and one model trained on all the training subsets from all the conditions. The four models trained on a single ego-motion condition were tested on the withheld test subset of the corresponding condition. The model trained on all four ego-motion conditions was tested on all four withheld test subsets independently. The purpose of this evaluation method was to discern the trade-off in performance between the two models per condition. Ideally for real-world applications, a single model could be used across multiple robot ego-motion conditions without a large drop in performance.

For all cross-validation experiments, model testing was performed based on individual breathing sessions. Each breathing session consisted of a single participant, in a single ego-motion condition, in a specific breathing cadence (e.g., participant 101 in the *Robot Not Breathing Table* condition breathing in a 3-2-3 cadence). Each session's data was passed frame-by-frame into the model recurrently along with the prior hidden and cell states from the LSTM. This methodology was chosen in order to support real-world applications of the algorithm, where data would be incoming in an online streaming manner. When evaluating each fold, frame-by-frame predictions from each session were concatenated together, and the F1 score was calculated on the entire list of predictions.

5 RESULTS

The results of our two cross-validation techniques are detailed below, along with a comparison to naive classifiers.

Table 1: Per-Individual Cross-Validation Groups and Results. "RNB" indicates the *Robot Not Breathing* conditions and "RB" indicates the *Robot Breathing* conditions.

Fold	# Indvls.	# RNB Table	# RNB Lap	# RB Table	# RB Lap	Test F1
1	9	13	13	15	15	0.79
2	9	13	15	14	14	0.81
3	9	15	12	15	12	0.74
4	9	15	15	14	12	0.76
5	10	9	11	18	18	0.81
					Avg.	**0.78**

5.1 Naive Classifiers

For comparison, we present the following F1 scores for naive classifiers based on our dataset. Because of the distribution of classes within the dataset, a classifier that always selects the dominant class (hold) would result in an F1 score of 0.15. A classifier that randomly samples from the distribution of classes (running the sampling 10 times) would result in an F1 score of 0.34. A classifier that used weighted sampling from the distribution of the classes (running the sampling 10 times) would result in an F1 score of 0.33.

5.2 Per-Individual Cross-Validation

Table 1 shows cross-validation results from splitting the data into five folds with unique participants. F1 scores for each fold ranged from 0.74 to 0.81 with an average of 0.78. Individual participant results can be found in Appendix A and range from .48 to .89 average F1. Of note, seven out of 47 participants (15%) resulted in average F1 scores below .68 (double the best performing naive classifier). An analysis of these participants is presented in Section 6.1.

5.3 Per-Condition Cross-Validation

Table 2 shows cross-validation results when splitting the data into folds by robot ego-motion condition. For models tested on the same condition from which they were trained on, F1 scores ranged from 0.78 to 0.84. In comparison, the model trained on all four conditions achieved F1 scores ranging from 0.76 to 0.83. Three out of four conditions perform slightly better when using a model trained on data from the same condition. However, a model trained on all four conditions performs comparably. The difference between testing on the all-conditions model and the models from the same conditions ranges from -0.04 to +0.01 per fold (see Δ F1 column in Table 2).

5.4 Ablation Study

We performed an ablation study to further understand the value of adding ego-motion inputs (inertial and motor position data) to the RGB inputs. After retraining the four conditions-based models using solely RGB inputs, we found that including ego-motion inputs, in addition to RGB inputs, marginally improved performance with an increase of F1 score ranging between .003 to .012. We suspect the relatively small increase is due to the fact that the image inputs received more complex feature processing compared to the motion signals. More sophisticated fusion mechanisms, such as attention [54], could potentially help the models better take advantage of ego-motion inputs in the future.

Table 2: Models Trained on Different Interaction Conditions. "Same" refers to a model trained only on data from the same motion condition as the test condition. "All" refers to a model trained on training data from all four motion conditions.

Test Condition	Same F1	All F1	Δ F1
Robot Breathing Table	0.84	0.81	-0.03
Robot Not Breathing Table	0.78	0.79	+0.01
Robot Not Breathing Lap	0.84	0.83	-0.01
Robot Breathing Lap	0.80	0.76	-0.04

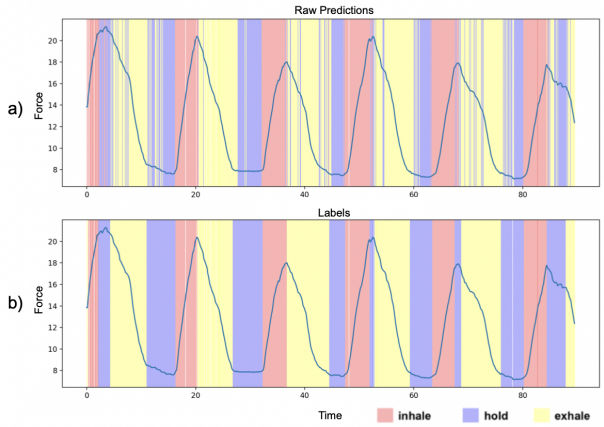

Figure 5: An example of a classified deep breathing session. (a) Shows the raw predictions from our classifier (F1 of 0.85) and (b) shows the ground truth labels.

6 DISCUSSION

6.1 Generalizing Across Individuals and Ego-Motion

One of our model training goals was to develop a robotic perception system that could generalize to unseen individuals. In practice, such generalization would mean that we would not need to train individual models for each robot user. The 5-fold cross-validation results of our learning architecture shows promise towards individual generalizability with an average F1 score of .78. However, as noted in Section 5.2, seven out of 47 participants resulted in poor performance (as characterized by an F1 score below two times the F1 score of the best naive classifier). A post-hoc analysis of these individuals resulted in a few insights as to why.

Of the seven individuals with average F1 scores below .68, three (PIDs 110_201[7], 104_202, 116) have a mixture of both high and low F1 scores across their breathing sessions. Upon visual observation of video data for these sessions, those with poor F1 performance do not exhibit large motion of the chest or shoulders in the field of

[7]Participants that performed deep breathing in both data collection phases are labeled with two PIDs but are considered to be one unique individual during analysis.

view. For these three participants, the sessions with high F1 scores exhibit visible shoulder motions while deep breathing.

Of the remaining four individuals with poor model performance, two had anomalies in their input data. One (PID 113) had irregular patterns in their ground truth data, possibly due to technical difficulties with the respiration belt. The other (PID 119) exhibited highly variable physical motions (beyond deep breathing) in comparison to the rest of the dataset population. These motions included head shaking and shifting seating positions. For the remaining two individuals (PIDs 108, 210), the root cause of poor model performance remained unclear. It is possible that our model is unable to generalize for these two individuals (4% of participants) based on some underlying latent characteristic.

We additionally sought to train a deep learning model that could generalize to multiple interaction conditions based on the robot's ego-motion. For a robot that can be used in multiple human-robot interaction settings (e.g., both on a table and in a lap, such as *Ommie* [34] or *Paro* [46]), utilizing a single on-board model reduces complexity in comparison to training multiple, condition-specific models. Our evaluation with per-condition cross-validation shows promise towards deploying robots with a single model for deep breathing phase classification. While we had naturally expected each motion condition to perform the best when training on its own condition, the performance of the model trained on data from all conditions was surprisingly comparable. We suspect this ability is due to the relative similarity in field of view that was possible from each condition. For robots that exist in just one ego-motion condition (e.g., CAKNA [2]), the model with the best performance for that condition can be implemented.

6.2 Haptic Interaction and Model Performance

A deeper analysis of both per-participant and per-condition cross-validation results shows that the *Robot Not Breathing Table* condition, the only without any haptic interaction, performs differently from the others. Based on the per-individual cross-validation F1 scores for each individual breathing session, *Robot Not Breathing Table* is disproportionately represented in poor performing sessions. The percentage of F1 scores below .68 is more than double in the *Robot Not Breathing Table* condition (26%) than in all three of the other conditions (11%). With per-condition cross-validation, the *Robot Not Breathing Table* condition was the one motion condition where performance increased slightly when trained on all conditions.

We suspect that the above patterns may be because of postural differences when individuals are physically handling the robot compared to when they are sitting in front of a robot. This difference can be observed visually in the videos of participants with a mixture of poor and high model performance. For these participants, the *Robot Not Breathing Table* videos show little chest and shoulder movement, while the *Lap* based videos show more visible motions. It may be that holding the robot in a lap, or placing one's hands on a robot on a table, encourages more physical motions of the chest and shoulders during deep breathing. It may also place the robot in a better position to observe these motions. A postural difference could also explain why testing our architecture on the *Robot Not Breathing Table* condition saw a slight benefit from training on all motion conditions. It is possible that this ego-motion condition benefits from training data with more exaggerated shoulder and chest motions from other conditions.

6.3 Limitations

There are several limitations to our work that are worth noting. First, the OOMDB dataset could be more diverse. The number of female individuals in our dataset is about three times the number of non-female individuals. Additionally, the majority of the individuals identify as White or Asian, with a small minority of individuals identifying as Black/African American or Hispanic/Latino/a. Second, despite our efforts to perform annotation as consistently and as objectively as possible, there is still inherent subjectivity in the ground truth labels of our dataset. Finally, our model relies on a fairly stable camera image that includes the participant's shoulders and at least part of their head. The participant must also exhibit clear motions in their shoulders or chest when deep breathing. An open question for our work is how to address deviations in ideal prediction conditions for in-the-wild robotic deployments. One option is to utilize probabilistic predictions in order for the robot to make interaction decisions based on the confidence level of the system. Other options could include adjusting the learning architecture to capture finer grained user movements, or experimenting with new input features based on posture recognition.

7 CONCLUSION

In this paper, we presented OMMDB, a multimodal dataset for deep breathing phase classification, and experimental results using the dataset with a recurrent deep learning architecture. We implemented a non-contact, robot-based perception system on an *Ommie* robot, which includes an RGB camera, IMU, and motor encoder. Ground truth labels of deep breathing phases were collected via a respiration belt and hand annotated. In alignment with our motivation of enabling new interaction modalities for social robots for mental health, OMMDB includes different recording conditions based on the ego-motion of the robot. We used this data to train a convolutional LSTM with optical flow and robot motion inputs for deep breathing phase classification. To our knowledge, we are the first to explore multimodal deep breathing phase classification with robot data, and our results show promise in applying deep learning techniques towards this aim. In particular, recurrent neural network models can be used to track the deep breathing phases of unique individuals across varying levels of robot ego-motion in human-robot interaction settings.

ACKNOWLEDGMENTS

We would like thank Nathan Tsoi, Kate Candon, and Alex Yuan for their assistance and feedback. This work was funded by the National Science Foundation (NSF) under grants No. 2106690, 1955653, and 1928448. Ellie Mamantov is supported by the NSF GRFP. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] Acconeer. [n. d.]. Acconeer/acconeer-python-exploration: Acconeer exploration tool. https://github.com/acconeer/acconeer-python-exploration

[2] Azizi Ab Aziz, Ali Saad Fahad, and Faudziah Ahmad. 2017. CAKNA: A personalized robot-based platform for anxiety states therapy. In *Intelligent Environments 2017*. IOS Press, 141–150.

[3] Ameen Bin Obadi, Ping Jack Soh, Omar Aldayel, Muataz Hameed Al-Doori, Marco Mercuri, and Dominique Schreurs. 2021. A Survey on Vital Signs Detection Using Radar Techniques and Processing With FPGA Implementation. *IEEE Circuits and Systems Magazine* 21, 1 (2021), 41–74. https://doi.org/10.1109/MCAS.2020.3027445

[4] S. Coşar, Z. Yan, F. Zhao, T. Lambrou, S. Yue, and N. Bellotto. 2018. Thermal Camera Based Physiological Monitoring with an Assistive Robot. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 5010–5013. https://doi.org/10.1109/EMBC.2018.8513201

[5] Mark É Czeisler, Rashon I Lane, Emiko Petrosky, Joshua F Wiley, Aleta Christensen, Rashid Njai, Matthew D Weaver, Rebecca Robbins, Elise R Facer-Childs, Laura K Barger, et al. 2020. Mental health, substance use, and suicidal ideation during the COVID-19 pandemic—United States, June 24–30, 2020. *Morbidity and Mortality Weekly Report* 69, 32 (2020), 1049.

[6] Gerard de Haan and Vincent Jeanne. 2013. Robust Pulse Rate From Chrominance-Based rPPG. *IEEE Transactions on Biomedical Engineering* 60, 10 (2013), 2878–2886. https://doi.org/10.1109/TBME.2013.2266196

[7] Justin R Estepp, Ethan B Blackford, and Christopher M Meier. 2014. Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography. In *2014 IEEE international conference on systems, man, and cybernetics (SMC)*. IEEE, 1462–1469.

[8] Justin R. Estepp, Ethan B. Blackford, and Christopher M. Meier. 2014. Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 1462–1469. https://doi.org/10.1109/SMC.2014.6974121

[9] Gunnar Farnebäck. 2003. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*. Springer, 363–370.

[10] Ronit Feingold Polak and Shelly Levy Tzedek. 2020. Social Robot for Rehabilitation: Expert Clinicians and Post-Stroke Patients' Evaluation Following a Long-Term Intervention. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 151–160. https://doi.org/10.1145/3319502.3374797

[11] Luay Fraiwan, Natheer Khasawneh, Khaldon Lweesy, Mennatalla Elbalki, Amna Almarzooqi, and Nada Abu Hamra. 2021. Non-contact spirometry using a mobile thermal camera and AI regression. *Sensors* 21, 22 (2021), 7574.

[12] Robin E Gearing, Lisa Townsend, Jennifer Elkins, Nabila El-Bassel, and Lars Osterberg. 2014. Strategies to predict, measure, and improve psychosocial treatment adherence. *Harvard review of psychiatry* 22, 1 (2014), 31–45.

[13] Renee D Goodwin, Andrea H Weinberger, June H Kim, Melody Wu, and Sandro Galea. 2020. Trends in anxiety among adults in the United States, 2008–2018: Rapid increases among young adults. *Journal of psychiatric research* 130 (2020), 441–446.

[14] Zixiong Han. 2021. *Respiratory Patterns Classification using UWB Radar*. Ph. D. Dissertation. Université d'Ottawa/University of Ottawa.

[15] Mirae Harford, Jacqueline Catherall, Stephen Gerry, John Duncan Young, and P Watkinson. 2019. Availability and performance of image-based, non-contact methods of monitoring heart rate, blood pressure, respiratory rate, and oxygen saturation: a systematic review. *Physiological measurement* 40, 6 (2019), 06TR01.

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[17] Susan I Hopper, Sherrie L Murray, Lucille R Ferrara, and Joanne K Singleton. 2019. Effectiveness of diaphragmatic breathing for reducing physiological and psychological stress in adults: a quantitative systematic review. *JBI Evidence Synthesis* 17, 9 (2019), 1855–1876.

[18] Preeti Jagadev and Lalat Indu Giri. 2020. Non-contact monitoring of human respiration using infrared thermography and machine learning. *Infrared Physics & Technology* 104 (2020), 103117.

[19] Shomik Jain, Balasubramanian Thiagarajan, Zhonghao Shi, Caitlyn Clabaugh, and Maja J Matarić. 2020. Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders. *Science Robotics* 5, 39 (2020), eaaz3791.

[20] Sooyeon Jeong, Sharifa Alghowinem, Laura Aymerich-Franch, Kika Arias, Agata Lapedriza, Rosalind Picard, Hae Won Park, and Cynthia Breazeal. 2020. A robotic positive psychology coach to improve college students' wellbeing. IEEE RO-MAN.

[21] Ravinder Jerath, Molly W Crawford, Vernon A Barnes, and Kyler Harden. 2015. Self-regulation of breathing as a primary treatment for anxiety. *Applied psychophysiology and biofeedback* 40, 2 (2015), 107–115.

[22] Ravinder Jerath, John W Edry, Vernon A Barnes, and Vandna Jerath. 2006. Physiology of long pranayamic breathing: neural respiratory elements may provide a mechanism that explains how slow deep breathing shifts the autonomic nervous system. *Medical hypotheses* 67, 3 (2006), 566–571.

[23] Cory D. Kidd and Cynthia Breazeal. 2008. Robots at home: Understanding long-term human-robot interaction. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 3230–3235. https://doi.org/10.1109/IROS.2008.4651113

[24] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

[25] Agni Kumar, Vikramjit Mitra, Carolyn Oliver, Adeeti Ullal, Matt Biddulph, and Irida Mance. 2021. Estimating Respiratory Rate From Breath Audio Obtained Through Wearable Microphones. *arXiv preprint arXiv:2107.14028* (2021).

[26] Egor Lakomkin, Mohammad Ali Zamani, Cornelius Weber, Sven Magg, and Stefan Wermter. 2018. On the Robustness of Speech Emotion Recognition for Human-Robot Interaction with Deep Neural Networks. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 854–860. https://doi.org/10.1109/IROS.2018.8593571

[27] Min Kyung Lee, Jodi Forlizzi, Sara Kiesler, Paul Rybski, John Antanitis, and Sarun Savetsila. 2012. Personalization in HRI: A longitudinal field experiment. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 319–326.

[28] Iolanda Leite, Carlos Martinho, and Ana Paiva. 2013. Social robots for long-term interaction: a survey. *International Journal of Social Robotics* 5, 2 (2013), 291–308.

[29] Daniel Leyzberg, Samuel Spaulding, and Brian Scassellati. 2014. Personalizing robot tutors to individuals' learning differences. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 423–430.

[30] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. 2020. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *arXiv preprint arXiv:2006.03790* (2020).

[31] Carlo Massaroni, Daniela Lo Presti, Domenico Formica, Sergio Silvestri, and Emiliano Schena. 2019. Non-contact monitoring of breathing pattern and respiratory rate via RGB signal measurement. *Sensors* 19, 12 (2019), 2758.

[32] Carlo Massaroni, Emiliano Schena, Sergio Silvestri, and Soumyajyoti Maji. 2019. Comparison of two methods for estimating respiratory waveforms from videos without contact. In *2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. 1–6. https://doi.org/10.1109/MeMeA.2019.8802167

[33] Carlo Massaroni, Emiliano Schena, Sergio Silvestri, Fabrizio Taffoni, and Mario Merone. 2018. Measurement system based on RBG camera signal for contactless breathing pattern and respiratory rate monitoring. In *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. 1–6. https://doi.org/10.1109/MeMeA.2018.8438692

[34] Kayla Matheus, Marnyel Vázquez, and Brian Scassellati. 2022. A Social Robot for Anxiety Reduction via Deep Breathing. IEEE RO-MAN.

[35] Lucia McBee. 2003. Mindfulness practice with the frail elderly and their caregivers: Changing the practitioner-patient relationship. *Topics in Geriatric Rehabilitation* 19, 4 (2003), 257–264.

[36] Kathleen Ries Merikangas, Jian-ping He, Marcy Burstein, Sonja A Swanson, Shelli Avenevoli, Lihong Cui, Corina Benjet, Katholiki Georgiades, and Joel Swendsen. 2010. Lifetime prevalence of mental disorders in US adolescents: results from the National Comorbidity Survey Replication–Adolescent Supplement (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry* 49, 10 (2010), 980–989.

[37] Venkata Srikanth Nallanthighal, Zohreh Mostaani, Aki Härmä, Helmer Strik, and Mathew Magimai-Doss. 2021. Deep learning architectures for estimating breathing signal and respiratory parameters from speech recordings. *Neural Networks* 141 (2021), 211–224.

[38] Samantha K Norelli, Ashley Long, and Jeffrey M Krepps. 2018. Relaxation techniques. (2018).

[39] Jelena Obradović, Michael J Sulik, and Emma Armstrong-Carter. 2021. Taking a few deep breaths significantly reduces children's physiological arousal in everyday settings: Results of a preregistered video intervention. *Developmental Psychobiology* 63, 8 (2021), e22214.

[40] Cara O'Brien, Molly O'Mara, Johann Issartel, and Conor McGinn. 2021. Exploring the Design Space of Therapeutic Robot Companions for Children. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21)*. Association for Computing Machinery, New York, NY, USA, 243–251. https://doi.org/10.1145/3434073.3444669

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[42] Valentina Perciavalle, Marta Blandini, Paola Fecarotta, Andrea Buscemi, Donatella Di Corrado, Luana Bertolo, Fulvia Fichera, and Marinella Coco. 2017. The role of deep breathing on stress. *Neurological Sciences* 38, 3 (2017), 451–458.

[43] Valentina Perciavalle, Marta Blandini, Paola Fecarotta, Andrea Buscemi, Donatella Di Corrado, Luana Bertolo, Fulvia Fichera, and Marinella Coco. 2017. The role of deep breathing on stress. *Neurological Sciences* 38, 3 (2017), 451–458.

[44] Nuerzati Resuli, Marjorie Skubic, and Jung Myungki. 2021. Noninvasive Respiration Monitoring of Different Sleeping Postures Using an RF Sensor. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 1485–1490.

[45] Hayley Robinson, Bruce MacDonald, and Elizabeth Broadbent. 2015. Physiological effects of a companion robot on blood pressure of older people in residential care facility: a pilot study. *Australasian journal on ageing* 34, 1 (2015), 27–32.

[46] Selma Šabanović, Casey C Bennett, Wan-Ling Chang, and Lesa Huber. 2013. PARO robot affects diverse interaction modalities in group sensory therapy for older adults with dementia. In *2013 IEEE 13th international conference on rehabilitation robotics (ICORR)*. IEEE, 1–6.

[47] Nicole Salomons, Tom Wallenstein, Debasmita Ghose, and Brian Scassellati. 2022. The Impact of an In-Home Co-Located Robotic Coach in Helping People Make Fewer Exercise Mistakes. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 149–154.

[48] Yasaman S. Sefidgar, Karon E. MacLean, Steve Yohanan, H.F. Machiel Van der Loos, Elizabeth A. Croft, and E. Jane Garland. 2016. Design and Evaluation of a Touch-Centered Calming Interaction with a Social Robot. *IEEE Transactions on Affective Computing* 7, 2 (2016), 108–121. https://doi.org/10.1109/TAFFC.2015.2457893

[49] Vinothini Selvaraju, Nicolai Spicher, Ju Wang, Nagarajan Ganapathy, Joana M Warnecke, Steffen Leonhardt, Ramakrishnan Swaminathan, and Thomas M Deserno. 2022. Continuous Monitoring of Vital Signs Using Cameras: A Systematic Review. *Sensors* 22, 11 (2022), 4097.

[50] Anup Sharma, Marna S Barrett, Andrew J Cucchiara, Nalaka S Gooneratne, and Michael E Thase. 2017. A breathing-based meditation intervention for patients with major depressive disorder following inadequate response to antidepressants: a randomized pilot study. *The Journal of clinical psychiatry* 78, 1 (2017), 493.

[51] Micol Spitale, Minja Axelsson, and Hatice Gunes. 2023. Robotic Mental Well-Being Coaches for the Workplace: An In-the-Wild Study on Form. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 301–310. https://doi.org/10.1145/3568162.3577003

[52] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. https://github.com/heartexlabs/label-studio Open source software available from https://github.com/heartexlabs/label-studio.

[53] Amin Ullah, Khan Muhammad, Javier Del Ser, Sung Wook Baik, and Victor Hugo C. de Albuquerque. 2019. Activity Recognition Using Temporal Optical Flow Convolutional Features and Multilayer LSTM. *IEEE Transactions on Industrial Electronics* 66, 12 (2019), 9692–9702. https://doi.org/10.1109/TIE.2018.2881943

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[55] Oriana Vesga-López, Franklin R Schneier, Samuel Wang, Richard G Heimberg, Shang-Min Liu, Deborah S Hasin, and Carlos Blanco. 2008. Gender differences in generalized anxiety disorder: results from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). *Journal of Clinical Psychiatry* 69, 10 (2008), 1606.

[56] Xiaorong Yang, Yuan Fang, Hui Chen, Tongchao Zhang, Xiaolin Yin, Jinyu Man, Lejin Yang, and Ming Lu. 2021. Global, regional and national burden of anxiety disorders from 1990 to 2019: results from the Global Burden of Disease Study 2019. *Epidemiology and psychiatric sciences* 30 (2021).

[57] Steve Yohanan and Karon E MacLean. 2012. The role of affective touch in human-robot interaction: Human intent and expectations in touching the haptic creature. *International Journal of Social Robotics* 4, 2 (2012), 163–180.

[58] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. 2016. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3438–3446.

[59] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. 2016. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3438–3446.