Are You Performing, Anticipating, or Observing? Role-Aware Physiological Sensing in Group Interactions

Sharmad Kalpande*
Indian Institute of Science Education and Research Bhopal
Bhopal, India
kalpande22@iiserb.ac.in

Nilesh Kumar Sahu*
Indian Institute of Science Education and Research Bhopal
Bhopal, India
nilesh21@iiserb.ac.in

Haroon R. Lone
Indian Institute of Science Education and Research Bhopal
Bhopal, India
haroon@iiserb.ac.in

Abstract

Structured group interactions such as presentations and sequential speaking tasks involve rapidly changing participant roles including performing, anticipating, and observing. Although participants share the same environment, these roles may induce distinct physiological responses, yet existing physiological sensing approaches largely treat co-present participants as experiencing a shared interaction context. We present a role-aware framework for modeling physiological responses during structured group interactions using wearable sensing. In a controlled study with 100 participants organized into groups of three, ECG and GSR signals were recorded during sequential speaking activities and labeled according to participant role: *Performer*, *Anticipator*, *Observer*, or *Baseline*. Using statistical analysis, machine learning, and deep learning approaches under a strict Leave-One-Group-Out protocol, we show that role-specific physiological responses can be differentiated. Deep learning models outperformed conventional approaches, with the best-performing model achieving 0.66 accuracy and 0.87 AUROC.

CCS Concepts

• **Human-centered computing → Empirical studies in ubiquitous and mobile computing**; *Ubiquitous computing*.

Keywords

Wearable Sensing, Physiological Signals, Group Interactions.

ACM Reference Format:

Sharmad Kalpande, Nilesh Kumar Sahu, and Haroon R. Lone. 2026. Are You Performing, Anticipating, or Observing? Role-Aware Physiological Sensing in Group Interactions. In *Proceedings of the 2026 ACM International Symposium on Wearable Computers (ISWC '26), October 11–15, 2026, Shanghai, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3830727.3834816

1 Introduction

Structured group interactions such as presentations, sequential speaking tasks, and evaluative discussions produce rapidly changing physiological demands across participants [1, 2]. In these settings, participants dynamically transition between three functional roles: *Performer*, the participant currently speaking; *Anticipator*, the participant awaiting their turn; and *Observer*, the participant who has already completed their turn and is observing others. Although participants share the same physical environment, these roles impose distinct cognitive and emotional demands that can lead to substantially different physiological states. Recent advances in wearable sensing have enabled continuous monitoring of such responses in both controlled and naturalistic environments using modalities such as electrocardiography (ECG) and galvanic skin response (GSR) [3, 4]. However, existing physiological sensing approaches largely analyze individuals in isolation or assume that all co-present participants experience a shared interaction context [5–8]. As a result, current methods lack **role-aware physiological modeling** in group interactions.

This limitation directly affects how physiological signals are labeled and modeled. Most prior work assigns a single label to an entire session or participant without accounting for role transitions during the interaction [5, 9]. In a sequential group task, the same participant may transition through anticipation, active performance, and post-performance observation within a session. Treating these temporally distinct phases as a single condition obscures the underlying physiological structure of the interaction. We therefore argue that physiological sensing in group settings requires **role-level segmentation**, where each signal window is labeled according to the participant's functional role at that moment. Such segmentation is important because anticipatory and post-performance phases may induce physiologically distinct responses despite occurring within the same shared environment [1].

A particularly underexplored question concerns the physiological responses associated with the Observer role. Existing work generally assumes that once participants complete their speaking task, their physiological activation returns toward a neutral condition [6, 10]. However, structured group interactions remain socially active even after an individual has finished speaking, as participants continue to observe an active performer while another anticipates evaluation. This shared interactional context may sustain physiological activation beyond the period of direct task engagement. If the physiological responses associated with the Observer role remain distinguishable from baseline, it would suggest that group interactions produce sustained effects extending beyond active participation, with implications for how physiological responses are modeled in social settings.

Beyond identifying role-dependent physiological responses, another challenge concerns how such systems are evaluated. Prior work in physiological sensing commonly evaluates models by holding out individual participants while training on data collected

*Both authors contributed equally.

ACM ISBN 979-8-4007-2872-3/2026/10
https://doi.org/10.1145/3830727.3834816

from the same group interactions [2, 11, 12]. In real-world deployment, however, models must generalize to entirely unseen groups with different interaction dynamics and participant compositions. Evaluating models at the group level, where all participants and their interaction context are unseen during training, therefore represents a more realistic and substantially stricter benchmark that has received limited attention in physiological sensing research.

Accordingly, this work investigates the following research questions: (i) **RQ1:** Do Performer, Anticipator, and Observer roles induce distinguishable physiological responses during structured group interactions? (ii) **RQ2:** Is the physiological response associated with the Observer role distinguishable from baseline despite the absence of active task participation? (iii) **RQ3:** Can wearable sensing models generalize role classification to entirely unseen interaction groups under Leave-One-Group-Out (LOGO) evaluation?

To address these questions, we formulate role identification in structured group interactions as a multiclass physiological sensing problem. We conducted a controlled study involving 117 participants organized into groups of three, where ECG and GSR signals were simultaneously recorded from all participants using wearable sensors. We model four interaction conditions corresponding to the Performer, Anticipator, Observer, and Baseline roles, and investigate whether these roles induce distinguishable physiological responses. Using statistical analysis, classical machine learning, and end-to-end deep learning architectures, we evaluate role classification under a LOGO protocol to assess generalization to unseen interaction groups.

The major contributions of this work are as follows:

- We introduce a role-aware formulation for physiological sensing in structured group interactions and propose role-level segmentation for modeling temporally varying interaction roles.
- We investigate whether Performer, Anticipator, Observer, and Baseline roles induce distinguishable physiological responses using multimodal wearable signals (ECG and GSR).
- We evaluate role classification using statistical analysis, classical machine learning, and deep learning approaches under a strict LOGO evaluation protocol designed to assess generalization to unseen groups.

2 Related Work

Physiological signals particularly ECG and GSR are well-established indicators of autonomic nervous system activity, with features derived from heart rate variability (HRV) and skin conductance reliably capturing variations in sympathetic and parasympathetic tone across cognitive, social, and evaluative demands [3–5, 13]. Multimodal fusion of these signals consistently outperforms unimodal approaches [4, 14–16], and deep learning architectures operating on raw signals have demonstrated strong discriminative performance without handcrafted features [14, 17, 18]. However, these works primarily assign labels based on the activity or stimulus experienced by an individual participant, without accounting for the participant's functional role within an ongoing group interaction.

A parallel line of work has extended physiological sensing to group and multi-participant settings [10, 19, 20], instrumenting multiple individuals simultaneously to study interpersonal physiological synchrony in collaborative, classroom, and audience settings [6, 7, 21]. These studies demonstrate that physiological signals are sensitive to the social dynamics of group interaction. However, the group is typically treated as the unit of analysis, with emphasis placed on synchrony or shared arousal across co-present participants rather than modeling individual physiological responses as a function of interaction role. In contrast, our work focuses on role-dependent physiological responses among participants sharing the same interaction environment.

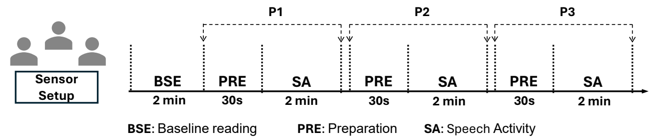

Figure 1: Data collection protocol for each participant, consisting of a 2 minute baseline recording (BSE), 30 second preparation time (PRE), and 2 minute speech activity (SA) repeated sequentially for each participant (P1, P2, P3).

Sequential public speaking tasks are widely used as controlled social evaluative settings in wearable physiological research [1, 2, 22]. Prior studies have shown that ECG and GSR responses differ across anticipation and active speaking phases, demonstrating that anticipatory periods represent physiologically meaningful states rather than neutral waiting conditions [1, 23]. These findings are directly relevant to our formulation because they suggest that temporally distinct phases within the same interaction can induce distinguishable physiological responses. However, existing work in this setting primarily instruments only the active speaker. The physiological responses of co-present participants who are anticipating their turn or observing another participant's performance remain unexplored.

A further limitation concerns how physiological states are defined and labeled in social interaction settings. Prior work examining anticipation, active experience, or post-event reflection has typically induced these phases explicitly through instructions or experimental prompts [23]. In these settings, all participants transition through the same experimental phase simultaneously. In contrast, anticipatory and observer states in sequential group interactions emerge organically from the interaction structure itself, where participants continuously respond to an unfolding social context without external prompting. This distinction is important because naturally occurring anticipation, sustained across another participant's performance, is structurally different from a short prompted reflection or preparation period. Despite this, existing labeling approaches continue to assign a single condition-level label to all participants across an interaction [5], overlooking the role-dependent physiological variations that occur within shared group interactions. Our work addresses this limitation through role-level segmentation, where physiological signals are labeled according to each participant's functional role during the interaction.

3 Methodology

3.1 Data Collection

We conducted a controlled study in which participants engaged in structured sequential speaking (i.e., speech) interactions in groups of three. ECG and GSR signals were recorded using the Shimmer3 wearable device at a sampling rate of 1024 Hz during speech performance and during periods when participants were not actively speaking. The study received institutional ethics approval, and all participants provided informed consent prior to participation. A total of 117 participants were recruited across 39 group interactions. During data analysis, recordings from 17 participants were excluded due to corrupted signals or due to dummy participants when scheduled participants did not attend the experiment. The final dataset consisted of 100 participants aged 17–30 years.

In each group interaction, participants first underwent a 2-minute baseline period during which they were asked to sit idle. Following the baseline, the research assistant (RA) assigned a speech topic to the participant, provided 30 seconds for preparation, and then asked the participant to speak for 2 minutes. After 2 minutes, the RA stopped the participant and assigned a new topic to the next participant (see Figure 1). The RA randomly selected topics from a predefined set of 29 topics, including examples such as "the importance of smart gadgets in your life" and "the merits and demerits of video games". At any given time during the activity, only one participant actively performs the task (speech delivery), while the remaining participants are present in the same environment but are not speaking (see Figure 2). Unlike prior work that treats all non-speaking participants as a single category, we explicitly differentiate their roles based on their position in the speaking order.

To formalize the sequential role structure of the speaking interaction, we define three scenarios (S1, S2, S3), each representing a different roles of participants within the group (see Figure 2). In each scenario, the participant delivering the speech is labeled as the Performer (Perf). Among the remaining participants, the individual who is next in the speaking order is labeled as Anticipator 1 (ATP1), indicating that they are preparing to speak next. The other participant, who will speak later, is labeled as Anticipator 2 (ATP2). Similarly, once a participant has completed their turn, they transition into an Observer. To distinguish between participants based on their position after speaking, we define Observer 1 (OBS1) and Observer 2 (OBS2), corresponding to different positions in the post-performance order. In addition to these interaction-derived roles, the resting signals recorded prior to the activity are assigned a Baseline (Base) label, serving as a neutral physiological reference.

Based on this formulation, role labels are assigned to all participants simultaneously for each scenario. For example, in S_1 when P_1 is performing, P_1 is assigned *Perf*, P_2 is assigned *ATP1*, and P_3 is assigned *ATP2*. As the interaction progresses across scenarios, each participant transitions through a distinct role trajectory P_1 transitions as *Perf* $\rightarrow$ *OBS1* $\rightarrow$ *OBS2*, P_2 as *ATP1* $\rightarrow$ *Perf* $\rightarrow$ *OBS1*, and P_3 as *ATP2* $\rightarrow$ *ATP1* $\rightarrow$ *Perf*, resulting in a structured mapping of roles across all participants and scenarios, ensuring that each participant contributes data under multiple role conditions.

3.2 Feature Extraction and Statistical Analysis

Following role assignment, physiological signals are segmented based on the duration of each speaking instance. For each scenario, the interval during which the performer delivers the speech is identified (SA in Figure 1), and the corresponding signal segments are extracted for all three participants in the group. This ensures

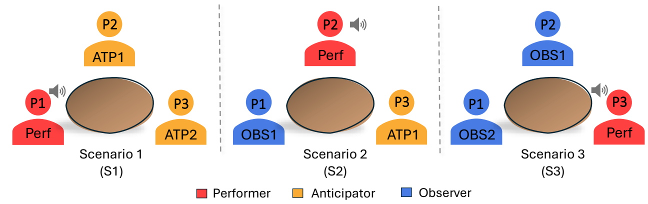

Figure 2: Experimental setup across three scenarios demonstrating systematic role transitions of participants (P1, P2, P3) among Performer (Perf), Anticipator (ATP), and Observer (OBS) roles.

that all group participants are analyzed under the same contextual condition defined by the active speech segment. Baseline segments are similarly extracted from the resting period (BSE in Figure 1).

For ECG, the signal segments was filtered and R-peaks were detected using the BioSPPy [24] library, and HRV-based features were subsequently extracted from the detected R-peaks using the NeuroKit2 library [25]. A total of 105 features[1] were extracted spanning time-domain, frequency-domain, nonlinear, and recurrence quantification analysis (RQA) measures. Time-domain measures capture overall variability and successive beat differences (e.g., SDNN, RMSSD), frequency-domain measures reflect autonomic balance across standard spectral bands, and nonlinear and RQA measures characterize the complexity and dynamical structure of cardiac activity. For GSR, the signal segments were resampled to 16 Hz prior to feature extraction, as GSR is a slow-varying signal whose relevant dynamics are well captured at lower sampling rates. The resampled signal was then cleaned using the GSR preprocessing functions provided in NeuroKit2, followed by phasic decomposition for feature extraction. A total of 46 features[2] were computed from the phasic (SCR) signal, covering peak characteristics (e.g., amplitude, count), temporal properties, and statistical descriptors.

Following feature extraction, a feature-wise statistical analysis was performed to identify features exhibiting significant differences across role conditions. Normality was assessed using the Shapiro-Wilk test [26], with parametric features evaluated using one-way ANOVA and non-parametric features using the Kruskal-Wallis test. Post-hoc pairwise comparisons were conducted using Tukey's HSD following ANOVA and Dunn's test with Bonferroni correction following Kruskal-Wallis, with significance set at $p_{adj} < 0.05$. Based on this analysis, consistent with prior work on physiological feature selection for autonomic state classification [23], a subset of 39 statistically significant features (19 ECG and 20 GSR) mentioned in Table 1 was retained for classification.

3.3 Segmentation

The complete signal segments (spanning all role conditions and Baseline) are further divided into fixed-length *windows* of 60 seconds with a 45-second overlap between consecutive windows. The selected window configuration ensures that sufficient temporal information, particularly from the slower-varying GSR signals, is preserved while also maintaining adequate ECG dynamics for analysis. Moreover, the overlapping windowing strategy increases the

[1] https://neuropsychology.github.io/NeuroKit/functions/hrv.html
[2] https://neuropsychology.github.io/NeuroKit/functions/eda.html

Table 1: List of statistically significant ECG and GSR features used in classification.

Signal	Features
ECG	MeanNN, SDNN, RMSSD, CVNN, CVSD, MedianNN, MadNN, IQRNN, Prc20NN, pNN50, pNN20, TINN, HTI, LF, HF, VHF, TP, LFHF, LnHF
GSR	*SCR:* Duration, Amplitude, RiseTime, RecoveryTime (mean, std each), Peak_count, mean_area *Phasic:* median, dynamic_range, slope, 1st_deriv (mean, std), 2nd_deriv (mean, std), Wavelet_4Hz (mean, median, std)

number of training windows available for modeling. The resulting dataset yields 547 Performer, 510 Anticipator (353 ATP1 and 157 ATP2), 506 Observer (354 OBS1 and 152 OBS2), and 461 Baseline windows, providing an approximately balanced class distribution across all four role conditions and sufficient samples for training and evaluating the classification models. The 39 statistically significant features identified in Section 3.2 were subsequently extracted from these windowed segments for classification.

3.4 Models

3.4.1 Machine Learning (ML) Models. To evaluate the effectiveness of the extracted features, we employ a set of standard ML classifiers [27] commonly used in physiological signal analysis. Specifically, we evaluated Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (XGBoost), and a Multi-Layer Perceptron (MLP) using standard configurations[3].

3.4.2 Deep Learning (DL) Models. While handcrafted features summarize physiological signals using explicit statistical measures, they may overlook subtle temporal dynamics associated with different interaction roles. To address this, we employ DL [28] architectures that operate on raw ECG and GSR signals alongside handcrafted features, enabling end-to-end learning of role-specific physiological representations. We evaluate two single-modality baselines and three multimodal fusion architectures to compare feature-based, raw-signal, and multimodal learning strategies for role classification.

Deep Neural Network (DNN) serves as a handcrafted-feature baseline operating exclusively on extracted ECG and GSR features. The model consists of multiple fully connected layers with Batch Normalization, ReLU activation, and Dropout regularization for stable training and improved generalization.

Dual-Branch ResNet (DBResNet) serves as a raw-signal baseline operating directly on ECG and GSR signals without handcrafted features. Separate 1D ResNet encoders process ECG and GSR independently to capture modality-specific temporal patterns. The learned representations are concatenated and passed to a fully connected classifier for role prediction.

We further evaluate three multimodal architectures, i.e., **M1**, **M2**, and **M3**, representing feature concatenation, attention-based fusion, and multi-scale attention-based fusion strategies, respectively.

[3]https://scikit-learn.org/stable/

***M1:** Multimodal ResNet1D with Feature-Level Concatenation:* M1 processes raw ECG and GSR signals using separate 1D ResNet encoders while handcrafted physiological features are encoded using a fully connected network. The learned embeddings from ECG, GSR, and handcrafted features are concatenated to form a unified multimodal representation, which is then passed through a fully connected classification head for role prediction.

***M2:** Dual ResNet1D with Attention-Based Feature Fusion:* M2 extends M1 by replacing fixed feature concatenation with an attention-based fusion mechanism. ECG, GSR, and handcrafted feature embeddings are first extracted independently and then adaptively weighted using an attention module to emphasize the most informative modalities for role classification. The fused representation is subsequently used for final prediction.

***M3:** Multi-Scale Encoder with Attention-Based Feature Fusion:* M3 follows the same attention-based fusion strategy as M2 but replaces the ResNet signal encoders with multi-scale convolutional encoders designed to capture temporal patterns at different receptive fields. Multi-scale feature representations from ECG and GSR are combined with handcrafted feature embeddings through the attention-based fusion module and used for final classification.

3.5 Evaluation

To evaluate the effectiveness of physiological features in distinguishing participant roles, we formulate the problem as a multi-class classification task. While the original role definitions distinguish between finer categories (e.g., ATP1/ATP2 & OBS1/OBS2), statistical analysis (Section 3.2) revealed no significant physiological differences between ATP1 & ATP2, or between OBS1 & OBS2. Accordingly, these sub-roles are merged into broader classes as ATP1 & ATP2 into a single Anticipator (ATP) class, and OBS1 & OBS2 into a single Observer (OBS) class, resulting in four classes: Baseline (Base), Performer (Perf), Anticipator (ATP), and Observer (OBS).

To rigorously assess model performance, we adopt a LOGO evaluation strategy. In this setting, each group (consisting of three participants) is held out as the test set, while the model is trained on the remaining groups. For DL models, 10% of the training groups (approximately 4 groups) are held out as a validation set for model selection. This ensures that all participants within a group are entirely unseen during training. All DL models are trained using a learning rate of 1×10^{-4}, a batch size of 16, a weight decay of 1×10^{-4}, and for 20 epochs.

Model performance is evaluated using accuracy, precision, recall, F1-score, and AUROC. Since the label distribution across the four role classes is approximately balanced, we report macro averaged precision, recall, and F1-score to treat each role class equally in the evaluation. AUROC is reported under a one-vs-rest scheme for each class, providing a threshold independent measure of how well the model separates each role's physiological signature from the remaining classes directly evaluating the core discriminability claim of our role-based formulation.

3.6 Baseline Normalization

Physiological signals such as ECG and GSR exhibit substantial inter-subject variability, where absolute signal magnitudes and patterns differ across individuals due to inherent physiological differences. To account for this variability, we apply a baseline normalization step to transform the data into a subject-relative representation.

For each participant, the baseline signal are used to compute a participant-specific reference for normalization. For the raw ECG and GSR signals, the mean baseline window is obtained by averaging across all baseline windows of that participant for each modality independently. This mean baseline ECG window and mean baseline GSR window are then subtracted from every corresponding window of that participant across all role conditions (Perf, ATP, OBS). Similarly, for the extracted features (Table 1), the mean baseline feature vector is computed by averaging the feature vectors extracted from all baseline windows of that participant. This mean baseline feature vector is then subtracted from the feature vectors of every window belonging to that participant across all role conditions.

This procedure transforms all modalities (raw ECG, GSR, and extracted features) into a participant-relative space, allowing the model to focus on deviations from each individual's resting physiological state rather than absolute signal values.

4 Results

Table 2 presents the classification performance of ML models under the LOGO evaluation. LR and SVM achieve the highest accuracies among all classifiers, both at 0.44. AUROC values range from 0.66 to 0.72, with LR achieving the highest AUROC, indicating modest separability across role classes. Overall, ML classifiers remain below 0.45 accuracy, suggesting that handcrafted features alone provide limited discriminability when generalizing to unseen groups.

Table 2: Classification performance of ML models using windowed segments.

Model	Accuracy	Precision	Recall	F1	AUROC
LR	**0.44 ± 0.12**	0.43 ± 0.13	**0.43 ± 0.12**	**0.39 ± 0.12**	**0.72 ± 0.09**
SVM	**0.44 ± 0.10**	**0.44 ± 0.12**	**0.43 ± 0.10**	**0.39 ± 0.10**	0.70 ± 0.12
RF	0.42 ± 0.11	0.42 ± 0.13	0.42 ± 0.12	0.38 ± 0.11	0.70 ± 0.11
XGB	0.43 ± 0.10	**0.44 ± 0.12**	**0.43 ± 0.12**	**0.39 ± 0.11**	0.70 ± 0.10
MLP	0.39 ± 0.11	0.41 ± 0.12	0.38 ± 0.11	0.36 ± 0.10	0.66 ± 0.09

DL baselines showed similar trends: DNN, trained on handcrafted features, achieves an accuracy of 0.43, while DBResNet, trained on raw signals, achieves 0.42 (Table 3). Both models reach a comparable AUROC of 0.71. Overall, both remain comparable to the ML baselines, suggesting that neither input type alone is sufficient for robust role classification under the LOGO setting.

Table 3: Classification performance of baseline DL models.

Model	Accuracy	Precision	Recall	F1	AUROC
DNN	**0.43 ± 0.12**	**0.45 ± 0.15**	**0.43 ± 0.12**	**0.39 ± 0.11**	**0.71 ± 0.11**
DBResNet	0.42 ± 0.14	0.43 ± 0.17	0.41 ± 0.13	0.37 ± 0.14	0.71 ± 0.11

To examine whether combining both inputs (handcrafted + raw signal) can better capture role-specific patterns, Table 4 presents the performance of the three multimodal DL architectures under the same LOGO evaluation. M3 achieves a performance of **0.48** accuracy and **0.76** AUROC, followed by M2 (accuracy: 0.47, AUROC: 0.76) and M1 (accuracy: 0.46, AUROC: 0.76). The DL architectures consistently outperform the ML baselines across all metrics, with M3 improving both accuracy and AUROC by approximately 4% over the best ML classifier. Despite this improvement, absolute performance remains modest, suggesting that while learned temporal representations from raw ECG and GSR signals capture role-specific dynamics better than handcrafted features, generalizing to entirely unseen groups remains a challenging problem.

Table 4: Classification performance of DL architectures without baseline normalization.

Model	Accuracy	Precision	Recall	F1	AUROC
M1	0.46 ± 0.13	0.48 ± 0.15	0.46 ± 0.13	0.42 ± 0.13	0.76 ± 0.08
M2	0.47 ± 0.12	0.48 ± 0.17	0.46 ± 0.12	0.42 ± 0.13	0.76 ± 0.09
M3	**0.48 ± 0.11**	**0.50 ± 0.12**	**0.47 ± 0.12**	**0.44 ± 0.12**	**0.76 ± 0.08**

Results with Baseline Normalization. Given the consistent improvement of DL architectures over ML and DL baselines, the three models were further evaluated under the same LOGO setting with baseline normalization applied. Table 5 presents the results. M3 achieves the strongest performance with an accuracy of 0.66 and AUROC of 0.87, followed by M1 (accuracy: 0.64, AUROC: 0.86) and M2 (accuracy: 0.63, AUROC: 0.86). The improvement over the non-normalized setting is substantial and consistent across all three models, with gains in accuracy of approximately 18% and AUROC of around 11%, confirming that anchoring the input to participants' resting physiological state allows the models to better capture role-specific patterns under a strictly group-independent evaluation.

5 Discussion

The low performance of ML classifiers under strict LOGO evaluation indicates that while the extracted features are statistically discriminative at the population level, they do not capture enough structural information about role-dependent temporal dynamics to generalize reliably across unseen groups, consistent with the difficulty of separating subtle physiological variations within a shared social evaluative context, unlike binary state detection, where classes such as stress and rest exhibit large autonomic contrast [3, 5, 22]. DL architectures address this limitation through end-to-end learning of temporal representations, with the attention mechanisms in M2 and M3 dynamically weighting ECG and GSR modalities rather than relying on fixed fusion; M3, with its multi-scale convolutional encoders capturing patterns across multiple temporal resolutions, achieves the best performance (0.48 accuracy, 0.76 AUROC).

A key factor limiting generalization to unseen groups is inter-individual physiological variability, where signals from new participants carry individual-specific characteristics not encountered during training. Baseline normalization, which expresses each participant's signals relative to their own resting state, reduces this variability and lets the model focus on role-dependent deviations rather than absolute signal magnitudes, yielding consistent gains across all three architectures (see Table 5). Table 5 further shows that individual modalities are not equally informative for role classification: ECG alone (0.62 accuracy) outperforms GSR alone (0.55

Table 5: Classification performance of DL architectures with baseline normalization

Model	Accuracy	Precision	Recall	F1	AUROC
M1	0.64 ± 0.11	0.65 ± 0.13	0.64 ± 0.10	0.61 ± 0.11	0.86 ± 0.06
M2	0.63 ± 0.11	0.65 ± 0.12	0.64 ± 0.11	0.61 ± 0.11	0.86 ± 0.06
M3	**0.66 ± 0.11**	**0.68 ± 0.12**	**0.66 ± 0.11**	**0.63 ± 0.11**	**0.87 ± 0.07**
M3 (ECG)	0.62 ± 0.12	0.61 ± 0.13	0.62 ± 0.10	0.59 ± 0.12	0.84 ± 0.08
M3 (GSR)	0.55 ± 0.12	0.57 ± 0.15	0.55 ± 0.12	0.52 ± 0.13	0.81 ± 0.09

accuracy), while the combined model achieves the highest accuracy (0.66), consistent with prior work showing that combining physiological signals provides a more accurate representation of arousal [29]. To examine whether these gains are distributed across roles or concentrated in the baseline condition, Figure 3 presents the per-class confusion matrix for the best-performing model, M3 (baseline normalized).

From Figure 3, we found that Performer is misclassified as Anticipator almost twice as often as Observer (0.19 vs. 0.11), even though Performer is structurally adjacent to both roles with comparable frequency by protocol design. This suggests a physiological rather than structural explanation: social self-preservation theory holds that the threat of being judged by others triggers a stress response as soon as a person expects to be evaluated, not only once the evaluation begins [30]. Which means that the arousal built up while waiting to speak likely carries over into the speech itself, but there is no similar reason for it to carry over once the speech has ended.

Further from Figure 3, we also found that Anticipator and Observer are the two roles most frequently confused with each other (0.26 and 0.23, respectively), despite never being temporally adjacent for any participant under the protocol's forward-only role progression, ruling out temporal proximity as an explanation. However, evaluation apprehension theory offers a possible account: the mere presence of others capable of evaluating a person's behavior is sufficient to produce arousal, independent of whether that person is currently performing, since it is the potential for evaluation, rather than the presence of others itself, that drives the response [31, 32]. Participants were never told they were being formally judged, yet they remained among the same peers throughout the interaction, whether awaiting their own turn or having just completed it, and these peers occupy a potentially evaluative position throughout. This shared exposure may explain why the two roles are misclassified as each other more often than as Performer, which is an actively scrutinized role.

What is less clear is why this similarity extends specifically into the Observer window, once the participant's own turn has already ended. One possible, but more uncertain, explanation comes from the perseverative cognition hypothesis: what sustains stress-related arousal is not only the stressful event itself but continuing to think about it, either before it happens or after it is over [33]. We treat this explanation as speculative: participants were never instructed to reflect on their own performance during the Observer phase, and since Observer always overlaps with another participant currently speaking, the sustained arousal observed in this window could instead, or additionally, reflect attentional engagement with that speaker rather than reflection on one's own completed turn. To our knowledge, no prior work has tested which of these explanations holds in a group setting such as ours, so we leave this as an open question rather than a firm conclusion.

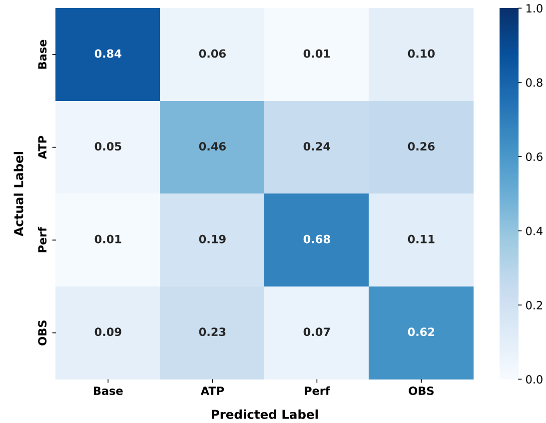

Figure 3: Average confusion matrix across roles (row-normalized). Diagonal values (bold) indicate per-class recall.

5.1 Limitations

The current work has several limitations that point to directions for future investigation. First, the role classification framework is evaluated exclusively within a controlled sequential speech activity, where the group structure, task duration, and interaction format are fixed. Whether role-specific physiological patterns generalize to less structured group interactions; such as collaborative tasks, meetings, or open discussions where roles are fluid and not formally defined, remains an open question. Second, in the current setting participants are aware of the speaking order in advance, meaning the anticipatory physiological response is informed by explicit foreknowledge of when they will perform. In naturalistic group interactions, turn-taking is dynamic and unpredictable, which may produce qualitatively different anticipatory responses. Extending the role-based framework to settings with dynamic role assignment represents a natural and important direction for future work.

6 Conclusion

We presented a role-aware framework for physiological sensing in structured group interactions using wearable sensing. In a controlled study involving 100 participants organized into small groups, ECG and GSR signals were simultaneously recorded across Performer, Anticipator, Observer, and Baseline conditions. Our findings demonstrate that interaction roles induce distinguishable physiological responses, including observer responses that remain separable from baseline despite the absence of active task participation. These results highlight the importance of role-level segmentation for modeling physiological responses in shared social environments.

Classification under a LOGO evaluation reveals a clear progression from ML to DL, with baseline normalization providing consistent gains across all models. M3 achieves the strongest performance with an accuracy of 0.66 and AUROC of 0.87, demonstrating that role-level segmentation combined with subject-relative normalization and end-to-end temporal learning provides a viable foundation for physiological modeling in group settings.

References

[1] Ruben T. Azevedo, Nell Bennett, Andreas Bilicki, Jack Hooper, Fotini Markopoulou, and Manos Tsakiris. The calming effect of a new wearable device during the anticipation of public speech. *Scientific reports*, 7(1):2285, 2017.

[2] Nilesh Kumar Sahu, Snehil Gupta, and Haroon R Lone. Exploring heart rate variability and heart rate dynamics using wearables before, during, and after speech activity: Insights from a controlled study in a low-middle-income country. In *Proceedings of the 2025 ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, pages 362–378, 2025.

[3] Varun Mishra, Sougata Sen, Grace Chen, Tian Hao, Jeffrey Rogers, Ching-Hua Chen, and David Kotz. Evaluating the reproducibility of physiological stress detection models. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 4(4):1–29, 2020.

[4] Fang Li and Dan Zhang. Multimodal physiological signals from wearable sensors for affective computing: A systematic review. *Intelligent Sports and Health*, 1(4):210–222, 2025.

[5] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[6] Shkurta Gashi, Elena Di Lascio, and Silvia Santini. Using unobtrusive wearable sensors to measure the physiological synchrony between presenters and audience members. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(1):1–19, 2019.

[7] Yang Liu, Tingting Wang, Kun Wang, and Yu Zhang. Collaborative learning quality classification through physiological synchrony recorded by wearable biosensors. *Frontiers in Psychology*, 12:674369, 2021.

[8] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–21, 2018.

[9] Emma R Toner, Mark Rucker, Zhiyuan Wang, Maria A Larrazabal, Lihua Cai, Debajyoti Datta, Haroon Lone, Mehdi Boukhechba, Bethany A Teachman, and Laura E Barnes. Wearable sensor-based multimodal physiological responses of socially anxious individuals in social contexts on zoom. *IEEE transactions on affective computing*, 2025.

[10] Ivo V Stuldreher, Nattapong Thammasan, Jan BF van Erp, and Anne-Marie Brouwer. Physiological synchrony in eeg, electrodermal activity and heart rate reflects shared selective auditory attention. *Journal of neural engineering*, 17(4):046028, 2020.

[11] Mouna Benchekroun, Pedro Elkind Velmovitsky, Dan Istrate, Vincent Zalc, Plinio Pelegrini Morita, and Dominique Lenne. Cross dataset analysis for generalizability of hrv-based stress detection models. *Sensors*, 23(4):1807, 2023.

[12] Emily Zhou, Mohammad Soleymani, and Maja J Matarić. Investigating the generalizability of physiological characteristics of anxiety. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 4848–4855. IEEE, 2023.

[13] Yekta Said Can, Bert Arnrich, and Cem Ersoy. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of biomedical informatics*, 92:103139, 2019.

[14] Maria Kyrou, Ioannis Kompatsiaris, and Panagiotis C Petrantonakis. Deep learning approaches for stress detection: A survey. *IEEE Transactions on Affective Computing*, 16(2):499–517, 2024.

[15] Kyriaki Kalimeri and Charalampos Saitis. Exploring multimodal biosignal features for stress detection during indoor mobility. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 53–60, 2016.

[16] Zhen Li, Yun Xing, Yao Pi, Mingzhe Jiang, and Lejun Zhang. A novel physiological feature selection method for emotional stress assessment based on emotional state transition. *Frontiers in Neuroscience*, 17:1138091, 2023.

[17] Fahad Khan, Xiaojun Yu, Zhaohui Yuan, and Atiq Ur Rehman. Ecg classification using 1-d convolutional deep residual neural network. *Plos one*, 18(4):e0284791, 2023.

[18] Anubhav Bhatti, Behnam Behinaein, Dirk Rodenburg, Paul Hungler, and Ali Etemad. Attentive cross-modal connections for deep multimodal wearable-based emotion recognition. In *2021 9th international conference on affective computing and intelligent interaction workshops and demos (ACIIW)*, pages 01–05. IEEE, 2021.

[19] Rafael Wespi, Andrea N Neher, Tanja Birrenbach, Stefan K Schauber, Marie Ottilie Frenkel, Helmut Schrom-Feiertag, Thomas C Sauter, and Juliane E Kämmer. Physiological team dynamics explored: physiological synchrony in medical simulation training. *Advances in Simulation*, 10(1):5, 2025.

[20] Patrícia Bota, Tianyi Zhang, Abdallah El Ali, Ana Fred, Hugo Plácido Da Silva, and Pablo Cesar. Group synchrony for emotion recognition using physiological signals. *IEEE Transactions on Affective Computing*, 14(4):2614–2625, 2023.

[21] Soundariya Ananthan, Nan Gao, and Flora D Salim. Understanding physiological responses of students over different courses. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers*, pages 104–110, 2024.

[22] Aamir Arsalan and Muhammad Majid. Human stress classification during public speaking using physiological signals. *Computers in biology and medicine*, 133:104377, 2021.

[23] Zhiyuan Wang, Maria A Larrazabal, Mark Rucker, Emma R Toner, Katharine E Daniel, Shashwat Kumar, Mehdi Boukhechba, Bethany A Teachman, and Laura E Barnes. Detecting social contexts from mobile sensing indicators in virtual interactions with socially anxious individuals. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 7(3):1–26, 2023.

[24] Patrícia Bota, Rafael Silva, Carlos Carreiras, Ana Fred, and Hugo Plácido da Silva. Biosppy: A python toolbox for physiological signal processing. *SoftwareX*, 26:101712, 2024.

[25] Dominique Makowski, Tam Pham, Zen J Lau, Jan C Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and SH Annabel Chen. Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior research methods*, 53(4):1689–1696, 2021.

[26] Shirley Dowdy, Stanley Wearden, and Daniel Chilko. *Statistics for research*. John Wiley & Sons, 2011.

[27] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[28] Simon JD Prince. *Understanding deep learning*. MIT press, 2023.

[29] Santiago Sosa, Adam K Fontecchio, Evangelia G Chrysikou, and Jennifer S Atchison. Beyond eda: A systematic review of multimodal sympathetic nervous system arousal classification for stress detection. *Sensors*, 26(5):1584, 2026.

[30] Sally S Dickerson and Margaret E Kemeny. Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. *Psychological bulletin*, 130(3):355, 2004.

[31] Nickolas B Cottrell, Dennis L Wack, Gary J Sekerak, and Robert H Rittle. Social facilitation of dominant responses by the presence of an audience and the mere presence of others. *Journal of personality and social psychology*, 9(3):245, 1968.

[32] Charles F Bond and Linda J Titus. Social facilitation: a meta-analysis of 241 studies. *Psychological bulletin*, 94(2):265, 1983.

[33] Jos F Brosschot, William Gerin, and Julian F Thayer. The perseverative cognition hypothesis: A review of worry, prolonged stress-related physiological activation, and health. *Journal of psychosomatic research*, 60(2):113–124, 2006.