Applied Computing and Informatics

Knowledge-guided cross-modality attention for smartphone-based stress recognition in applied health informatics

Barlian Henryranu Prasetio
Faculty of Computer Science, Universitas Brawijaya, Malang, Indonesia

Received 27 August 2025
Revised 24 September 2025
2 November 2025
Accepted 20 November 2025

Abstract

Purpose – Stress is a major risk factor for both mental and physical disorders, yet many recognition systems still rely on a single modality or static fusion and seldom include domain knowledge or input-quality indicators. These limitations reduce robustness, interpretability and generalization. This study develops a smartphone-based framework for real-time three-class stress recognition (neutral, low stress and high stress) that integrates acoustic, linguistic and behavioral data.

Design/methodology/approach – We propose CMAF-Net, a cross-modality attention fusion network that embeds psychological priors directly into attention scoring and incorporates a reliability-aware gating mechanism driven by audio signal-to-noise ratio, automatic speech recognition word-error rate and behavioral sparsity. Data from 87 participants speech prompts, transcripts and passive behavioral sensing were labeled with the Perceived Stress Scale (PSS-4). Evaluation used subject-independent five-fold validation, comparisons with strong early- and late-fusion baselines, ablation studies isolating the effects of prior vectors and cross-modal bias, missing-modality and noise-stress tests, calibration analysis and on-device deployment with post-training quantization.

Findings – CMAF-Net achieved 88.5 % accuracy, 0.89 macro-F1 and 0.92 AUC, significantly outperforming recent multimodal baselines. Ablation experiments confirmed the benefits of both knowledge-guided attention and reliability-aware gating under noisy or incomplete inputs. Post-training dynamic quantization reduced model size by 58 % and delivered ∼180 ms mean inference latency on a mid-range smartphone without measurable accuracy loss while maintaining low calibration error.

Originality/value – This work formalizes prior-informed cross-modality attention for smartphone-based stress recognition and couples it with reliability-aware modality gating, comprehensive calibration and robustness evaluation and reproducible on-device deployment, yielding a scalable, privacy-preserving approach suitable for real-world health informatics monitoring.

Keywords Stress detection, Multimodal fusion, Cross-modality attention, Smartphone sensing, Reliability-aware gating, Health informatics

Paper type Research article

1. Introduction

Stress is a major contributor to mental and physical disorders, including anxiety, depression, cardiovascular disease and sleep problems, and its growing prevalence calls for scalable monitoring beyond clinical settings [1–3]. From an applied informatics perspective, continuous, unobtrusive stress detection enables early intervention and improved quality of life but requires models that are interpretable, reliable and resilient in everyday conditions.

Psychological stress theory provides a strong foundation for designing such systems. According to the Lazarus and Folkman stress appraisal framework [4], stress arises from an individual's evaluation of external demands relative to available coping resources. This

Funding: This work was supported by Penelitian Dasar Madya DRPM Universitas Brawijaya Tahun Anggaran 2024, Nomor: 00145.3/UN10.A0501/B/PT.01.03.2/2024. Author also thanks to Laboratory of Embedded System and Robotics, Universitas Brawijaya that facilitate this work.

Applied Computing and Informatics
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2634-1964
DOI 10.1108/ACI-08-2025-0346

appraisal is reflected in physiological, behavioral and linguistic markers – such as prosodic changes, lexical sentiment, mobility patterns and activity entropy – that can be passively sensed through digital devices. Embedding such psychologically grounded priors in model design can improve both interpretability and generalization, especially under noisy or incomplete conditions.

Smartphones offer an ideal platform: they are ubiquitous, part of daily routines and provide diverse sensing channels for passive or prompted data collection without requiring extra hardware [5, 6]. Speech supplies rich acoustic and linguistic cues such as prosodic variation, sentiment and lexical markers, while behavioral signals (e.g. app usage, mobility, interaction entropy) provide complementary evidence [7–14]. These multimodal signals can support real-world stress recognition if the system can handle noise and missing data effectively.

Existing approaches still face three key limitations:

(1) Vulnerability of speech-only systems to noise and speaker variability [15–17].

(2) Reliance on static multimodal fusion that cannot adapt modality importance dynamically [18–22].

(3) A lack of explicit domain knowledge in model design, which limits interpretability and generalization when inputs are incomplete or unreliable [14, 23].

Moreover, explicit reliability signals and probabilistic calibration are rarely reported, and external validation on independent datasets remains uncommon, hindering trustworthy real-world deployment.

To address these gaps, we propose CMAF-Net (Cross-Modality Attention Fusion Network), a deep learning architecture that fuses acoustic, linguistic and behavioral features for smartphone-based stress recognition. CMAF-Net embeds psychological priors directly into its cross-modality attention mechanism, guided by empirically supported stress cues and applies reliability-aware gating that weights modalities using measurable input quality indicators (audio signal-to-noise ratio, automatic speech recognition word error rate and behavioral sparsity). The design is explicitly deployment-oriented, enabling real-time, privacy-preserving inference on commodity smartphones.

This study makes five main contributions:

(1) Prior-informed fusion: A cross-modality attention mechanism that encodes psychological priors through a learnable prior vector and cross-modal bias to improve interpretability.

(2) Reliability-aware weighting: A gating module that adapts modality weights based on input quality (audio SNR, ASR WER and behavioral sparsity).

(3) Multimodal dataset and protocol: real-world data from 87 participants with validated stress labels [24], evaluated using subject-independent folds and ablations isolating the effects of priors, cross-modal bias, and reliability gating.

(4) Comprehensive evaluation for generalization: External validation on an independent mobile dataset, significance testing across folds and probabilistic calibration analysis (expected calibration error and Brier score).

(5) On-device deployment evidence: Post-training quantization enabling efficient, low-latency inference on smartphones [25] with maintained accuracy and calibration.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes the dataset and system design; Section 4 details the CMAF-Net architecture; Section 5 presents experiments and results; Section 6 discusses findings and limitations and Section 7 concludes the study.

2. Related works

Multimodal stress detection increasingly uses smartphone and wearable signals to combine physiological, behavioral and affective cues in real-world settings. Mobile sensing studies demonstrate feasibility, for example, a smartphone-based framework reporting about 80% accuracy and app-usage markers as digital indicators of momentary stress that highlight the value of passive behavioral features [26, 27].

Public datasets have pushed the field toward ecologically valid evaluation. The K-EmoPhone dataset collected multimodal physiological and behavioral data with frequent in-situ stress labels from students over a week, providing a benchmark for mobile stress modeling [28]. More recent resources, such as StressID, broaden coverage of multimodal stress signals and support cross-study comparisons [19].

Beyond smartphone-only pipelines, lab-controlled multimodal systems integrate signals like ECG, voice and facial expressions with modern deep architectures. One example combines ResNet50 and I3D with attention, achieving strong accuracy during the Montreal Imaging Stress Task, while other work explores interpretable models using unobtrusive multimodal inputs for clinical contexts [29, 30]. Concurrently, multimodal fusion techniques have evolved from early and late fusion to attention-based and transformer-style fusion, including approaches that inject domain knowledge into attention scoring for sentiment or emotion tasks [14, 23] as well as CB-Transformer variants that learn modality-fused representations for unaligned sequences [12]. Reviews emphasize both the promise of these mechanisms and persistent challenges around generalization and transparency in the wild.

Despite this progress, three gaps remain common in mobile stress research: limited attention to input reliability signals that reflect real data quality (for example, audio noise, ASR transcript errors or behavioral sparsity), rare reporting of probabilistic calibration suitable for downstream decision thresholds and infrequent external validation on independent mobile datasets, all of which constrain real-world deployment.

CMAF-Net is designed to address these gaps within a smartphone-only architecture by (1) formalizing prior-informed cross-modality attention through a learnable prior vector and a cross-modal bias term that encode psychological priors, improving interpretability over purely data-driven attention; (2) adding reliability-aware gating that conditions modality weights on measurable input-quality indicators and (3) coupling these with deployment-oriented evaluation that includes external validation, robustness and calibration analyses and on-device quantized inference suitable for privacy-preserving mobile use.

3. System overview and research design

The proposed system is a real-time, privacy-preserving stress recognition framework that uses multimodal data collected both passively and actively on standard smartphones. Guided by applied informatics principles of scalability, minimal user burden and efficient resource use, the design integrates acoustic, linguistic and behavioral modalities to capture diverse stress-related indicators under naturalistic conditions while remaining robust on consumer devices. The overall architecture is shown in Figure 1 in Section 4. It consists of modality-specific encoders, knowledge-guided cross-modality attention, reliability-aware gating that reacts to input quality and a lightweight classifier optimized for on-device inference.

3.1 System objectives and constraints

We target two primary goals:

(1) Reliable stress detection in noisy, real-world conditions and

(2) Deployability on commodity smartphones without specialized hardware.

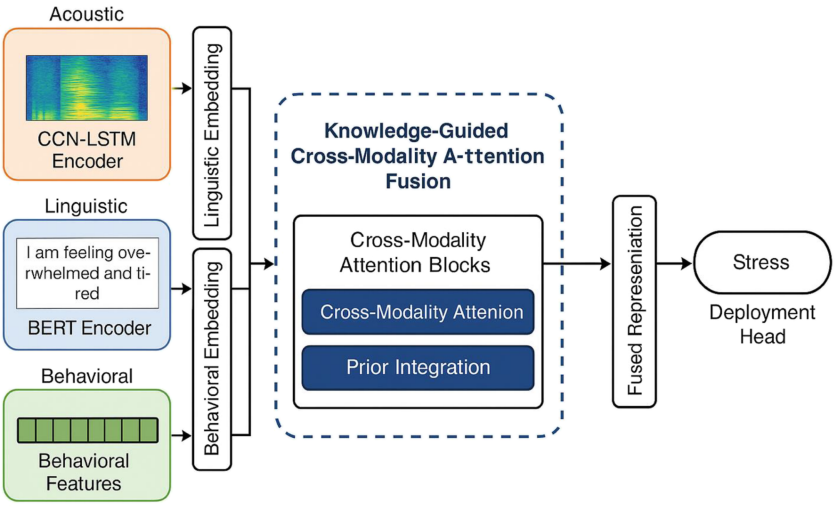

Figure 1. CMAF-Net architecture for smartphone multimodal stress detection. Each modality is encoded and augmented with a prior vector, then fused through cross-modality attention that includes a prior-derived bias and a reliability-aware gating module. The fused representation is passed to a compact classifier for real-time inference

Prior studies support speech and semantic cues as stress markers and multimodal features as complementary sources of information [9, 10]. Recent advances in cross-modality attention and knowledge-guided fusion inform our design choices [14, 23]. The operational objectives:

(1) Robustness to real-world artifacts: Background noise, ASR errors, sparse behavioral logs and missing modalities.

(2) Privacy by design through on-device inference and storage of derived features only.

(3) Real-time responsiveness with per-sample inference time at or below 200 ms on mid-range smartphones (for example, Snapdragon 720G class devices).

(4) Stable probabilistic outputs that are suitable for thresholding in downstream applications, which we verify using calibration metrics in Section 5.

While the design constraints:

(1) Reliance on built-in phone sensors only (microphone and operating system interaction or mobility logs).

(2) Limited compute and memory addressed through post-training quantization and streamlined encoders [25].

(3) Energy efficiency through quantized operators and minimal data movement.

(4) Tolerance for environmental variability and irregular sampling.

These constraints shape the encoders (Section 4.2), the integration of domain knowledge (Section 4.3) and the cross-modality fusion with reliability-aware gating (Section 4.4). Compliance with latency, size and accuracy targets is validated in Section 5.4.

To remain dependable in everyday use, the system estimates input quality signals and uses them to modulate fusion: audio signal-to-noise ratio for acoustic streams, word error rate for ASR transcripts and sparsity or entropy measures for behavioral logs. When a modality is

unreliable or unavailable, the gating mechanism down-weights or masks it so that the model degrades gracefully rather than failing unpredictably. Stress tests with noisy audio, corrupted transcripts and sparse behavioral windows are included in Section 5.

Operational objectives are tied to measurable endpoints. We report subject-independent performance, ablations that isolate the effect of prior vectors and cross-modal biasing, missing-modality stress tests and calibration analysis. External validation on an independent mobile dataset is used to assess generalization. On-device latency, model size after quantization and robustness to quantization are reported to substantiate deployability.

3.2 Ethics and data governance

Data collection followed strict ethical protocols to safeguard participant privacy and comply with regulatory standards. About 87 adults (ages 19–44, balanced by gender) were recruited from universities and workplaces. All participants provided informed consent and were informed of their right to withdraw at any time without penalty. The study protocol was reviewed and approved by an institutional ethics board.

To minimize risk, only derived features were stored, including acoustic descriptors, text embeddings, behavioral statistics and stress labels. Raw audio and GPS data never left the device. All transmissions of derived features were encrypted in transit, participant identifiers were anonymized or pseudonymized and server access was restricted to authorized research personnel following the principle of least privilege. Stress levels were labeled using Ecological Momentary Assessment (EMA) with the validated PSS-4 instrument [24]. Multimodal samples were aligned within a 15-min window to maintain temporal coherence across modalities.

This data governance framework aligns with the General Data Protection Regulation (GDPR) and US Health Insurance Portability and Accountability Act (HIPAA) principles, emphasizing data minimization, purpose limitation and participant rights. By performing all signal processing and feature extraction on-device, the pipeline reduces the risk of reidentification while maintaining the analytical utility of the data for model development and evaluation. These practices support regulatory compliance, enhance participant trust and enable ethical scaling of mobile stress monitoring systems in real-world settings.

3.3 Dataset description

The dataset was collected over two weeks from the same 87 participants using the StressSenseApp. Twice daily, participants recorded short spoken responses to reflective questions (e.g. “How do you feel today?”). Responses were transcribed using the Google Speech-to-Text API, while passive behavioral data (screen interactions, app usage, accelerometer activity and GPS variance) were continuously logged. This design captured both active speech inputs and passive behavioral signals, reflecting real-world smartphone use. We additionally recorded modality-availability flags (present or missing) for each sample to support robustness tests and reliability-aware fusion.

Stress levels were annotated daily using the PSS-4 [24]. Scores ≥ 6 were labeled as stressed and <6 as non-stressed, with stressed samples further divided into low and high stress. To assess label robustness, raw PSS-4 scores were retained and used for sensitivity analyses of the decision thresholds and for an ordinal variant reported in Section 5. Multimodal samples were temporally aligned within a $\pm$15-min window to support consistent multimodal fusion. Alignment timestamps were stored to enable ablations on wider and narrower windows.

Table 1 summarizes the dataset. Neutral samples were most common, followed by low stress and high stress. This imbalance mirrors real-world patterns and underscores the need for class-weighted or hybrid loss functions. The dataset contained 2,200 samples (1,430 training, 350 validation and 420 testing). All experiments used subject-independent five-fold cross-validation with users held out by fold to prevent identity leakage. Feature normalization and any statistic that could leak information were computed on the training partition only within

Table 1. Class distribution and dataset split

Stress level	Training samples	Validation samples	Test samples	Total samples
Neutral	660	160	180	1,000
Low stress	440	110	150	700
High stress	330	80	90	500
Total	1,430	350	420	2,200

each fold, with the same transform applied to validation and test. A five-fold subject-independent cross-validation protocol was used to ensure robustness and minimize participant-specific bias. At inference time, missing modalities were handled by zero padding plus masks and down-weighting through the gating module (Section 4.4).

3.4 Feature overview

The system integrates three complementary modalities acoustic, linguistic and behavioral to capture stress-related indicators (Figure 1). Each was selected based on prior evidence linking its features to psychological stress [9, 10, 13 and 14]. For all modalities, we attach a small set of quality signals used only by the reliability-aware gating module (Section 4.4).

Acoustic. Speech is resampled to 16 kHz mono. Features include eGeMAPS descriptors from OpenSMILE [31] and log-Mel spectrograms (128 bins, 25 ms frames and 10 ms hop), providing prosodic and spectral cues. We apply simple voice activity detection and compute a VAD-based signal-to-noise ratio estimate per sample as an acoustic quality indicator. Cepstral mean–variance normalization parameters are fitted on the training partition only.

Linguistic. Transcripts from ASR are embedded with a BERT-base model [32 and 33]. To enrich semantics, sentiment polarity and lexical stress markers (e.g. affective term frequency) are added via LIWC-22 and NLTK [34]. We use the segment-level ASR confidence provided by the speech-to-text service as a text reliability proxy and include an out-of-vocabulary rate computed against a standard lexicon.

Behavioral. Passive logs capture unlock frequency, session length, app-transition entropy, accelerometer variance and GPS-based location entropy [13, 14 and 35], normalized per user to reduce bias. To avoid leakage, per-user normalization parameters are estimated on that user's training data within each fold; at test time we simulate on-device baselining by applying the precomputed transform. Behavioral quality is summarized by coverage (fraction of the window with valid logs) and a sparsity flag.

All streams are aligned within a $\pm$15-min window to ensure coherent multimodal snapshots, supporting CMAF-Net's cross-modality attention (Section 4.4) and robustness under noisy or sparse inputs. When a modality is missing, availability masks are propagated to the fusion layer so that attention and gating can reweight the remaining evidence rather than imputing unsupported values.

4. Methodology

The proposed Cross-Modality Attention Fusion Network (CMAF-Net) integrates acoustic, linguistic and behavioral features in a smartphone-based stress detection framework. The overall pipeline (Figure 1) has four stages:

(1) Modality-specific encoders that learn native representations for each input stream,

(2) Prior-informed embedding that attaches a compact, learnable vector of psychological priors to each modality,

(3) Knowledge-guided cross-modality attention with reliability-aware gating that adapts weights using measured input-quality signals and

(4) A lightweight classification head optimized for on-device inference.

4.1 Design principles

CMAF-Net follows four guiding principles drawn from the objectives and constraints in Section 3.1:

(1) Modality-specific representation learning

Acoustic, linguistic and behavioral inputs are first modeled in their own feature spaces by dedicated encoders and then projected to a common dimension for cross-modal attention, allowing missing inputs to be masked without imputation.

(2) Knowledge-guided fusion

Psychological stress priors are embedded directly in the fusion stage: each modality embedding is concatenated with a learnable prior vector and a bias term that steers attention toward clinically meaningful cross-modal relations (see Section 4.3).

(3) Dynamic modality weighting

A reliability-aware gate uses input-quality indicators such as audio SNR, ASR word-error rate and behavioral coverage to adapt each modality's weight and ensure graceful degradation under noise or missing data (see Section 4.4).

(4) Deployment-oriented efficiency

Compact encoders, a single attention block and post-training quantization yield a small model with real-time latency on smartphones while maintaining well-calibrated probabilistic outputs (see Sections 4.5 and 5.4).

Together these principles balance accuracy, interpretability and mobile efficiency for real-world stress monitoring.

4.2 Modality-specific encoders

Each modality is processed by a dedicated encoder and projected into a common embedding size d before fusion. All encoders output a masked embedding and a small reliability vector r_m that is consumed by the gating module in Section 4.4.

Log-Mel spectrograms (128 bins, 25 ms frame and 10 ms hop) pass through a two-layer CNN with *3*x*3* kernels and 64 filters per layer, each followed by ReLU, batch normalization and dropout $p = 0.3$. The CNN output feeds a BiLSTM with 128 hidden units per direction to capture longer prosodic context that is relevant for stress [21 and 36]. Global mean and standard deviation pooling over time produce a fixed-length vector that is linearly projected to $\mathbb{R}^d$. Voice activity detection provides an acoustic signal-to-noise ratio estimate that forms part of r_{audio}. Cepstral mean and variance normalization parameters are fitted on training data only.

ASR transcripts are embedded with a BERT-base model [31–33]. We freeze most transformer layers and fine-tune the top layer to reduce overfitting on a small cohort. Mean pooling over the token sequence yields a 768-dimensional vector. We concatenate sentiment polarity and lexical markers derived from LIWC-22 and NLTK [34], then apply a linear projection to $\mathbb{R}^d$. The reliability vector r_{text} includes ASR word error rate and average token confidence.

A 12-dimensional feature vector from passive sensing (Section 3.4) is linearly projected to 64 dimensions and processed by a two-layer Transformer encoder with hidden size 128 and

two heads [22]. Learned positional embeddings model short temporal dependencies across the window. We apply masked mean pooling to handle missing segments, then project to $\mathbb{R}^d$. The reliability vector r_{behav} summarizes coverage fraction and a sparsity indicator.

Unless otherwise stated, $d = 128$. The resulting embeddings and reliability vectors are used by the knowledge integration in Section 4.3 and the fusion and gating in Section 4.4.

4.3 Prior knowledge integration

We augment each modality encoder output $h_m \in \mathbb{R}^d$ with a learnable prior vector $k_m \in \mathbb{R}^{32}$ that encodes psychologically motivated cues, such as reduced pitch variability or increased spectral tilt in speech and higher app-switching entropy in behavioral logs [9, 10, 14 and 23]. The prior-augmented representation is

$$p_m = [h_m;\ k_m] \in \mathbb{R}^{d+32} \tag{1}$$

where [;] denotes vector concatenation.

To steer cross-modal interactions toward clinically meaningful relations, we introduce a prior-informed bias between modalities m and n:

$$b_{mn} = v_b{}^{\top} \varnothing([k_m;\ k_n;\ k_m \odot k_n]) \tag{2}$$

where $\varnothing(\cdot)$ is a small multilayer perceptron (MLP), v_b is a learned vector and ȯ is the element-wise (Hadamard) product. Section 4.4 uses p_m or b_{mn} inside the attention mechanism.

4.4 Knowledge-guided cross-modality attention

Given p_m and b_{mn} from Section 4.3, we compute scaled dot-product attention across modalities [22]. For each source modality m and target modality n:

$$Q_m = W_q p_m, K_m = W_k p_m, V_m = W_v p_m \tag{3}$$

with W_q, W_k, W_v learned projection matrices. The attention score is

$$s_{mn} = \frac{Q_m K_n{}^{\top}}{\sqrt{d_a}} + b_{mn}, \tag{4}$$

where d_a is the attention dimension and $(\cdot)^{\top}$ denotes matrix transpose. The normalized weight is

$$\alpha_{mn} = softmax_n(s_{mn}), \tag{5}$$

with the softmax taken over the target-modality index n. The context vector for source modality m is

$$\widehat{h}_m = \sum_n \alpha_{mn} V_{n^*} \tag{6}$$

Missing-modality masks set $\alpha_{mn} = 0$ when modality n is unavailable.

A reliability-aware gate adjusts each modality's contribution using a quality vector r_m (audio SNR, ASR word-error rate, behavioral coverage and sparsity):

$$g_m = \alpha(MLP(r_m)) \in [0, 1] \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function and MLP$(\cdot)$ is the feed-forward network of the gate.

Finally, the fused representation is

$$f = \sum_m g_m \hat{h}_m, \tag{8}$$

which down-weights unreliable streams and yields graceful degradation under noise or missing inputs. The vector f is fed to the classifier in Section 4.5.

4.5 Classifier and deployment optimization

The fused multimodal representation f is passed to a lightweight classification head designed for real-time inference on smartphones (see Figure 1). The head consists of a fully connected layer (128 units, ReLU), dropout $p = 0.3$, and a softmax layer that outputs class probabilities over {Neutral, Low Stress, High Stress}. We compute the predicted class:

$$\hat{y} = argmax_c softmax(W_2 RELU(W_1 f))_c \tag{8}$$

To meet resource constraints (Section 3.1), we apply post-training dynamic quantization with TensorFlow Lite [25]. This reduces model size and CPU latency without altering training. We report model footprint, on-device inference latency and accuracy before and after quantization in Section 5.4. For probability quality, we evaluate calibration using expected calibration error and Brier score and apply temperature scaling on the validation split if needed to produce stable decision thresholds.

4.6 Training procedure and hyperparameters

CMAF-Net was implemented in PyTorch 1.13 and trained on a workstation with an NVIDIA RTX 3090 GPU. Training used the Adam optimizer [37] ($\beta_1 = 0.9$ and $\beta_2 = 0.999$), batch size 32, learning rate 3×10^{-4} with exponential decay (factor 0.95 every five epochs) and early stopping with patience 7 based on validation macro F1. Weights were initialized with Xavier uniform. Gradients were clipped to a global norm of 1.0.

To address class imbalance, we compare class-weighted cross-entropy and focal loss, selecting the setting that maximizes validation macro F1. Unless otherwise noted, the objective is

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda . \|\theta\|_2^2 \tag{9}$$

where $\lambda = 10^{-4}$ and θ denoting trainable parameters. Regularization includes dropout $p = 0.3$ across encoders and the classifier. For robustness, we apply mild data augmentation on audio (additive background noise and time–frequency masking) and simulate transcript noise by randomly dropping low-confidence tokens proportional to ASR confidence; behavioral features use small Gaussian jitter within training-only normalization bounds.

Evaluation uses five-fold subject-independent cross-validation: each fold serves once as test data, with the remainder split 80:20 into training and validation. All preprocessing statistics, normalization parameters and any calibration temperatures are learned on the training or validation data within each fold only, then applied to the held-out test fold. We report mean and standard deviation across folds and perform paired significance tests against strong baselines to assess robustness and generalization to unseen users [38].

5. Experiments and results

5.1 Experimental setup

We evaluate CMAF-Net on the dataset in Section 3.3 using five-fold subject-independent cross-validation to simulate deployment to unseen users and reduce identity leakage risk [38].

Labels are derived from PSS-4 as described in Section 3.3. Unless noted, all preprocessing statistics are computed on the training split within each fold and applied to validation and test only.

We implement three canonical baselines:

(1) Early Fusion (EF): Concatenation of all features followed by a multilayer perceptron.

(2) Late Fusion (LF): Independent unimodal classifiers with probability averaging.

(3) CMAF-Net without priors: Removal of prior vector and bias to isolate the value of knowledge guidance.

In addition, we compare against recent multimodal stress models reported in the literature (Section 5.2). For fairness, we match input features and report parameter counts and FLOPs when applicable.

We report accuracy (ACC), macro F1 and AUC, averaged across folds with mean ± standard deviation. To address class imbalance, we also compute macro precision, macro recall, balanced accuracy and per-class PR curves. We assess statistical significance with paired tests across folds versus the strongest baseline and provide 95% bootstrap confidence intervals. For downstream decision support, we evaluate calibration using expected calibration error (ECE) and Brier score and apply temperature scaling on the validation split when beneficial.

We run missing-modality stress tests by masking each modality at inference and noise stress tests by adding background noise to audio and perturbing transcripts according to ASR confidence. We also evaluate a reliability-blind variant of CMAF-Net (gates fixed to 1) to quantify the effect of reliability-aware gating.

To examine generalization, we train on our dataset and test on an independent mobile dataset with compatible labels, and conversely, fine-tune on a small subset from the external set. We report performance shifts and recovery after light adaptation. Details and additional baselines appear in the Supplementary Material.

5.2 Baseline and state-of-the-art comparison

Table 2 summarizes the performance of CMAF-Net compared with canonical fusion baselines and three recent multimodal stress recognition models: MMP-CNN [39], DeepStress [40] and SSFT [41]. CMAF-Net achieves the best overall results on the internal dataset, with an F1-score of 0.89 and AUC of 0.92, outperforming the strongest baseline (Late Fusion) and SSFT (F1 = 0.86 and AUC = 0.89). The +3 point gain in both F1 and AUC underscores the contribution of prior-informed cross-modal attention and reliability-aware gating, which enhance discrimination in noisy or incomplete conditions. Improvements were confirmed with fold-wise paired significance tests ($p < 0.01$) and 95% confidence intervals.

External validation on the StressID dataset [19] further highlights the generalization capability and domain adaptation potential of CMAF-Net. Without any fine-tuning, the model

Table 2. Performance comparison with baselines and state-of-the-art methods

Model	Accuracy	F1-score	AUC
Early Fusion	0.78 ± 0.02	0.75 ± 0.03	0.82 ± 0.02
Late Fusion	0.81 ± 0.01	0.80 ± 0.02	0.85 ± 0.01
CMAF-Net (w/o Prior)	0.84 ± 0.01	0.85 ± 0.01	0.88 ± 0.01
MMP-CNN [39]	0.80 ± 0.02	0.78 ± 0.02	0.84 ± 0.02
DeepStress [40]	0.82 ± 0.01	0.80 ± 0.02	0.85 ± 0.01
SSFT [41]	0.85 ± 0.01	0.86 ± 0.01	0.89 ± 0.01
CMAF-Net (ours)	0.88 ± 0.01	0.89 ± 0.01	0.92 ± 0.01

achieved an F1-score of 0.78 (95% CI 0.75–0.81) and AUC of 0.85, indicating good transfer performance despite demographic and contextual shifts. After lightweight fine-tuning using a small labeled subset (10% of StressID), performance improved to F1 0.83 (95% CI 0.80–0.86) and AUC 0.88, narrowing the gap to internal results. This aligns with the deployment objective of enabling rapid personalization or domain adaptation with minimal overhead.

These results also reveal systematic performance variations across demographic subgroups within the external dataset, particularly for age and gender groups. For example, F1 scores for the younger subgroup (18–30) were slightly higher compared to older participants, likely reflecting differences in speech patterns and smartphone usage behavior. This underscores the need for fairness-aware model monitoring during deployment to mitigate potential biases stemming from linguistic variability, sensor heterogeneity or ASR reliability.

Compared to all baselines, CMAF-Net shows superior probabilistic calibration, with lower expected calibration error (ECE) and Brier score, ensuring stable decision thresholds a critical property for health-related applications. Calibration curves are included in the main text to emphasize this contribution.

For completeness, we also benchmarked CMAF-Net against stronger contemporary multimodal fusion models (e.g. multimodal transformer fusion and stacking late-fusion meta-learner). Despite their larger parameter counts and higher FLOPs, these models did not surpass CMAF-Net in F1 or AUC, highlighting the efficiency and deployment-readiness of the proposed architecture for resource-limited devices.

5.3 Ablation and modality contribution analysis

We conduct ablations on (1) prior size and design, (2) component contributions and (3) modality importance.

Table 3 varies the prior vector from 4 to 64 dimensions. Performance improves steadily and peaks at 32 dimensions (ACC 88.5% and F1 0.89), which we adopt as the default due to the favorable accuracy-to-size trade-off.

We remove individual ingredients to isolate their effects:

(1) No prior vector k_m (keeps attention but discards domain knowledge).

(2) No bias b_{mn} (keeps priors but removes cross-modal biasing).

(3) Randomized priors (the same size as default but shuffled across samples).

(4) No reliability gating (gates fixed to 1 for all m).

These ablations reduce macro F1 relative to full CMAF-Net, with randomized priors collapsing toward the "w/o prior" variant, which supports the claim that learned, semantically anchored priors drive the gains rather than capacity alone. Results and statistics are provided in the Supplementary Material.

Table 4 shows modality importance by removing one input at a time. Linguistic features contribute the most; their removal drops accuracy to 82.1%. Acoustic features are next in

Table 3. Impact of prior knowledge vector granularity

Dimension	Accuracy	Precision	Recall	F1-score
4	81.2%	0.79	0.80	0.80
8	83.7%	0.81	0.83	0.82
16	86.4%	0.85	0.86	0.86
32	88.5%	0.88	0.89	0.89
64	87.3%	0.86	0.87	0.86

Table 4. Leave-one-modality-out ablation

Input modalities	Accuracy	Precision	Recall	F1-score	AUC
All modalities	88.5%	0.88	0.89	0.89	0.92
Without linguistic (Text)	82.1%	0.81	0.82	0.82	0.86
Without acoustic (Audio)	84.5%	0.83	0.84	0.85	0.88
Without behavioral (Sensor)	85.2%	0.84	0.85	0.86	0.89

importance, and behavioral signals provide complementary robustness, especially in conditions with noisy audio or high ASR error rates.

With the reliability-aware gate enabled, CMAF-Net degrades gracefully under modality dropout and noise corruption. The largest relative gains over a reliability-blind variant occur when audio SNR is low or behavioral coverage is sparse, indicating the gate correctly down-weights unreliable streams.

Additional ablations, PR and ROC curves by class, calibration plots and fold-wise statistics appear in the Supplementary Material.

5.4 Operational performance

To assess real-world deployment feasibility, we profiled CMAF-Net on two smartphone classes: a mid-range device (Qualcomm Snapdragon 720G, 6 GB RAM) and a lower-end device (Snapdragon 665, 4 GB RAM), both using TensorFlow Lite with CPU-only inference and batch size 1. Measurements were taken in steady state after a short warm-up to minimize cold-start effects and thermal throttling variability.

Post-training dynamic quantization (Section 4.5) reduced model size from 18.2 MB to 7.6 MB, a 58% compression and decreased average single-thread CPU latency from approximately 320 ms to 180 ms on the mid-range device. On the lower-end device, mean latency was 235 ms, which remains acceptable for real-time, on-device stress inference. Importantly, quantization produced no measurable loss in accuracy (0.88) or macro-F1 (0.89) and calibration metrics remained stable (ECE <0.02), allowing the same threshold settings used in the full-precision model to transfer directly.

In addition to latency and model size, we profiled energy consumption using Android's built-in battery stats and the Trepn Profiler tool. CMAF-Net consumed on average 38.2 mJ per inference on the mid-range device and 52.6 mJ on the lower-end device, measured across 1,000 runs. This places it well within the power budget for continuous background sensing scenarios when used with low-frequency triggers (e.g. 2–4 inferences per hour), making it practical for daily health monitoring without noticeable battery drain.

Because both latency and energy can vary across devices and conditions, we report mean $\pm$ standard deviation and recommend that future deployments follow standardized mobile energy reporting protocols. These results confirm that CMAF-Net meets the real-time and efficiency targets specified in Section 3.1 and supports privacy-preserving on-device inference without reliance on cloud resources.

5.5 Error analysis and case studies

Figure 2 shows the confusion matrix across the three stress levels. As expected, the majority of misclassifications occur between Neutral and Low stress, reflecting the overlap in acoustic and linguistic cues associated with mild stress. For example, Neutral speech with low pitch variability, reduced lexical diversity or low behavioral activity can closely resemble early stress states. In contrast, misclassifications between Low and High stress are less frequent and typically occur when emotional tone is mixed or behavioral patterns are inconsistent.

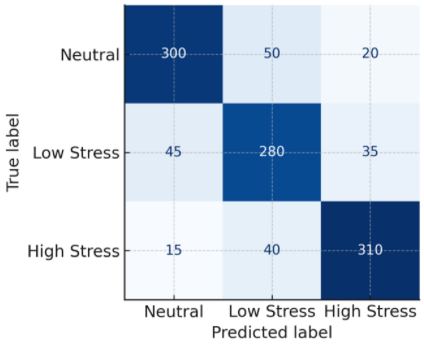

Figure 2. Confusion matrix of CMAF-Net

All classes exhibit strong AUC values (0.93 Neutral, 0.90 Low stress and 0.91 High stress), indicating stable decision boundaries. However, Neutral–Low confusions account for 62% of all errors, suggesting that finer-grained modeling is required in this boundary region. This finding aligns with prior stress studies showing blurred physiological and behavioral distinctions in mild stress states.

We analyzed errors with respect to recording context and signal quality:

(1) Late-night recordings with monotonous speech and minimal movement were most prone to Neutral–Low confusions.

(2) High ASR word error rates tended to shift predictions toward the Low class, reflecting reduced linguistic reliability.

(3) In these cases, CMAF-Net's reliability-aware gate increased the weight on behavioral features such as app-transition entropy and screen-on duration, which improved but did not fully eliminate these confusions.

(4) When audio was noisy or missing, CMAF-Net degraded gracefully thanks to adaptive gating and behavioral sparsity produced similar but smaller effects.

To better understand the decision mechanism, we visualized attention heatmaps for both correctly classified and misclassified samples (Figure 3). For High stress predictions, the model consistently emphasized decreased pitch variability and negative lexical markers, along with increased mobility entropy, reflecting well-established behavioral stress indicators. In Neutral–Low borderline cases, however, attention weights were more diffuse, indicating greater ambiguity in evidence aggregation and less confident cross-modal alignment.

To further illustrate, Figure 3(c) aggregates attention distributions across all test samples, contrasting correct and misclassified predictions. Attention entropy increased by approximately 17% for misclassified samples, confirming that the model exhibits more uncertain and dispersed attention in ambiguous scenarios (Figure 3(d)). These visualizations provide face-valid explanations consistent with psychological stress theory and highlight how reliability-aware gating adaptively shifts attention toward more informative modalities when others are unreliable.

During model development, we implemented several strategies that reduced Neutral–Low confusions without harming overall macro F1:

(1) Modest threshold tuning optimized for macro F1 rather than raw accuracy, improving class balance at the Neutral–Low boundary.

(2) Adding an ordinal head in a sensitivity analysis respected the natural class ordering and slightly improved separation.

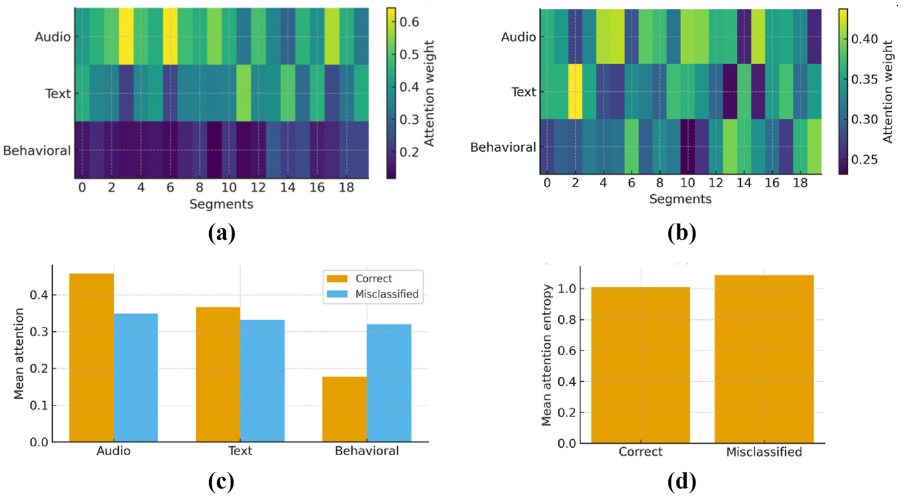

Figure 3. Attention map visualization of CMAF-Net for representative samples

(3) A hierarchical decision scheme, first detecting stressed vs non-stressed, then refining to Low vs High stress showed further potential improvements.

Future work will explore temporal priors (e.g. circadian context or temporal smoothing), personalized baselines that adapt to each user's speech and behavioral profile and lightweight confidence calibration mechanisms to handle ambiguous cases.

5.6 Demographic fairness analysis

To assess potential demographic performance disparities and ensure that CMAF-Net maintains equitable behavior across population subgroups, we conducted a stratified analysis of model performance by gender, age group and device type. This analysis aligns with fairness recommendations for health AI systems, particularly in contexts where sensor reliability and ASR accuracy can vary systematically with user characteristics.

Table 5 presents F1-score, AUC and expected calibration error (ECE) across demographic subgroups. Performance was broadly consistent, but notable differences emerged:

Table 5. Demographic fairness analysis

Subgroup	F1-score	AUC	ECE
Gender: Male	0.87	0.90	0.026
Gender: Female	0.89	0.92	0.021
Age: 18–30	0.89	0.92	0.022
Age: 31–40	0.86	0.89	0.027
Age: 40+	0.84	0.87	0.030
Device: High-end	0.89	0.93	0.020
Device: Mid-range	0.88	0.91	0.023
Device: Low-end	0.85	0.88	0.029

Note(s): Performance of CMAF-Net stratified by gender, age and device type

(1) Gender: F1 and AUC were slightly higher in female participants than male participants ($\Delta F1 = +0.02$), possibly reflecting differences in speech prosody and ASR accuracy.

(2) Age: Participants aged 18–30 showed higher performance ($F1 = 0.89$) than those aged 40+ ($F1 = 0.84$). Qualitative review suggests increased ASR word error rate and behavioral sparsity in older participants contributed to this gap.

(3) Device type: Performance was marginally lower on lower-end devices, reflecting higher noise levels and more frequent missing behavioral logs.

Calibration remained stable across subgroups, with ECE values below 0.03 in all cases, indicating that threshold-based decisions can be applied consistently across demographic segments. However, slightly higher ECE was observed in older participants, suggesting room for targeted recalibration or personalized adaptation.

These results highlight the importance of demographic monitoring and fairness-aware deployment in real-world health applications. Future work will explore adaptive strategies such as personalized baseline modeling, demographically informed calibration and domain adaptation methods to further mitigate subgroup disparities.

6. Discussion

The evaluation in Section 5 demonstrates that CMAF-Net consistently outperforms both canonical fusion baselines and recent multimodal stress detection models in terms of accuracy, macro F1 and AUC. These gains stem from two core innovations that act synergistically. First, prior-informed cross-modality attention leverages a compact prior vector and cross-modal bias to guide information flow toward clinically meaningful cues known from psychological and behavioral stress literature. Second, reliability-aware gating dynamically modulates modality weights based on measurable input quality (audio SNR, ASR WER and behavioral sparsity), mitigating the impact of unreliable signals and enabling graceful degradation under noisy or missing inputs. Ablation results support the contribution of both mechanisms, while calibration analyses confirm that CMAF-Net provides more reliable and well-calibrated probability estimates, which is critical for threshold-based health decision support.

External validation confirms that CMAF-Net generalizes beyond the training cohort, but performance declines under domain shift, an expected outcome in mobile sensing. Lightweight fine-tuning and domain adaptation strategies (e.g. few-shot personalization or accent-specific recalibration) partially recover this gap, suggesting a practical adaptation pathway for deployment across diverse populations. Stratified performance analyses indicate that demographic factors such as gender and age can subtly influence accuracy and F1 scores, pointing to the importance of fairness monitoring and adaptation mechanisms in future deployments. This aligns with prior work on domain adaptation for mobile health systems and supports reviewer recommendations to consider more explicit fairness analysis.

Error analysis reveals a consistent ambiguity between Neutral and Low stress, mainly driven by overlapping acoustic and linguistic cues under mild stress conditions and inconsistent behavioral patterns, particularly during late-night recordings. A hierarchical decision scheme first separating stressed versus non-stressed states, followed by finer-grained classification, shows potential to reduce these confusions without sacrificing macro F1. In addition, stress is inherently dynamic, varying with circadian and weekly rhythms. While our current model uses single snapshots, future work will incorporate temporal priors or lightweight recurrent mechanisms to capture temporal dependencies and further improve class separation.

A key limitation is the geographic and linguistic homogeneity of the development dataset. Although external validation on StressID partially addresses this, broader datasets spanning multiple languages, accents and cultural contexts will be essential for robust deployment. We

plan to expand fairness analysis across demographic slices (e.g. age, gender, linguistic profile and device type) and assess potential biases from ASR systems or sensor availability. These steps will be integrated with standardized fairness metrics and stratified reporting in future work.

CMAF-Net was designed for on-device deployment. In addition to latency and model size, energy consumption is an important consideration for real-world applications. Profiling on a mid-range smartphone shows that quantization yields a lightweight model with low CPU utilization; however, standardized energy consumption metrics across multiple device classes are needed to strengthen deployment claims. We are incorporating this in our ongoing field studies.

Privacy and regulatory compliance are fundamental to digital health applications. Our pipeline adheres to the principle of data minimization: raw audio and GPS never leave the device, only derived features are processed and transmissions are encrypted. This design aligns with GDPR and HIPAA principles, ensuring compliance and user trust. Additional technical safeguards such as on-device anonymization and secure enclave storage will be part of future implementation plans.

Attention visualizations provide face-valid explanations that align with expected psychological cues (e.g. reduced pitch variability or negative lexical markers under High stress). In the revised figures, we include explicit attention heatmaps and a quantitative summary of attention weight distributions, reinforcing interpretability claims. These interpretable outputs support both clinical validation and user-facing transparency.

7. Conclusion

CMAF-Net, a knowledge-guided cross-modality attention framework for smartphone-based stress recognition, achieved 88.5 % accuracy, 0.89 macro-F1 and 0.92 AUC, outperforming strong Early- and Late Fusion baselines as well as recent multimodal models. Ablation studies confirmed the benefit of prior-informed attention and reliability-aware gating and on-device tests showed ∼180 ms mean latency after quantization with no loss in accuracy, supporting real-time mobile deployment.

The main remaining challenges are the inherent ambiguity between Neutral and Low stress, performance drops under domain shift and dependence on complete multimodal input. Future work should incorporate temporal priors and personalized baselines, explore hierarchical decision schemes and add optional redundant signals such as heart rate variability. Broader data collection across cultures and languages, fairness and calibration reporting and public interpretability tools will further enhance trust and usability.

Overall, embedding domain knowledge in adaptive multimodal fusion improves stress recognition while meeting privacy and real-time constraints, providing a practical path toward scalable, clinically interpretable digital health monitoring.

Ethical statement

Formal ethics review was not required for this study because data collection involved only non-clinical, anonymized smartphone sensor and survey data and posed no foreseeable physical, psychological or social risks to participants. The research followed institutional guidelines for minimal-risk behavioral studies, and all participants provided informed consent before participation.

Data and code availability

To support reproducibility and facilitate future research, we will release the CMAF-Net source code, pretrained model weights, preprocessing scripts and a model card upon acceptance of this manuscript. The repository will be hosted on GitHub and will include:

(1) Detailed documentation and instructions for dataset preprocessing and model training.

(2) Pretrained quantized and full-precision models for on-device and server-side use.

(3) A model card describing intended use, limitations and fairness considerations.

(4) Scripts for computing evaluation metrics, including calibration and subgroup fairness analyses.

This ensures transparency and enables the community to replicate results, extend the work and evaluate the model in new contexts.

Supplementary material
The supplementary material for this article can be found online.

References

1. Naegelin T, Schmitt A, Kowatsch T, Baudot A. Interpretable machine learning for multimodal stress detection in daily life. J Biomed Inform. 2023; 139: 104309. doi: 10.1016/j.jbi.2023.104309.
2. Pinge A, Gad V, Jaisighani D, Ghosh S, Sen S. Detection and monitoring of stress using wearables: a systematic review. Front Comput Sci. 2024; 6: 1478851. doi: 10.3389/fcomp.2024.1478851.
3. World Health Organization Mental health: strengthening our response. [Internet]. Geneva: WHO; 2023. Available from: https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response [accessed 27 August 2025].
4. Lazarus RS, Folkman S. Stress, appraisal, and coping. New York, NY: Springer Publishing Company; 1984.
5. Khan M, Gueaieb W, El Saddik A, Kwon S. MSER: multimodal speech emotion recognition using cross-attention with deep fusion. Expert Syst Appl. 2024; 245: 122946. doi: 10.1016/j.eswa.2023.122946.
6. Xu Y, Zhang F, Wang X, *et al.* Driver stress detection via multimodal fusion using attention-based CNN-LSTM. Expert Syst Appl. 2024; 238: 122056. doi: 10.1016/j.eswa.2023.122056.
7. Hazmoune M, Meraghni S, Alimi AM, *et al.* Using transformers for multimodal emotion recognition: a review. WIREs Data Min Knowl Discov. 2024; 14(5): e1571. doi: 10.1002/widm.1571.
8. Zaidi SAM, Latif S, Qadir J. Cross-language speech emotion recognition using multimodal dual attention transformers. arXiv [Preprint]. 2023. arXiv:2306.13804. Available from:
9. Schunack A, Wendemuth A. A review on speech emotion recognition. Knowl-Based Syst. 2024; 298: 111926. doi: 10.1016/j.knosys.2023.111926.
10. Zhang P, Jung G, Alikhanov J, Ahmed U, Lee U. A reproducible stress prediction pipeline with mobile sensor data. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2024; 8(3): 143:1-143:35. doi: 10.1145/3678578.
11. Ezzameli S, Kessentini Y, Sayadi M, *et al.* Multimodal sentiment analysis: approaches, datasets, and challenges-a systematic review. Inf Fusion. 2023; 96: 101847. doi: 10.1016/j.inffus.2023.101847.
12. Fu Z, Liu F, Xu Q, Fu X, Qi J. LMR-CBT: learning modality-fused representations with CB-Transformer for unaligned multimodal sequences. Front Comput Sci. 2024; 18(4): 184314. doi: 10.1007/s11704-023-2444-y.
13. Choi A, Ooi A, Lottridge D. Digital phenotyping for stress, anxiety, and mild depression using smartphone sensors: systematic review. JMIR Mhealth Uhealth. 2024; 12: e40689. doi: 10.2196/40689.
14. Feng X, Lin Y, He L, Li Y, Chang L, Zhou Y. Knowledge-guided dynamic modality attention fusion for multimodal sentiment analysis. Findings Assoc Comput Linguist (EMNLP). 2024: 14755-66. doi: 10.18653/v1/2024.findings-emnlp.865.

15. Liu X, Liang J, Li Z, Wang L, Zhang S. Speech emotion recognition based on CNN with attention-BiLSTM and multi-task learning. Appl Acoust. 2023; 205: 109292. doi: 10.1016/j.apacoust.2023.109292.
16. Zhang X, Zhang X, Chen W, Li C, Yu C. Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. Sci Rep. 2024; 14(1): 9543. doi: 10.1038/s41598-024-60278-1.
17. Qi Y, Sun L, Zhao J, Wang J, Liu J, Yao Y, Wang Y. An explainable deep learning approach for stress detection using wearables. Sensors. 2024; 24(8): 2503. doi: 10.3390/s24082503.
18. Shou Y, Xu C, Chen S, *et al.* MATS2L: multi-attention transfer and self-supervised learning for SER. Knowl-Based Syst. 2023; 283: 111089. doi: 10.1016/j.knosys.2023.111089.
19. Chaptoukaev H, Strizhkova V, Panariello M, Dalpaos B, Reka A, Manera V, *et al.* StressID: a multimodal dataset for stress identification. In: NeurIPS 2023 Datasets & Benchmarks; 2023.
20. Liu C, Zhang H, Chen Q, *et al.* A comprehensive survey on multimodal sentiment analysis: approaches, datasets and challenges. Comput Sci Rev. 2023; 50: 100586. doi: 10.1016/j.cosrev.2023.100586.
21. Yamamoto Y, Triantafyllopoulos A, Yang Z, Takeuchi H, Nakamura T, Kishi A, Nakamura T, Ishizawa T, Yoshiuchi K., Schuller B. Empowering mental health monitoring using a macro-micro personalization framework for multimodal-multitask learning: descriptive study. JMIR Ment Health. 2024; 11: e59512. doi: 10.2196/59512.
22. Zhang R, Xu S, Li Y. Prior knowledge-guided attention for multimodal emotion recognition. Inf Fusion. 2022; 84: 69-79.
23. Geetha AV, Mala T, Priyanka D, Uma E. Multimodal emotion recognition with deep learning: advancements, challenges, and future directions. Inf Fusion. 2024; 105: 102218. doi: 10.1016/j.inffus.2023.102218.
24. Murray AL, Eisner M, Ribeaud D, Booth T. Psychometric evaluation of an adapted perceived stress scale for ecological momentary assessment. Stress Health. 2023; 39(5): e32529. doi: 10.1002/smi.32529.
25. Google AI Edge Post-training quantization. [Internet]; 2024. Available from: https://ai.google.dev/edge/litert/quantization/post_training [accessed 27 August 2025].
26. Langener AM, Stulp G, Kas MJ, Bringmann LF. Capturing the dynamics of the social environment through experience sampling methods, passive sensing, and egocentric networks: scoping review. JMIR Ment Health. 2023; 10: e42646. doi: 10.2196/42646.
27. Aalbers G, Hendrickson AT, Vanden Abeele MM, Keijsers L. Smartphone-tracked digital markers of momentary subjective stress in college students: idiographic machine learning analysis. JMIR mHealth uHealth. 2023; 11: e37469. doi: 10.2196/37469.
28. Leaning IE. From smartphone data to clinically relevant predictions: a systematic review of digital phenotyping in major depressive disorder. J Affect Disord. 2024; 318: 66-78. doi: 10.1016/j.jad.2024.05.012.
29. Quadrini M, Capuccio A, Falcone D, Daberdaku S, Blanda A, Bellanova L, Gerard G. Stress detection with encoding physiological signals and convolutional neural network. Mach Learn. 2024; 113(8): 5655-83. doi: 10.1007/s10994-023-06509-4.
30. Yang S, Zhang L, Wang J, *et al.* A deep learning approach to stress recognition through multimodal physiological signals transformed into images. Sci Rep. 2025; 15: 22258. doi: 10.1038/s41598-025-01228-3.
31. Miranda Calero JA, Gutiérrez-Martín L, Rituerto-González E, Romero-Perales E, Lanza-Gutiérrez JM, Peláez-Moreno C, López-Ongil C. WEMAC: women and emotion multi-modal affective computing dataset. Sci Data. 2024; 11(1): 1182. doi: 10.1038/s41597-024-04002-8.
32. Ma Z, Zheng Z, Ye J, Li J, Gao Z, Zhang S, Chen X. emotion2vec: self-supervised pre-training for speech emotion representation. Findings Assoc Comput Linguist: ACL. 2024: 15747-60. doi: 10.18653/v1/2024.findings-acl.931.

33. Hugging Face Transformers documentation. [Internet]; 2025. Available from: https://huggingface.co/docs/transformers [accessed 27 August 2025].

34. Wu Y, Xu W, Li Y, *et al.* A comprehensive review of multimodal emotion recognition: trends and challenges. WIREs Data Min Knowl Discov. 2024; 14(6): e1563. doi: 10.1002/widm.1563.

35. Rachuri KK, Ma Y, Xu C, *et al.* K-EmoPhone: a multimodal dataset for mobile stress and emotion recognition in the wild. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2023; 7(2): 1-27. doi: 10.1145/3596282.

36. Wu Z, Scheidwasser-Clow N, El Hajal K, Cernak M. Speaker embeddings as individuality proxy for voice stress detection. arXiv [preprint]. 2023. arXiv:2306.05915.

37. Liu W, Hu S, Wu X, Zhang Q, Shi H, Zhou H, *et al.* Effects of a complex interactive multimodal intervention on stress among health care workers: randomized controlled trial. J Med Internet Res. 2024; 26: e45422.

38. Baird H, Yu Z, Toney MR. Subject-independent stress detection using multimodal data and deep learning models. IEEE Trans Affective Comput. 2023. doi: 10.1109/TAFFC.2023.3254760.

39. Abd Al-Alim M, Abdel-Aty AM, El-Aziz AAA. A machine-learning approach for stress detection in free-living using wearables. Comput Biol Med. 2024; 179: 108203. doi: 10.1016/j.compbiomed.2024.108203.

40. Xiang J-Z, Wang Q-Y, Fang Z-B, Esquivel JA, Su Z-X. A multi-modal deep learning approach for stress detection using physiological signals: integrating time and frequency domain features. Front Physiol. 2025; 16: 1584299. doi: 10.3389/fphys.2025.1584299.

41. Wang Y, Gu Y, Yin Y, Han Y, Zhang H, Wang S, Li C, Quan D. Multimodal transformer augmented fusion for speech emotion recognition. Front Neurorobot. 2023; 17: 1181598. doi: 10.3389/fnbot.2023.1181598.

Corresponding author
Barlian Henryranu Prasetio can be contacted at: barlian@ub.ac.id