Just Breathe: Harnessing Pretrained Audio Models for Mood Awareness in a Gamified Breathing Training App

Akshaye Shenoi*
ETH Zurich
Switzerland
Future Health Technologies,
Singapore-ETH Centre
Singapore
akshaye.shenoi@ethz.ch

Gisbert W. Teepe*
University Hospital of Old Age
Psychiatry and Psychotherapy,
University of Bern
Switzerland
gisbert.teepe@unibe.ch

Yanick X. Lukic
School of Engineering, Zurich
University of Applied Sciences
Winterthur
Switzerland
University of St. Gallen
Switzerland
yanick.lukic@zhaw.ch

Samarth Negi
ETH Zurich
Switzerland
Future Health Technologies,
Singapore-ETH Centre
Singapore
samarth.negi@sec.ethz.ch

Jacqueline Louise Mair
Future Health Technologies,
Singapore-ETH Centre
Singapore
jacqueline.mair@sec.ethz.ch

Elgar Fleisch
ETH Zurich
Switzerland
University of St. Gallen
Switzerland
efleisch@ethz.ch

Tobias Kowatsch
University of St. Gallen
Switzerland
Univerity of Zurich
Switzerland
tobias.kowatsch@unisg.ch

Abstract

Mood disorders can significantly impair an individual's ability to function in daily life. One common symptom observed in early stages of several mood disorders is *Agitation*. Early detection and regulation of agitation can thus help prevent the progression of severe mood disorders. Digital Health Interventions such as guided slow-paced breathing apps have demonstrated their efficacy in reducing agitation. However, such interventions are often hindered by poor adherence in the long term. We hypothesize that an awareness of the individual's current mood states would allow for a more personalized intervention (e.g., adaptive session duration), thereby improving adherence. In this paper, we explore the feasibility of predicting real-time agitation in an individual as they engage with a gamified biofeedback-based breathing training app *Breeze*. In a controlled lab experiment ($n = 30$), we collect voice and breathing samples from *Breeze* alongside self-reports of perceived agitation. Two pretrained audio models, VGGish and OPERA, are used to extract feature embeddings from raw audio signals, and downstream classifiers are subsequently trained to predict agitation levels. Initial results show that (1) respiratory audio is indeed a reliable predictor of agitation, and (2) pretrained models are capable of extracting meaningful features from respiratory audio. Our findings lay the technical groundwork to further investigate and evaluate the effect of mood-aware personalization strategies to improve user adherence, and ultimately, lead to better mental health outcomes.

CCS Concepts

• **Applied computing** → **Health informatics**; • **Human-centered computing** → **Interactive systems and tools**; *Ubiquitous and mobile computing systems and tools.*

ACM Reference Format:
Akshaye Shenoi, Gisbert W. Teepe, Yanick X. Lukic, Samarth Negi, Jacqueline Louise Mair, Elgar Fleisch, and Tobias Kowatsch. 2025. Just Breathe: Harnessing Pretrained Audio Models for Mood Awareness in a Gamified Breathing Training App. In *Proceedings of (CHI '25 Workshop on Envisioning the Future of Interactive Health).* ACM, New York, NY, USA, 4 pages.

1 Introduction

Mood disorders—including bi-polar and depressive disorders—are a category of mental health disorders characterized by persistent disturbances in an individual's emotional state [1]. A common mood symptom observed in patients with mood disorders is heightened *Agitation* which refers to a state of increased restlessness, irritability, or emotional disturbance. Clinicians regularly screen for self-experienced agitation in initial consultations and during the

*Both authors contributed equally to the paper.

CHI '25 Workshop on Envisioning the Future of Interactive Health, Yokohama, Japan

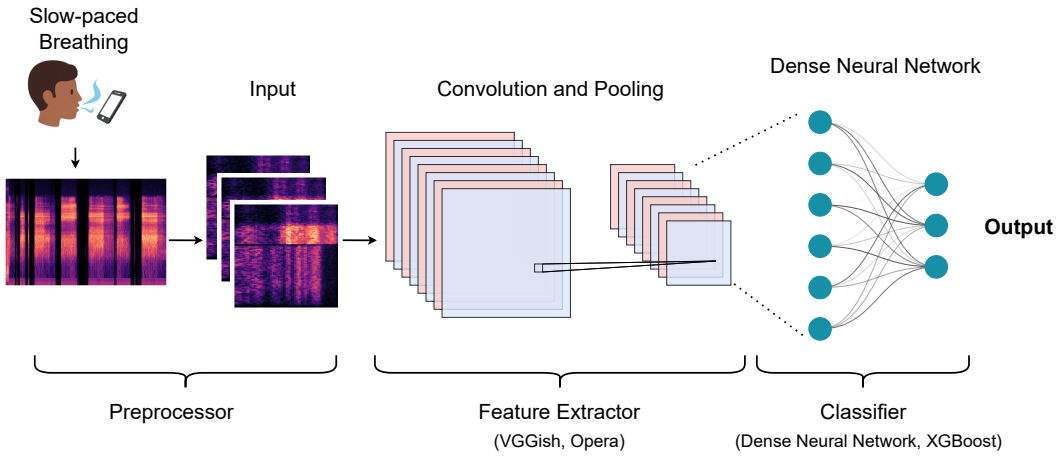


Figure 1: The Prediction Pipeline. Respiratory audio is captured from a smartphone mic as the user completes a slow-paced breathing session with *Breeze*. The Preprocessor cleans, crops and transforms the raw audio signal into a suitable format for the Feature Extractor. At this stage, pretrained audio models such as VGGish and OPERA extract semantically meaningful feature embeddings that are passed through a lightweight Classifier to obtain the final predictions.

course of treatment. Early diagnosis and intervention during heightened states of agitation can be instrumental in preventing mood disorders and improving outcomes. Digital health interventions, such as guided slow-paced breathing apps, have shown promise in this regard [6, 13]. However, despite their benefits, these interventions often face high dropout rates and poor adherence, limiting their effectiveness.

One possible reason for this challenge might be the lack of personalization—most interventions follow a one-size-fits-all approach, failing to adapt to an individual's momentary mood states. We hypothesize that integrating real-time *mood awareness* into such interventions could enhance engagement and adherence. Examples include dynamically adjusting the duration of the session or lowering the frequency based on user's agitation level. To explore this, we investigate the feasibility of predicting agitation using respiratory audio collected during interactions with Breeze [11], a gamified biofeedback-based breathing app (§2.1). We make use of transfer learning [16], leveraging two pretrained audio models—VGGish [8] and OPERA [15]—to automatically extract meaningful features from raw respiratory sounds (§2.3).

Our findings suggest that respiratory audio serves as a reliable predictor of agitation and that pretrained models can effectively extract meaningful features for mood classification (§3). These results establish a technical foundation for integrating mood-aware personalization strategies into digital health interventions, ultimately aiming to improve engagement, adherence, and mental health outcomes.

2 Methodology

2.1 Breeze

Breeze [11] is a slow-paced breathing training app that combines gamification, sensing and biofeedback to provide an immersive and engaging interaction. It adopts a maritime theme guiding participants to sail a boat across a river by *filling* the sail (exhaling) and *emptying* the sail (inhaling). The smartphone mic is used to detect inhalation/exhalation stages in real time. Several studies have shown the efficacy of *Breeze* in improving physiological and relaxation-related outcomes [9, 10].

2.2 Data Collection

A controlled lab study was conducted with two objectives: (1) measuring the effectiveness of *Breeze* in decreasing agitation and (2) predicting perceived pre- and post-agitation for each session. This paper focuses solely on the latter objective.

Ground Truth. To measure the ground truth, i.e., Perceived agitation, we used the *Calm–Restless* dimension from the Multidimensional Mood State Questionnaire (MDMQ) [14]. The subscale consists of five items: *composed*, *relaxed* and *absolutely calm* constituting the positive items, and *restless* and *uneasy* constituting the negative ones. A 6-point Likert scale (responses ranging from *not at all* [+1] to *very much* [+6]) assigns a score to each item. A single agitation score is computed as:

$$\begin{aligned} Agitation = &(Restless + Uneasy) \\ &- (Composed + Relaxed + 2 * AbsolutelyCalm) \end{aligned}$$

The resultant raw score in range $[-22, 8]$ is linearly scaled by adding 22 to obtain the final agitation score $y \in [0, 30]$.

Participants. Participants were recruited from ETH Zurich via e-mail and personal approach. The sessions were conducted in a controlled lab environment given the exploratory nature of the study. Upon providing informed consent and reading the information sheet, the participants proceeded as follows: a) answer the five MDMQ subscale items; b) utter three short voice commands to start the session; c) perform the slow-paced breathing for three minutes;

d) utter three voice commands to end the session; e) answer the MDMQ subscale items again.[1]

Data. At the study's conclusion, we collected the following data from the participants, each performing one *Breeze* session:

(1) *Agitation Scores:* Distinct pre- and post-session perceived agitation scores.
(2) *Breathing:* 3-minute long breathing clips. Intuitively, a participant's mood state is bound to change as they progress through the slow-paced breathing session. As a close approximation, we therefore associate the pre-session agitation scores to only the first 60 seconds of the 3-minute clip, and the post-session scores to the last 60 seconds only.
(3) *Voice:* 3 voice commands both before and after each session. We directly associate the pre- and post-agitation scores to the voice clips. On average, the combined voice commands per session last 6-8 seconds.

2.3 Mood Prediction: System Design

We illustrate our prediction pipeline in Figure 1.

Feature Extraction using Pretrained Audio Models. A key component of our proposal is the use of pretrained audio models that *automatically* extract high quality features from raw audio signals. Trained on large amounts of audio data[2], state-of-the-art models such as VGGish [8, 12] and OPERA [15] learn meaningful representations of audio signals. With minimal fine-tuning, these representations can be transferred to other downstream tasks that involve audio signals. This method stands in contrast to *manual* feature extraction where "handcrafted" features (e.g., jitter, shimmer, tone, etc.) are extracted from a signal and fed into the classifier directly. Such an approach may risk overfitting on the training dataset.

We use standard configurations for VGGish which takes 960ms clips as input per 128-dimension output feature embedding. For OPERA, we use the OPERA-CT variant—the best performing model according to the authors' evaluations [15]. We find that an input length of 4s per 768-dimension feature embedding produces best results.

Classification. Once the features are extracted from the raw signals, we train lightweight classifiers to classify the severity of agitation. We implement the following models: Dense Neural Network (DNN), Random Forest (RF) [2] and XGBoost (XGB) [3]. Given the skewed distribution of our dataset, we only attempt to classify the input as **low** agitation if $y \in [0, 10]$ or **moderate-to-high** agitation if $y \in [11, 30]$.

Metrics. The weighted average F1-score is used to measure the performance. We employ the Leave-One-Out Cross Validation (LOOCV) scheme to validate the performance.

3 Results

Participants. We recruited $n = 30$ participants, however, due to a technical error, we were unable to use the breathing audio for one participant. Thus, we ended up with $29 \times 2 = 58$ distinct pre/post audio samples and agitation scores (Mean = 9.51, SD = 4.59, Median = 9).

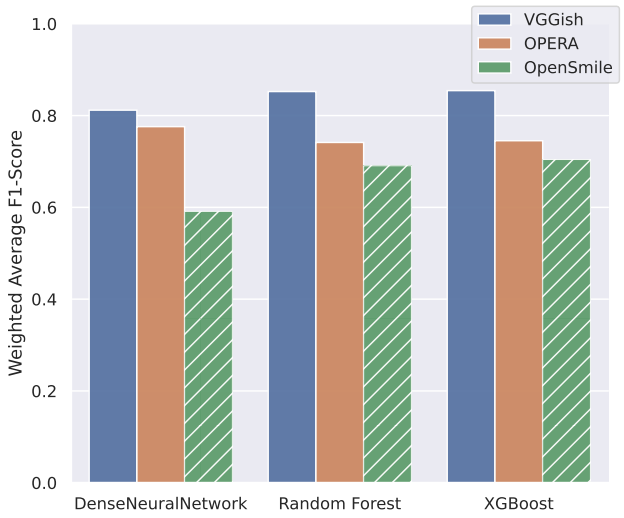


Figure 2: Weighted Average F1-Score for Perceived Agitation classification. VGGish and OPERA are automated feature extractors making use of extensive pretraining, whereas OpenSmile extracts features using a known acoustic feature set.

Baseline. To compare the performance of automated feature extraction using pretrained models, we added a commonly used acoustic feature set—eGeMAPS [4] with the *OpenSmile* toolkit [5]—as a baseline feature extractor. The input signal is split into 1-second clips, and 88 features are extracted for each clip. We use the same classifiers described in §2.3.

RQ1. Can breathing sounds accurately predict perceived agitation? We extract features using VGGish, OPERA and OpenSmile, and train the three downstream classifiers. Figure 2 plots the results. The weighted average F1-Score is consistently over 0.75 when the automated feature extractors are used, indicating 1) that breathing audio is indeed predictive of perceived agitation, and 2) the suitability of pretraining-based feature extraction over manual feature extraction for respiratory audio. When using the VGGish feature extractor along with the XGB classifier, we observe an F1-score of 0.83. Interestingly, VGGish outperforms OPERA despite the fact that OPERA is trained exclusively on respiratory audio.

RQ2. How does the breathing modality compare to voice? We only use VGGish to extract voice features given OPERA's non-suitability for non-respiratory audio. Similar to RQ1, we train three downstream classifiers. Results in Figure 3 suggest that both voice and breathing modalities have comparable performance (F1-Score $> .80$). However, note that compared to the 60 seconds of breathing audio per sample, voice samples are only 6-8 seconds long, which might limit the classifier's learning ability. Despite the limitation, our results indicate that the breathing modality is promising in it's own right.

[1] Participants also answered other surveys unrelated to the scope of this paper. Details of these surveys are omitted here for brevity.
[2] VGGish is trained on the AudioSet dataset [7] which contains 5.24 million hours of audio clips from YouTube, whereas OPERA is trained on 440 hours of respiratory audio exclusively.

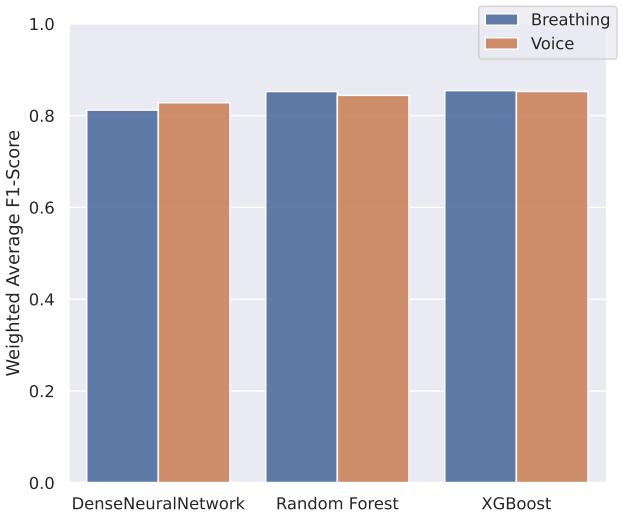


Figure 3: Enter Caption

4 Discussion

Our findings suggest that respiratory audio can effectively classify users' agitation severity as they interact with the slow-paced breathing app *Breeze*. Importantly, this ability opens up opportunities for real-time mood aware personalization in breathing interventions. Such an adaptive strategy can dynamically adjust parameters of the intervention (e.g., breathing pace) to improve long-term user adherence, and as a consequence, help limit the progression of mood disorders.

We also note that, from a methodological perspective, transferring learned representations encoded in large pretrained audio models to specific downstream tasks is highly efficient, especially when the training data size is limited.

While this study provides promising preliminary insights, several limitations must be considered. First, the sample size is relatively small, and the study was conducted in a controlled lab setting. More "in-the-wild" data are required to validate our findings for real-world scenarios. Second, the feasibility of running the prediction pipeline directly on the smartphone device must be evaluated. Lastly, we trained our models to classify severity as binary classes. A more granular, regression-based approach might enable finer and more personalized adjustments to the intervention parameters. We plan to address these limitations in future work.

References

[1] DSMTF American Psychiatric Association, DS American Psychiatric Association, et al. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*. Vol. 5. American psychiatric association Washington, DC.

[2] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[3] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. doi:10.1145/2939672.2939785

[4] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[5] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[6] Guy William Fincham, Clara Strauss, Jesus Montero-Marin, and Kate Cavanagh. 2023. Effect of breathwork on stress and mental health: A meta-analysis of randomised-controlled trials. *Scientific Reports* 13, 1 (2023), 432.

[7] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*. New Orleans, LA.

[8] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017-03). 131–135. doi:10.1109/ICASSP.2017.7952132

[9] Yanick Xavier Lukic, Shari Shirin Klein, Victoria Brügger, Olivia Clare Keller, Elgar Fleisch, and Tobias Kowatsch. 2021. The impact of a gameful breathing training visualization on intrinsic experiential value, perceived effectiveness, and engagement intentions: between-subject online experiment. *JMIR serious Games* 9, 3 (2021), e22803.

[10] Yanick Xavier Lukic, Alvaro Hernandez Reguera, Amanda Cotti, Elgar Fleisch, Tobias Kowatsch, et al. 2021. Physiological responses and user feedback on a gameful breathing training app: within-subject experiment. *JMIR serious Games* 9, 1 (2021), e22802.

[11] Yanick Xavier Lukic, Gisbert Wilhelm Teepe, Elgar Fleisch, and Tobias Kowatsch. 2022. Breathing as an Input Modality in a Gameful Breathing Training App (Breeze 2): Development and Evaluation Study. 10, 3 (2022), e39186. doi:10.2196/39186

[12] Karen Simonyan and Andrew Zisserman. 2015. *Very Deep Convolutional Networks for Large-Scale Image Recognition.* doi:10.48550/arXiv.1409.1556 arXiv:1409.1556 [cs]

[13] Patrick R Steffen, Tara Austin, Andrea DeBarros, and Tracy Brown. 2017. The impact of resonance frequency breathing on measures of heart rate variability, blood pressure, and mood. *Frontiers in public health* 5 (2017), 222.

[14] Rolf Steyer, Peter Schwenkmezger, Peter Notz, and Michael Eid. 2003. *Development of the Multidimensional Mood State Questionnaire (MDBF). Primary dataEntwicklung des Mehrdimensionalen Befindlichkeitsfragebogens (MDBF). Primärdatensatz.* doi:10.5160/PSYCHDATA.SRRF91EN15

[15] Yuwei Zhang, Tong Xia, Jing Han, Yu Wu, Georgios Rizos, Yang Liu, Mohammed Mosuily, Jagmohan Chauhan, and Cecilia Mascolo. 2024. *Towards Open Respiratory Acoustic Foundation Models: Pretraining and Benchmarking.* doi:10.48550/arXiv.2406.16148 arXiv:2406.16148 [cs, eess]

[16] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *Proc. IEEE* 109, 1 (2020), 43–76.

Just Breathe: Harnessing Pretrained Audio Models for Mood Awareness in a Gamified Breathing Training App

Akshaye Shenoi*, *Gisbert W. Teepe**, *Yanick X. Lukic, Samarth Negi, Jacqueline Louise Mair, Elgar Fleisch, Tobias Kowatsch*

Background

- Mood disorders, such as bipolar and depressive disorders, involve persistent emotional disturbances in mood states
- **Agitation** is a common early symptom across mood disorders; early detection and intervention for agitation can be crucial in preventing disorder progression
- Guided slow-paced breathing interventions can help manage agitation

Interventions often struggle with high dropouts and poor adherence

- Possible reason: inability to adapt to an individual's momentary mood states
- Real-time *mood awareness* could improve engagement, e.g., by dynamically adjusting session duration based on current agitation

Objective

Investigate the feasibility of predicting agitation from respiratory audio collected during interactions with a gamified breathing training app.

Methods

Breeze App
A maritime-themed breathing app that uses biofeedback to guide users through slow-paced breathing sessions [1].

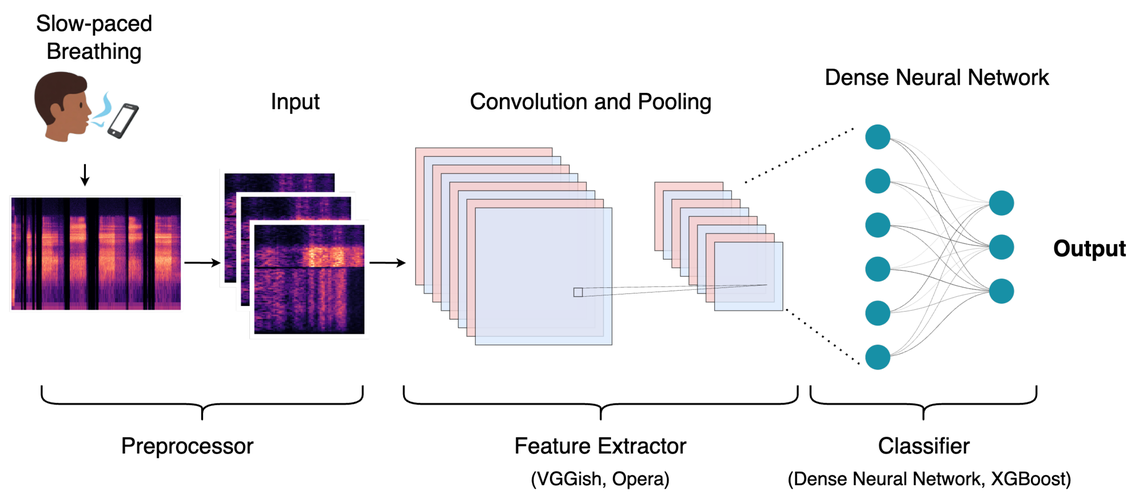
Participants
30 individuals from ETH Zurich completed one lab-based Breeze session each.

Ground Truth
- Self-reported *Perceived Agitation* as measured by the Calm-Restless dimension of the *Multidimensional Mood State Questionnaire* (MDMQ) scale
- 5 items rated on a 6-point Likert scale
  - Positive: composed, relaxed and absolutely-calm
  - Negative: restless and uneasy
- Single agitation score computed and scaled: $y \in [0, 30]$

Data Collection
- Participants completed one Breeze session each in a controlled lab study
- Following data were collected:
  1. *Agitation score:* Distinct pre- and post-session agitation scores.
  2. *Respiratory audio:* 3-minute-long breathing clips. First 60 seconds associated to pre-session agitation score, and last 60 seconds to post-session agitation score.
  3. *Voice commands:* 3 voice commands uttered in the beginning and end of each session.

[1] Lukic YX, Teepe GW, Fleisch E, Kowatsch T. *Breathing as an Input Modality in a Gameful Breathing Training App (Breeze 2): Development and Evaluation Study.* JMIR Serious Games 2022

Mood Prediction



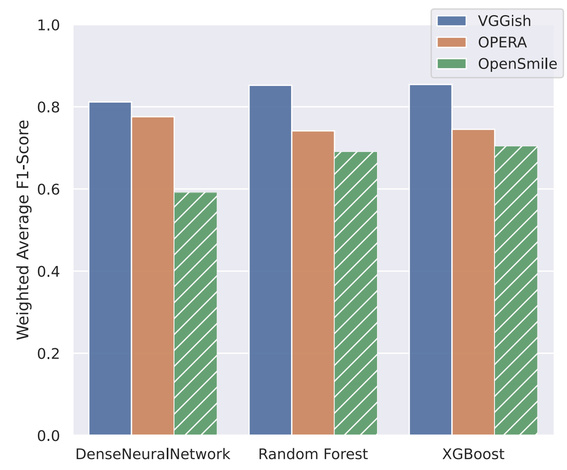
Preprocessing
Clean, crop and transform the raw audio signal into a suitable format for the Feature Extractor

Feature Extraction using Pretrained Audio Models
- Pretrained models—such as VGGish [2] (trained on general audio) and OPERA [3] (trained on respiratory audio exclusively)—learn meaningful representations of audio signals
- Allows *automatic* extraction of high-quality features from raw audio signals
- Contrasted with handcrafted feature extraction, e.g., pitch, jitter, loudness
- 60-second input is segmented before feature extraction:
  - VGGish: 1s chunks → 128-D embeddings
  - OPERA: 4s chunks → 768-D embeddings

Mood Classification
- Transfer learned representations to predict agitation
- Classify input as *low* agitation if $y \in [0, 10]$ or *moderate-to-high* if $y \in [11, 30]$
- Train three lightweight classifiers: Dense Neural Network, Random Forest and XGBoost

Results

Perceived Agitation classification
- VGGish and OPERA both predicted agitation with F1-score > 0.75
- VGGish + XGBoost performed best (F1-score = 0.83)
- Pretrained models outperformed a handcrafted feature-set extractor: *OpenSmile*
- Voice performance comparable to breathing audio (F1-score = 0.8)



Conclusion

- Respiratory and voice audio can be leveraged to infer real-time mood states, specifically agitation
- Transferring learned representations in pretrained audio models to predict mood states is an effective approach
- Opens up opportunity for real-time mood-aware personalization in breathing interventions, ultimately improving adherence

Limitations
- Study was conducted in a controlled lab setting
- Small sample size

Future work
- Collect "in-the-wild" data to validate generalizability
- Shift from binary classification to a regression model
- Investigate the effect of real-time mood-aware personalization in interventions

[2] Simonyan K, and Andrew Z. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv, April 10, 2015.
[3] Zhang, Yuwei, Tong Xia, Jing Han, Yu Wu, Georgios Rizos, Yang Liu, Mohammed Mosuily, J. Ch, and Cecilia Mascolo. *Towards open respiratory acoustic foundation models: Pretraining and benchmarking.* Advances in Neural Information Processing Systems (2024)

ETH zürich (SEC) SINGAPORE-ETH CENTRE (FHT) FUTURE HEALTH TECHNOLOGIES CREATE Campus for Research Excellence And Technological Enterprise Centre for Digital Health Interventions