VibraHealth: Pervasive Health Sensing via Speech-Evoked Multimodal Biosignals

1st Yuanhao Feng
The University of Electro-Communications
Tokyo, Japan
fengyuanhao@uec.ac.jp

2nd Jinyang Huang
Hefei University of Technology
Anhui, China
hjy@hfut.edu.cn

3rd Zhi Liu
The University of Electro-Communications
Tokyo, Japan
liu@ieee.org

***Abstract*—Realizing the vision of Society 5.0 requires seamless integration of physical space and cyberspace to support human-centric well-being. A critical component of this vision is pervasive health monitoring that is both intelligent and privacy-preserving. However, existing solutions often rely on specialized sensing instrumentation or cloud-based audio analysis, which can compromise user comfort and data privacy. We propose *VibraHealth*, an edge-native sensing system that leverages natural speech as an active probe for physiological assessment. Instead of analyzing speech audio, VibraHealth infers health states from speech-evoked biosignals, including Electromyography (EMG), Electrocardiography (ECG), Electroencephalography (EEG), and Electrooculography (EOG). To handle heterogeneous modalities, condition-dependent salience, and strong cross-subject variability, VibraHealth uses modality-specific temporal encoders with modality-aware attention fusion and domain-adversarial training. In a controlled study with 30 participants across six minor health conditions, VibraHealth achieves 87.2% macro-F1 under subject-disjoint evaluation. These results highlight the potential of on-device multimodal fusion for reliable and unobtrusive health intelligence in future smart societies.**

***Index Terms*—Society 5.0, Multimodal Fusion, Edge Intelligence, Cyber-Physical Systems, Wearable Sensing.**

I. Introduction

The deep integration of data, operational, information, and communication technologies is reshaping Industry 4.0 and Society 5.0. Society 5.0 envisions a "super-smart" society in which physical space and cyberspace are tightly coupled to address social challenges and improve quality of life. A key enabler is *human-centric intelligence*: wearable devices and edge computing nodes collaboratively sense human states and deliver timely feedback [1]–[4]. Achieving this vision requires health sensing systems that are accurate, unobtrusive, and privacy-preserving, so that they can be deployed safely in everyday environments [5]–[7].

Despite rapid progress in wearable sensing, continuously detecting *minor yet frequent* health anomalies (e.g., early respiratory infections, fatigue, and mild cardiac discomfort) remains challenging. Clinical measurements can provide high-fidelity signals but are episodic and inconvenient for daily monitoring. In contrast, consumer wearables often rely on limited modalities (e.g., heart rate alone) and may miss systemic physiological changes [3], [8]. Meanwhile, acoustic health sensing approaches commonly process voice recordings and, in some cases, upload audio to the cloud, which raises privacy concerns about speech content and user identity [9], [10]. Therefore, Society 5.0 calls for edge-native architectures that can fuse heterogeneous biosignals into actionable health insights while minimizing exposure of sensitive data [4], [11], [12].

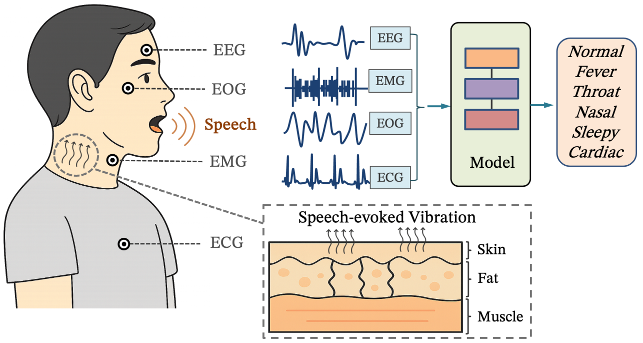


Fig. 1: System overview of *VibraHealth*. During natural speech, wearable sensors capture speech-evoked biosignals, and a local inference pipeline detects minor health conditions without analyzing acoustic content.

In this paper, we propose *VibraHealth*, a privacy-preserving sensing system that infers health conditions from *speech-evoked multimodal biosignals*. As shown in Fig. 1, the system leverages vocalization as an everyday, repeatable physiological probe that elicits coordinated responses across multiple subsystems. VibraHealth captures these responses through EMG, ECG, EEG, and EOG, and performs inference without accessing the semantic or acoustic content of speech.

Designing VibraHealth raises three technical challenges. **First**, biosignals are heterogeneous in sampling characteristics, frequency content, amplitude scales, and noise profiles; effective fusion requires modality-specific feature extraction that preserves informative temporal patterns. **Second**, the diagnostic relevance of each modality is condition-dependent (e.g., sore throat may manifest in neck EMG, while fatigue may be reflected in EEG rhythms), making static fusion strategies unreliable. **Third**, inter-subject variability in physiology and sensor placement can cause models trained on a fixed cohort to overfit subject-specific patterns, limiting generalization to unseen users.

The work of Zhi Liu was partially supported by JST ASPIRE JPMJAP2432.

To address **Challenge 1** (heterogeneous signal characteristics), VibraHealth uses event-centered windowing around detected speech onsets and applies lightweight *modality-specific* temporal encoders (Conv1D+BiGRU) to extract morphology- and rhythm-aware representations under different frequency bands and noise profiles. To address **Challenge 2** (condition-dependent modality relevance), we introduce a *Modality-Aware Attention Fusion* module that learns instance-wise modality weights and aggregates embeddings into a unified health representation, preventing informative cues from being diluted by irrelevant or noisy streams. To address **Challenge 3** (inter-subject variability), we incorporate GRL-based *Adversarial Domain Adaptation*—together with supervised contrastive regularization during training—to reduce subject-specific shortcuts and encourage subject-invariant yet health-discriminative features that generalize to unseen users without per-user retraining.

Our contributions are summarized as follows:

- We present *VibraHealth*, a privacy-preserving sensing paradigm that leverages speech-evoked multimodal biosignals for health monitoring, bridging everyday speech with local cyber-physical inference.
- We design a multimodal learning architecture that integrates heterogeneous biosignals via modality-aware attention and improves cross-subject robustness through domain-adversarial training.
- We evaluate VibraHealth on a dataset of 30 participants covering six health conditions and achieve 87.2% macro-F1, demonstrating feasibility for unobtrusive health intelligence in future smart societies.

II. Physiological Basis of Speech-Evoked Biosignals

VibraHealth builds on the observation that speech is a coordinated physiological maneuver rather than a purely acoustic output. Vocalization couples respiratory control, laryngeal vibration, articulatory muscle activation, and cognitive planning, and therefore mild perturbations (e.g., inflammation, autonomic stress, or fatigue) can modulate biosignals elicited during speech [5], [7]. We thus view vocalization as an *active physiological probe*: compared to passive resting measurements, the task-evoked dynamics around speech onset make subtle deviations more observable.

Speech-evoked responses manifest across complementary subsystems. On the neuromuscular side, phonation and articulation recruit respiratory muscles, intrinsic laryngeal muscles, and the orofacial complex, producing repeatable activation patterns that can shift under minor illness [5], [13]. Throat or respiratory inflammation can increase perilaryngeal tension and prolong activation, yielding higher surface-EMG RMS and longer bursts during speech. Nasal obstruction changes upper-airway impedance and resonance control (e.g., during /m/ and /n/ production), which may trigger compensatory facial muscle recruitment and altered spatial EMG patterns. On the systemic side, controlled exhalation during speech modulates beat-to-beat intervals in ECG; sympathetic activation associated with fever or discomfort may elevate heart rate and reduce HRV, providing a global cue that complements localized EMG changes.

Speech also depends on cognitive control for lexical retrieval and motor sequencing, reflected in frontal EEG rhythms and oculomotor behavior captured by EOG [6], [7]. Fatigue can increase frontal theta-to-beta power ratios during speech [7] and destabilize oculomotor control (e.g., more blinks or altered saccade dynamics) [6]. However, single modalities are often ambiguous: elevated EMG may arise from throat discomfort or general fatigue, and elevated heart rate may reflect fever or psychological stress. By fusing neuromuscular (EMG), systemic (ECG), and neuro-cognitive (EEG/EOG) responses measured for the same speech event, VibraHealth forms a more discriminative and robust health representation for downstream classification [11].

III. System Methodology

VibraHealth is an end-to-end cyber-physical pipeline that transforms raw, heterogeneous biosignals into actionable health-state predictions. As shown in Fig. 2, the pipeline comprises four tightly coupled modules: (A) synchronized acquisition and preprocessing, (B) modality-specific temporal encoding, (C) modality-aware attention fusion, and (D) multi-objective robust learning. Formally, given a set of multimodal time-series signals $X = \{X_m\}_{m\in\mathcal{M}}$ (with $\mathcal{M} = \{\mathrm{EMG}, \mathrm{ECG}, \mathrm{EEG}, \mathrm{EOG}\}$), we learn a function $f_\theta : \mathcal{X} \to \mathcal{Y}$ to predict a health label $y \in \mathcal{Y} = \{1, \ldots, K\}$, while encouraging subject-invariant representations to support cross-user deployment.

A. Synchronized Signal Acquisition and Preprocessing

A key design goal is to align physiological responses to the *speech event* rather than relying on arbitrary sliding windows. This event-centric alignment makes downstream modeling focus on task-evoked dynamics that are most informative for health inference.

1) Vocalization-Triggered Windowing: VibraHealth follows a trigger–response paradigm using voice activity detection (VAD). In our prototype, a bone-conduction microphone provides a robust trigger signal that is less sensitive to ambient noise than airborne microphones. Let t_{start} denote the detected onset of a vocalization (e.g., a sustained phoneme). We extract a symmetric window

$$W = [t_{\mathrm{start}} - \delta,\ t_{\mathrm{start}} + \delta], \qquad \delta = 1.0\ \mathrm{s}. \tag{1}$$

This 2-second window covers three phases commonly present in speech-evoked physiology: (i) *pre-phonation preparation* (motor planning and inhalation), (ii) *phonation execution* (laryngeal vibration and articulatory muscle activation), and (iii) *early recovery* (immediate autonomic adjustment). Centering on the trigger provides a consistent temporal reference across samples and users.

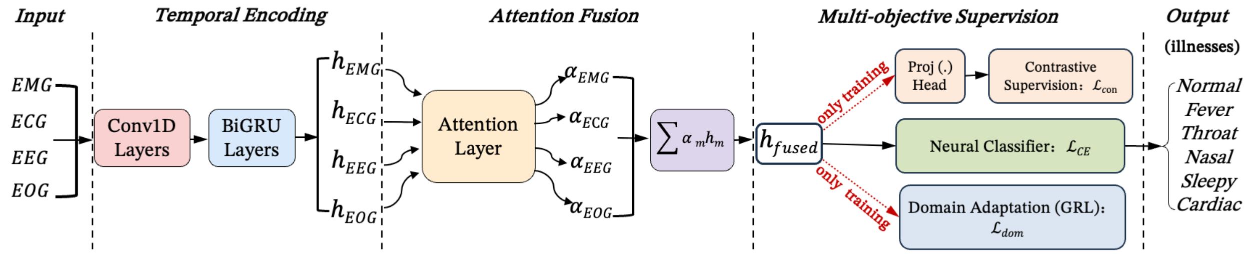


Fig. 2: **System architecture of VibraHealth.** Speech-evoked biosignals (EMG/ECG/EEG/EOG) are windowed and encoded per modality, fused via modality-aware attention, and trained with multi-objective supervision (classification, supervised contrastive learning, and GRL-based domain adaptation). Only the classification branch is used for inference.

2) Signal Conditioning and Normalization: Raw biosignals are subject to baseline drift, power-line interference, motion artifacts, and cross-modality contamination. We therefore apply modality-specific filtering based on typical frequency content:

- **EMG (20–450 Hz):** bandpass filtering to preserve motor-unit activity while suppressing motion artifacts and out-of-band noise.
- **ECG (0.5–40 Hz):** bandpass filtering to retain morphological components used for heart-rate and HRV estimation while attenuating drift and high-frequency interference.
- **EEG (1–45 Hz):** bandpass filtering to cover delta–beta rhythms while reducing high-frequency muscle noise.
- **EOG (0.1–10 Hz):** low-frequency preservation to capture eye movements and blinks while attenuating unrelated high-frequency noise.

After filtering, we apply per-channel z-score normalization within each window, $x' = (x - \mu)/\sigma$, to reduce amplitude variability caused by skin impedance and sensor placement. Finally, to enable uniform tensorization and shared downstream hyperparameters, all modalities are resampled to a common rate $f_s = 250\,\mathrm{Hz}$.

3) Tensor Construction: We organize each window into a set of modality tensors

$$X = \{X_m\}_{m\in\mathcal{M}}, \qquad X_m \in \mathbb{R}^{C_m \times T}, \tag{2}$$

where C_m is the channel count for modality m (e.g., multiple facial/neck EMG channels and a single-lead ECG), and $T = 2\delta f_s = 500$ is the number of samples in the window. This structured representation preserves within-modality spatial structure (channels) and temporal dynamics (samples), and serves as the input to the neural encoder.

B. Modality-Specific Temporal Encoding

Biosignals differ substantially in temporal statistics and informative patterns: EMG contains high-frequency stochastic bursts, ECG includes sparse morphological events (e.g., QRS complexes), EEG is dominated by oscillatory rhythms, and EOG contains slow movements with blink transients. Using a shared encoder can lead to feature interference and suboptimal representations. VibraHealth therefore adopts a split-stream design with an independent temporal encoder per modality.

For modality m, we first apply a 1D temporal convolution to extract local morphological patterns:

$$Z_m = \mathrm{ReLU}(\mathrm{Conv1D}(X_m; W_m) + b_m)\,, \tag{3}$$

where we use 64 filters with kernel size 5. This stage captures short-range patterns such as ECG complexes or blink-related transients. To model longer-range evolution across preparation, execution, and recovery, we feed Z_m to a bidirectional GRU:

$$\vec{h}_t = \mathrm{GRU}(z_t, \vec{h}_{t-1}), \qquad \overleftarrow{h}_t = \mathrm{GRU}(z_t, \overleftarrow{h}_{t+1}). \tag{4}$$

We then form a fixed-dimensional modality embedding

$$h_m = [\vec{h}_T; \overleftarrow{h}_1] \in \mathbb{R}^d, \qquad d = 128, \tag{5}$$

which summarizes modality-specific dynamics around the speech trigger.

C. Modality-Aware Attention Fusion

In realistic settings, the diagnostic utility of modalities is condition-dependent. For example, EMG can be highly informative for throat-related conditions, while EEG/EOG may be more informative for fatigue. Uniform fusion (e.g., plain concatenation) can dilute salient cues with irrelevant or noisy modalities. To address this, VibraHealth uses a modality-aware attention mechanism that assigns an importance weight to each modality embedding.

Given $\{h_m\}_{m\in\mathcal{M}}$, we compute an attention logit for each modality:

$$e_m = v_a^\top \tanh(W_a h_m + b_a), \tag{6}$$

where W_a, b_a, and v_a are learnable parameters shared across modalities. We normalize logits via softmax to obtain weights:

$$\alpha_m = \frac{\exp(e_m)}{\sum_{k\in\mathcal{M}} \exp(e_k)}, \qquad \sum_{m\in\mathcal{M}} \alpha_m = 1. \tag{7}$$

The fused representation is computed as

$$h_{\text{fused}} = \sum_{m\in\mathcal{M}} \alpha_m \, h_m. \tag{8}$$

Beyond performance gains, the attention weights provide a lightweight interpretability signal by exposing which physiological subsystems the model emphasized for a given decision.

D. Multi-Objective Robust Learning

To support reliable sensing across diverse users and noisy daily environments, VibraHealth is trained with a multi-objective loss that jointly promotes (i) classification accuracy, (ii) class-wise compactness/separation, and (iii) subject invariance.

1) Health-State Classification Loss: A classification head $f_{\text{cls}}(\cdot)$ predicts the health distribution $\hat{y}$ from h_{fused}, $\hat{y} = f_{\text{cls}}(h_{\text{fused}})$. We minimize categorical cross-entropy:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{I}(y_i = k)\log\hat{y}_{i,k}, \tag{9}$$

where N is the batch size and $\mathbb{I}(\cdot)$ is the indicator function.

2) Supervised Contrastive Loss: To increase separability between subtle conditions, we adopt supervised contrastive learning. A projection head $g(\cdot)$ maps h_{fused} to $z = g(h_{\text{fused}})$, and we optimize an NT-Xent style objective:

$$\mathcal{L}_{\text{con}} = \sum_{i\in I}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(\text{sim}(z_i, z_p)/\tau)}{\sum_{a\in A(i)}\exp(\text{sim}(z_i, z_a)/\tau)}, \tag{10}$$

where $P(i)$ denotes same-class positives for anchor i, $A(i)$ denotes all comparison samples in the batch, τ is a temperature, and $\text{sim}(\cdot,\cdot)$ is cosine similarity.

3) Adversarial Domain Adaptation Loss: Biosignals exhibit strong inter-subject variability due to physiology and sensor placement. To reduce subject-specific shortcuts, we incorporate adversarial domain adaptation with a gradient reversal layer (GRL). A subject discriminator $D(\cdot)$ predicts subject identity $d_i \in \{1,\ldots,N_{\text{subj}}\}$ from $\text{GRL}(h_{\text{fused}})$:

$$\mathcal{L}_{\text{dom}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{s=1}^{N_{\text{subj}}}\mathbb{I}(d_i = s)\log D(\text{GRL}(h_{\text{fused},i}))_s. \tag{11}$$

Minimizing $\mathcal{L}_{\text{dom}}$ trains D to identify subjects, while the GRL reverses gradients to encourage the encoder to learn subject-invariant embeddings that remain discriminative for health classification.

4) Overall Objective: We train the entire model end-to-end by minimizing

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{dom}}\mathcal{L}_{\text{dom}}. \tag{12}$$

In our experiments, we set $\lambda_{\text{con}} = 0.3$ and $\lambda_{\text{dom}} = 0.2$. This objective encourages representations that are accurate for health inference, well-structured in the latent space, and robust to cross-subject domain shift.

IV. Evaluation

A. Experimental Setup

We evaluate VibraHealth in a controlled user study with 30 university participants under IRB approval.

Data collection and protocol. Fig. 3 shows the experimental setup. Each participant completed a 2-minute read-aloud task using a phonetically balanced script. This protocol reliably engages respiratory control, vocal articulation, and cognitive processing while avoiding additional task-induced workload. We recorded four biosignals synchronously at 250 Hz using research-grade wearable modules: EMG (neck), ECG (chest), EEG (frontal), and EOG (periocular). A bone-conduction microphone provided voice activity detection (VAD) timestamps. Following our methodology (Sec. III), we extract event-centered windows (2 s) around detected vocal onsets. Here, a “vocal onset” corresponds to the start of a voiced segment identified by VAD during read-aloud. Unless otherwise stated, *each window is treated as one sample* for model training and evaluation.

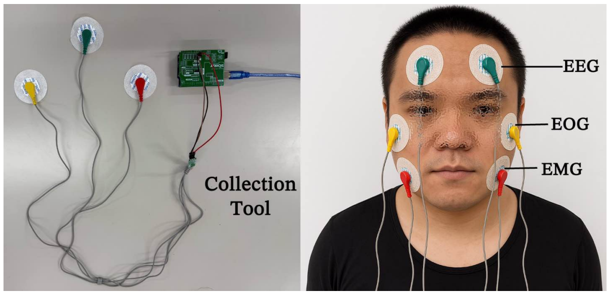


Fig. 3: Data collection setup for synchronized multimodal biosignal acquisition during a read-aloud task.

Health labels. Participants were assigned to six health states using screening criteria based on self-reports, infrared thermometry, and standardized questionnaires (e.g., Karolinska Sleepiness Scale, KSS). Table I summarizes the label definitions. The participant-level distribution is: *Normal* (n=7), *Sleepy* (n=6), *ThroatResp* (n=5), *NasalCong.* (n=5), *Fever* (n=4), and *Cardiac* (n=3). We treat this dataset as a feasibility study covering common minor conditions.

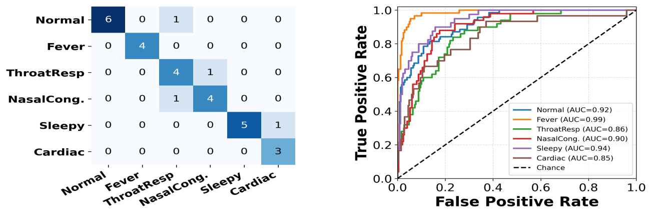
TABLE I: Definition criteria for health condition labels.

Label	Validation Criteria
Normal	No reported symptoms; baseline physiology.
Fever	Body temperature ≥ 37.5°C (infrared thermometry).
ThroatResp	Sore throat (pain score ≥ 3) and/or dyspnea/coughing symptoms.
NasalCong.	Nasal obstruction score ≥ 3 with rhinitis symptoms.
Sleepy	KSS ≥ 5 or prior-night sleep < 6 hours.
Cardiac	Palpitations/tightness (score ≥ 3) with atypical HR/HRV patterns during speech (screening cue).

Implementation details. We implement VibraHealth in PyTorch and train on a single NVIDIA RTX 3090 GPU. Each modality uses a Conv1D layer (kernel size 5, 64 filters) followed by a BiGRU (hidden size 64). The fused embedding dimension is $d = 128$. We optimize with Adam (learning rate 10^{-3}, batch size 64). For supervised contrastive learning, we set $\tau = 0.07$. We set $\lambda_{\text{con}} = 0.3$ and $\lambda_{\text{dom}} = 0.2$ based on a lightweight grid search on the training folds.

B. Performance Analysis

We report macro-F1, accuracy, and one-vs-rest ROC-AUC. To avoid leakage across multiple windows from the same participant, we perform **5-fold cross-validation with subject-disjoint splits**. Due to the small cohort and imbalanced labels (e.g., only 3 *Cardiac* participants), folds are constructed to balance classes *as much as possible* rather than enforcing strict stratification. Unless otherwise specified, performance metrics are computed at the **window level** by pooling predictions from all test folds.



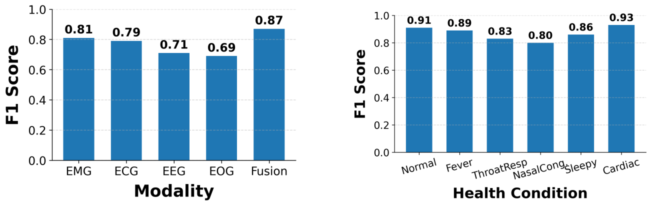
(a) Confusion Matrix

(b) ROC Curve (One-vs-Rest)

Fig. 4: **Overall classification performance.** (a) Subject-level confusion matrix over six health states (N=30), where each participant contributes a single prediction after aggregating window-level outputs. (b) One-vs-rest ROC curves computed from window-level predictions under subject-disjoint cross-validation.

Overall detection performance (window-level). VibraHealth achieves a macro-F1 of **0.87** and an accuracy of **0.88** on pooled test windows. Fig. 4(b) reports one-vs-rest ROC curves computed at the *window level*. Across the six classes, VibraHealth achieves a mean AUC of **0.91**, indicating robust sensitivity for early health screening.

Subject-level confusion analysis. Fig. 4(a) reports a *subject-level* confusion matrix. For each held-out participant, we aggregate predictions across all test windows by averaging predicted class probabilities and assign a single label using $\arg\max$. The results show that *Normal*, *Fever*, and *Cardiac* are identified with high precision and recall, suggesting that speech-evoked biosignals capture discriminative systemic cues. The primary confusion occurs between *ThroatResp* and *NasalCong.*, which is expected because both conditions affect the upper airway and can induce similar compensatory articulation patterns.



(a) Modality-wise Macro-F1

(b) Per-Class F1

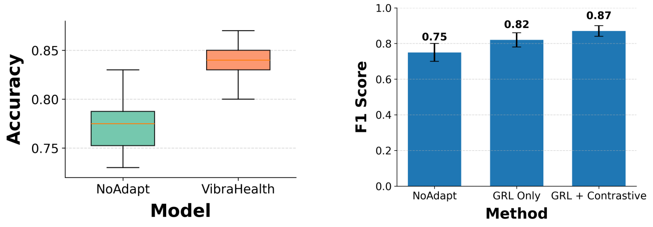
Fig. 5: **Detailed performance breakdown.** (a) Multimodal fusion improves macro-F1 over all unimodal baselines. (b) Per-class F1 scores remain consistently high across conditions.

Benefit of multimodal attention fusion. Fig. 5(a) compares VibraHealth with unimodal baselines using the same encoder backbone. EMG achieves the strongest unimodal performance (macro-F1 $=$ 0.81), consistent with its coupling to vocal-tract muscle activity, while ECG follows (0.79) due to its sensitivity to systemic stress (e.g., fever). EEG (0.71) and EOG (0.69) are weaker in isolation, but provide critical cues for cognitive states such as *Sleepy*. By fusing modalities, VibraHealth reaches macro-F1 $= 0.87$, improving by **0.06** over the best unimodal baseline, indicating that modality-aware attention effectively leverages complementary cues.

Fig. 5(b) further breaks down performance by class. The lowest F1 values are observed for the upper-respiratory conditions (*ThroatResp* and *NasalCong.*), matching the confusion patterns in Fig. 4(a), whereas systemic conditions such as *Normal*, *Fever*, and *Cardiac* remain highly separable. These results suggest that fusion improves not only average accuracy but also class-wise robustness under heterogeneous symptom profiles.

C. Generalization and Robustness

A practical Society 5.0 deployment requires robustness to unseen users without per-user calibration. We evaluate cross-subject generalization using leave-one-subject-out (LOSO) validation, where all windows from the held-out participant are used only for testing.



(a) Subject Generalization (LOSO)

(b) Ablation of Robust Objectives

Fig. 6: **Robustness analysis.** (a) LOSO accuracy distribution across held-out participants. (b) Ablation of robust learning objectives; error bars denote variability across LOSO runs.

Cross-subject generalization. Fig. 6(a) shows LOSO accuracy across test subjects. Compared to a non-adapted baseline (*NoAdapt*), VibraHealth reduces performance variance and improves average accuracy from **0.75** to **0.86**. This suggests that the learned representation relies less on subject-specific signal idiosyncrasies.

Impact of robust learning objectives. Fig. 6(b) ablates the multi-objective components. Adding GRL-based domain adaptation improves macro-F1 from 0.75 to 0.82, demonstrating that subject shift is a dominant challenge for biosignals. Adding supervised contrastive learning further improves macro-F1 to 0.87 by refining class separation in the embedding space. Together, these objectives yield a robust model suitable for cross-subject deployment.

V. Related Work

Biosignal-based physiological sensing. Wearable and mobile computing systems increasingly rely on bioelectrical signals to infer human states and behaviors [5]–[7], [13]. Prior work has demonstrated that EMG can capture fine-grained neuromuscular dynamics for activity and gesture recognition (e.g., EMGSense and sEMG-based pipelines) [5], [13], while EOG has been repurposed for unobtrusive affect and attention sensing [6], [14]. EEG-driven systems further enable cognitive state estimation and neuro-behavioral inference through rhythm features and signal synchrony [15]–[17]. In contrast to these efforts, which typically assume explicit tasks (e.g., gestures) or target affect/cognition as primary outcomes, VibraHealth focuses on *minor, systemic health conditions* and uses *speech as an active physiological probe* without analyzing acoustic content.

Multimodal fusion and robustness. Multimodal fusion has become a standard strategy to improve reliability and coverage in mobile health sensing, where a single modality is often insufficient under noise, missing data, or heterogeneous symptom manifestations [1], [11], [12], [18]–[21]. Representative systems combine multiple biosignals (e.g., EMG/PPG/GSR) to estimate workload, habits, and health-related behaviors [1], [2]. Recent models further explore attention-based fusion and domain generalization to mitigate inter-subject variability and sensor placement effects [3], [4]. VibraHealth builds on these lines of work but differs in both *problem setting* and *trigger*: it aligns and fuses *speech-evoked* biosignals for privacy-preserving illness screening, and combines modality-aware attention, supervised contrastive learning, and adversarial adaptation to support cross-subject deployment.

VI. Conclusion

We presented *VibraHealth*, a privacy-preserving health sensing system that leverages speech as an active physiological probe. Instead of analyzing voice content, VibraHealth fuses speech-evoked EMG/ECG/EEG/EOG via modality-specific encoding, attention-based fusion, supervised contrastive learning, and adversarial domain adaptation. In a 30-participant study across six conditions, VibraHealth achieves 87.2% macro-F1 and strong cross-subject generalization. Future work will optimize ultra-low-power edge deployment and enable longitudinal monitoring with next-generation networks.

References

[1] E. Park, D. Lee, Y. Han, J. Diefendorff, and U. Lee, "Hide-and-seek: Detecting workers' emotional workload in emotional labor contexts using multimodal sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 3, pp. 1–28, 2024.

[2] G. J. Fernandes, J. Zheng, M. Pedram, C. Romano, F. Shahabi, B. Rothrock, T. Cohen, H. Zhu, T. S. Butani, J. Hester *et al.*, "Habitsense: A privacy-aware, ai-enhanced multimodal wearable platform for mhealth applications," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 3, pp. 1–48, 2024.

[3] B. Zhai, Y. Guan, M. Catt, and T. Plötz, "Ubi-sleepnet: Advanced multimodal fusion techniques for three-stage sleep classification using ubiquitous sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–33, 2021.

[4] L. Meegahapola, H. Hassoune, and D. Gatica-Perez, "M3bat: Unsupervised domain adaptation for multimodal mobile sensing with multi-branch adversarial training," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 2, pp. 1–30, 2024.

[5] D. Duan, H. Yang, G. Lan, T. Li, X. Jia, and W. Xu, "Emgsense: A low-effort self-supervised domain adaptation framework for emg sensing," in *2023 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2023, pp. 160–170.

[6] S. Rostaminia, A. Lamson, S. Maji, T. Rahman, and D. Ganesan, "W! nce: Unobtrusive sensing of upper facial action units with eog-based eyewear," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 1, pp. 1–26, 2019.

[7] Z. Rui, Y. Li, Z. Dong, L. Hao, B. Chen, F. Chang, and Z. Gu, "Eeg, eog, likert scale, and interview approaches for assessing stressful hazard perception scenarios," *International Journal of Human–Computer Interaction*, pp. 1–26, 2024.

[8] K. Saha, T. Grover, S. M. Mattingly, V. D. Swain, P. Gupta, G. J. Martinez, P. Robles-Granda, G. Mark, A. Striegel, and M. De Choudhury, "Person-centered predictions of psychological constructs with social media contextualized by multimodal sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–32, 2021.

[9] X. Xu, E. Nemati, K. Vatanparvar, V. Nathan, T. Ahmed, M. M. Rahman, D. McCaffrey, J. Kuang, and J. A. Gao, "Listen2cough: Leveraging end-to-end deep learning cough detection model to enhance lung health assessment using passively sensed audio," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–22, 2021.

[10] M. Clark and A. Doryab, "Sounds of health: Using personalized sonification models to communicate health information," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 4, pp. 1–31, 2023.

[11] K. Sun, C. Xia, X. Zhang, H. Chen, and C. J. Zhang, "Multimodal daily-life logging in free-living environment using non-visual egocentric sensors on a smartphone," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, pp. 1–32, 2024.

[12] D. Jeong and K. Han, "Precyse: Predicting cybersickness using transformer for multimodal time-series sensor data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 2, pp. 1–24, 2024.

[13] Y. Zhang, Y. Chen, H. Yu, X. Yang, R. Sun, and B. Zeng, "A feature adaptive learning method for high-density semg-based gesture recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–26, 2021.

[14] C. Belkhiria, A. Boudir, C. Hurter, and V. Peysakhovich, "Eog-based human–computer interface: 2000–2020 review," *Sensors*, vol. 22, no. 13, p. 4914, 2022.

[15] B. Zhang, X. Li, Y. Zhou, J. Liu, C. Zhou, W. Liu, and Y. Bian, "Are we in the zone? exploring the features and method of detecting simultaneous flow experiences based on eeg signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 4, pp. 1–42, 2024.

[16] S. Bian, P. Kang, J. Moosmann, M. Liu, P. Bonazzi, R. Rosipal, and M. Magno, "On-device learning of eegnet-based network for wearable motor imagery brain-computer interface," in *Proceedings of the 2024 ACM International Symposium on Wearable Computers*, 2024, pp. 9–16.

[17] J. Berent, E. Vereycken, N. Colenbier, P. van Mierlo, and C. Neuray, "Quality assessment and neurophysiological signal resemblance of ear eeg," in *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2024, pp. 691–696.

[18] Y. Feng, D. Dai, J. Huang, P. Yang, X.-Y. Li, F. Han, and L. Yang, "Battery-free monitoring of micron-level vibrations with sub-hertz frequency accuracy: Toward robust and accurate industrial sensing," *IEEE Transactions on Mobile Computing*, 2025.

[19] Y. Feng, Y. Zhang, P. Yang, H. Zhou, H. Du, and X.-Y. Li, "Rf-ear+: A mechanical identification and troubleshooting system based on contactless vibration sensing," *IEEE Transactions on Mobile Computing*, vol. 22, no. 12, pp. 7310–7326, 2022.

[20] J. Huang, J.-X. Bai, X. Zhang, Z. Liu, Y. Feng, J. Liu, X. Sun, M. Dong, and M. Li, "Keystrokesniffer: An off-the-shelf smartphone can eavesdrop on your privacy from anywhere," *IEEE Transactions on Information Forensics and Security*, 2024.

[21] J. Huang, B. Liu, C. Miao, X. Zhang, J. Liu, L. Su, Z. Liu, and Y. Gu, "Phyfinatt: An undetectable attack framework against phy layer fingerprint-based wifi authentication," *IEEE Transactions on Mobile Computing*, 2023.