State of the Art of Multimodal Mobile Stress Detection

Evaluating Multimodal Sensor Fusion, Machine Learning Models and

Mobile Feasibility in Stress Detection Systems

Melanie Maier
it251510@ustp-students.at
University of Applied Sciences St. Pölten
St. Pölten, AUSTRIA

ABSTRACT

Over the last few years, mobile and wearable stress detection has evolved rapidly. This is due to maturing of sensor technology, machine learning (ML) and multimodal fusion strategies. Todays fast paced world critically increases stress-induced illnesses, therefor stress detection becomes more and more important. This State-of-the-Art article critically compares 11 recent and influential papers (2022-2025) about multimodal stress detection.

The analysis examines datasets, preprocessing strategies, fusion architecture, machine learning approaches, evaluation protocols and the extent, to which current models are feasible for real-world mobile deployment. Findings reveal strong progress in multimodal fusion and deep learning, but limited attention to on-device constraints, poor cross-subject generalization, and an overreliance on laboratory datasets. Clear research gaps and design recommendations are identified to guide future work toward robust, scalable, and ecologically valid mobile stress-detection systems.

KEYWORDS

Stress Detection, Mobile, Machine Learning, Sensors, Mobile Feasibility

1 Introduction

Stress is a multidimensional physiological and psychological response involving changes in the autonomic nervous system, endocrine activity, metabolism, behavior, cognition and movement. As of today, wearables, like smartwatches, enabled continuous monitoring of the human physiological state (Zhao, Masood, & Ning, 2025). Wearables and mobile sysrems have shown to be promising platforms for continuous stress monitoring due to high user acceptance, broad availability and the increasing fidelity of embedded sensors.

However, mobile stress detection in real-world settings remains challenging. Signals from wearable sensors are easily affected by motion artifacts, environment conditions, sensor placement variability and individual physiological differences. Approaches that rely on a single sensor, such as PPG (detects changes in blood volume using a pulse oximeter) or EDA (measures the electrical conductivity of the skin) generally struggle with robustness under real-world conditions.

Multimodal stress detection attempts to resolve this by combining heterogenous sensor modalities (e.g. PPG, EDA, ACC, Temp). The resulting sensor fusion pipelines can capture complementary aspects of the stress response), making them more reliable and generalizable.

This SOTA analyzes eleven key papers from 2022-2025, comparing their multimodality, ML models, fusion techniques, dataset characteristics and mobile feasibility. The goal is to build a structured understanding of the field's strengths, limitations and future directions.

2 Multimodal Datasets and Data Characteristics

Multimodal Datasets are datasets that contain multiple types of data (modalities) collected from different sources, in this case different sensors. Instead of relying on one kind of information multimodal datasets combine several complementary data types to provide a richer and more robust understanding of whatever is studied. Each analyzed paper was either based on an established or a custom dataset.

2.1 WESAD as the benchmark

5 of the reviewed papers make use of the WESAD dataset. It's a laboratory-based multimodal dataset capturing EDA, ECG, BVP/PPG, accelerometer data, respiration, EMG and skin temperature during a Trier Social Stress Test (TSST) scenario. It provides a broad set of sensor data connected to stress and therefore allows for recognition of stress in individuals using a combination of data inputs.

Despite its popularity, WESAD has key limitations:

- Only 15 subjects
- Heavily lab-controlled
- Limited ecological validity

Nevertheless, its multimodal nature makes it valuable for fusion research.

2.2 Rare real-world datasets

Only **two** studies (Islam & Washington, 2023) (Darwish, et al., 2025) include real-life data. These capture stress under natural conditions (e.g daily life, workload), but suffer from event uncertainty, self-report noise and device variability.

2.3 Emerging multimodal datasets

A major 2025 contribution is EmpathicSchool (Hosseini, et al., 2025)which is a large multimodal dataset. It combines:

- Facial video
- PPG, ECG, EDA
- Accelerometer
- Structured academic stressors

This dataset expands the field toward a richer and visually enhanced multimodality, though it's not yet optimized for mobile devices.

3 Preprocessing and Feature Engineering

Preprocessing and **Feature Engineering** are two essential steps in machine learning, that transform raw data from datasets into a form that machine learning models can understand and learn from effectively. They happen before training the model and often determine whether the model performs poorly or achieves the desired results.

The analyzed papers have shown that most studies follow a consistent preprocessing pipeline:

1. Signal cleaning through band-pass filtering or artifact removal:
2. Normalization:
3. Segmentation (typical window sizes 30-60 seconds)
4. Feature extraction:
    a. Physiological: HRV metrics, EDA peaks, spectral features
    b. Behavioral: ACC-derived activity levels
    c. Visual: CNN-extracted features
    d. Contextual: metadata, event labels

A reoccurring issue is inconsistent preprocessing documentation, which reduces transparency and reproducibility. Only a few papers (e.g. (Md Santo, et al., 2025) (Zhao, Masood, & Ning, 2025)) describe their pipelines in sufficient detail for replication.

4 Multimodal Fusion Strategies

Fusion strategies define how different modalities are combined to yield a stress prediction. Three main paradigms dominate.

4.1 Early Fusion

Early fusion means that features or raw data are chained before feeding them to a model.

Advantages: simple, efficient
Disadvantages: sensitive to missing modalities

Used in:

- TEANet (Md Santo, et al., 2025)
- Image-encoding CNN models (Ghosh, Kim, Ijaz, Singh, & Mahmud, 2022)

4.2 Late Fusion

Late fusion means that each modality is processed independently and then predictions are combined.

Advantages: robust to missing or weak signals
Disadvantages: limited modeling of cross-modality relationships

Used in:

- From lab to real-life (Darwish, et al., 2025)
- EmpathicSchool Dataset (Hosseini, et al., 2025)

4.3 Hybrid / Cross-Modality Fusion

In hybrid or cross-modality fusion, ML models learn relationships between modalities through different methods, e.g.:

- Transformer cross-attention (Oliver & Dakshit, 2025)
- Privileged modality learning (Zhao, Masood, & Ning, 2025)
- Context-aware adaptive fusion (Rashid, Mortlock, & Al Faruque, 2023)

These methods achieve **the best results** but also have the highest computational cost, making **mobile deployment difficult** without compression techniques.

5 Machine Learning Model Landscape

5.1 Classical ML

Classical machine learning models are for example Random Forests, SVMs and logistic regression. They:

- Have low computational cost
- Depend on hand-crafted features and
- Are less effective on high-dimensional multimodal signals

An example for such a ML model would be the Global HRV + RF model (Dahal, Bogue-Jimenez, & Doblas, 2023). It demonstrates a high mobile feasibility but low multimodal richness.

5.2 Deep Learning Models

The analyzed studies have shown different sorts of deep learning models:

5.2.1 CNNs and CNN-LSTM hybrids. CNN (Convolutional Neural Network) is a deep learning model originally developed for images but widely used for signal processing. CNN-LSTM combines that with a Long Short-Term Memory network for modeling long-term time dependencies. That means that a CNN

captures **instant patterns** and LSTM captures **temporal evolution** across windows.

Strength: automatic feature extraction
Weakness: mobile computational limitations

5.2.2 Transformers. Transformers are a deep learning architecture designed to handle sequences of data (like language, signals, video, sensor time series). They learn global structure while CMMs learn local features and LSTMs learn sequential dependencies. They outperform RNNs and CNNs in cross-modality modeling (Oliver & Dakshit, 2025). They are based on self attention, a mechanism that allows the model to look at any part of the input at any time, and parallel processing, which lets them process all time steps simultaneously, not one-by-one. This makes them extremely fast on GPUs and very good for long signals, multimodal fusion, cross-modal attention (Oliver & Dakshit, 2025) and complex temporal patterns. However, they have a high memory footprint, high inference latency and therefore are unsuitable for on-device execution without quantization or pruning.

5.2.3 Self-Supervised Learning. Self-Supervised Learning is an approach in which the model derives its own learning objectives from the data in order to learn useful representations, without relying on costly human-labeled data. It is particularly powerful when large amounts of unlabeled data are available, as is often the case with physiological or sensor-based stress data. Islam and Washington show that SSL pretraining significantly improves person-specific stress detection, reducing training data requirements (Islam & Washington, 2023).

5.2.4 Autoencoders. Autoencoders are a class of neural networks designed to learn efficient representations of data, typically for purposes like dimensionality reduction, feature learning, or denoising. They are **self-supervised** in nature because they use the input data itself as the target for training, meaning no external labels are required. TEANet (Md Santo, et al., 2025) uses a transpose-enhanced autoencoder for feature compression, promising for low-resource mobile inference.

6 Mobile Feasibility Analysis

Mobile Feasibility refers to whether a machine-learning model and sensor setup, a stress-detection system, can realistically run on a mobile device such as:

- A smartphone
- A smartwatch
- A fitness tracker
- An embedded IoT health device

A crucial dimension often missing in published research is evaluation on actual mobile or wearable hardware.

From the papers analyzed:

- Only **3** report any on-device considerations
- **None** provide full mobile benchmarking such as latency, battery and resource usage

Paper / Study	**Sensors Used for Training**	**Sensors Required at Deployment (Inference)**	**Comment**
PULSE (Zhao, Masood, & Ning, 2025)	EDA, PPG/BVP, ECG, ACC, Temperature	PPG + ACC only (EDA used only during training)	Privileged knowledge transfer enables strong sensor reduction
TEANet (Md Santo, et al., 2025)	PPG/BVP, EDA, ACC	PPG/BVP + ACC	EDA used for reconstruction during training; not required at inference
Cross-Modality Transformer (Oliver & Dakshit, 2025)	ECG, EDA, EMG, RESP, TEMP, ACC	All modalities required	No sensor reduction, computationally heavy
Individualized SSL (Islam & Washington, 2023)	HRV (ECG/PPG), EDA, ACC	HRV + ACC	EDA optional; personalization reduces training burden
SELF-CARE (Rashid, Mortlock, & Al Faruque, 2023)	PPG, EDA, ACC, contextual signals (GPS, phone logs)	PPG + ACC + contextual signals	EDA helpful but optional, high sensor complexity
From Lab to Real-Life (Darwish, et al., 2025)	PPG, EDA, ACC	PPG + ACC	Classical ML models, mobile-feasible
Image-Encoding CNN (Ghosh, Kim, Ijaz, Singh, & Mahmud, 2022)	PPG, EDA, ECG	PPG + EDA + ECG	No reduction, image-encoding pipeline requires all modalities
Global HRV + RF (Dahal, Bogue-Jimenez, & Doblas, 2023)	ECG-derived HRV features	ECG or PPG HRV only	Only true single-sensor solution
EmpathicSchool (Hosseini, et al., 2025)	Facial video, EDA, ECG, PPG, ACC	Video + physiological signals	Very sensor-intensive, not suitable for mobile deployment
Stressor Type Matters (Prajod, Mahesh, & André, 2024)	ECG, EDA, ACC, RESP	All modalities required	Focus on generalization, no deployment optimization

Table 1: Training and Deployment Sensors

6.1 Critical barriers

The analyzed studies have shown that there are the following critical barriers for the mobile feasibility of the presented stress-detection systems:

1. Large model sizes (Transformers, CNN-LSTM hybrids)
2. High inference latency for multimodal pipelines
3. Sensor synchronization issues in mobile settings
4. Energy consumption rarely measured
5. Multimodal dropout (missing modalities in real life)

6.2 Promising mobile-oriented techniques

There are still some promising mobile-oriented techniques:

1. Model compression (quantization, pruning)
2. Teacher-student learning (Zhao, Masood, & Ning, 2025)
3. Lightweight multimodal autoencoders (Md Santo, et al., 2025)
4. Contextual gating (Rashid, Mortlock, & Al Faruque, 2023)

These approaches are promising but still underexplored.

7 Key Comparative Insights

Across the reviewed studies, five clear patterns emerge:

1. **Multimodal fusion consistently outperforms unimodal approaches**, but its computational cost is restrictive for mobile implementation
2. **Cross-subject generalization remains weak**, person-specific models still perform better
3. **Transformer architectures lead in accuracy**, but remain unsuitable for real-time mobile inference without compression
4. **Real-world datasets are severely lacking**, leading to limiting ecological validity
5. **Mobile feasibility is the largest research gap**, almost entirely unaddressed in current literature

8 Research Gaps and Future Directions

1. **Multimodal real-world datasets.** The field urgently needs datasets collected under natural conditions, with motion artifacts, lighting changes and real-world stressors to test stress-detection systems under real conditions.
2. **Standardization of preprocessing pipelines.** Current inconsistency makes cross-paper comparisons unreliable.
3. **Explicit modeling of modality dropout.** In mobile contexts, sensors fail frequently.
4. **Energy-aware model design.** Most published architectures are unfitting for wearables.
5. **Fairness and personalization.** Little work exists on how stress-detection performance varies across gender, age, or physiology.
6. **Cross-dataset generalizability.** Transfer learning and domain adaptation are required for practical real-world deployment.

9 Conclusion

Multimodal mobile stress detection is progressing rapidly, particularly in fusion architectures and learning models. However, the field remains dominated by laboratory datasets and computationally expensive models. Practical mobile feasibility, real-world robustness, subject variability and missing-modality sensitivity are insufficiently addressed.

Future research must integrate lightweight multimodal models, mobile-optimized architectures and real-world datasets to build usable, scalable and ethically sound stress-detection systems.

Study	Dataset	Modalities Used	Fusion Type	ML Model	Metrics	Mobile Feasibility	Key Limitation
PULSE (Zhao, Masood, & Ning, 2025)	WESAD (and derivatives)	EDA, ECG, BVP/PPG, ACC, Temperature	Privileged modality (teacher-student) fusion	Deep Learning (teacher-student privileged knowledge transfer)	Accuracy, F1-score	Medium (model compression intended; not fully evaluated on-device)	Small lab dataset; EDA required during training; limited real-world evaluation
TEANet (Md Santo, et al., 2025)	WESAD	BVP/PPG, EDA, ACC	Early fusion (feature-level); autoencoder compression	Transpose-enhanced Autoencoder (DL)	Accuracy, F1-score, Kappa	Medium (feature compression promising; no full mobile benchmark)	Limited reporting on multimodality and energy/latency metrics
Cross-Modality Investigation (Oliver & Dakshit, 2025)	WESAD	ECG, EDA, EMG, RESP, TEMP, ACC	Attention-based cross-modality fusion (Transformer)	Transformer (cross-modal)	Accuracy, F1-score	Low (high compute & memory; not mobile-friendly)	High computational cost; lacks on-device evaluation
Individualized SSL (Islam & Washington, 2023)	Custom wearable + phone (real-world)	HRV (from PPG/ECG), EDA, ACC	Late fusion (separate encoders, combined features)	Self-Supervised Learning encoder + classifier	F1-score, ROC-AUC	Medium (personalized models reduce data needs; mobile deployment possible but not fully demonstrated)	Less generalizable across users; requires SSL pretraining

SELF-CARE (Rashid, Mortlock, & Al Faruque, 2023)	Custom context-aware dataset	PPG, EDA, ACC, contextual sensors (phone logs/GPS)	Hybrid/adaptive fusion (context-aware gating)	DNN + context module (hybrid)	Accuracy, F1-score	Medium (adaptive methods promising; requires many sensors)	High system complexity; many sensors reduce deployability
From Lab to Real-Life (Darwish, et al., 2025)	Wearable study (lab + real-world)	PPG, EDA, ACC	Late fusion (decision-level aggregation)	Classical ML (Random Forest, SVM) and simple ensembles	Accuracy (real-world vs lab comparisons)	High (simpler models validated in real-world; low latency/energy reported)	Decrease in accuracy in wild conditions; device variability
Image-encoding CNN (Prajod, Mahesh, & André, 2024)	WESAD + other datasets	PPG, EDA, ECG (time series encoded as images)	Early fusion via image-encoding (GAF/Recurrence plots)	CNN on image-encoded signals	Accuracy	Low (image encoding + CNN pipeline is compute-heavy)	Complex pipeline; not optimized for on-device inference
Global HRV + RF (Dahal, Bogue-Jimenez, & Doblas, 2023)	Custom HRV dataset (real-world)	HRV only (ECG-derived features)	Single-modality (no fusion)	Random Forest (classical ML)	Accuracy, F1-score	High (lightweight; mobile-ready)	Limited to HRV signal; lacks multimodal robustness
Recent Advances Review (Ghonge, Shukla, Pradeep, & Solanki, 2025)	Multiple (review)	Multiple (physiological, contextual, visual)	Survey of fusion strategies (various)	Survey (various ML/DL approaches)	N/A (review)	N/A (review paper; discusses feasibility conceptually)	Not empirical; synthesizes literature only
EmpathicSchool (Hosseini, et al., 2025)	EmpathicSchool (new multimodal dataset)	Facial video, EDA, ECG, PPG, ACC	Late multimodal fusion (visual + physiological)	CNN (visual) + physiological models; late fusion	Accuracy, F1-score	Low (video processing heavy; not optimized for wearables)	High sensor cost and compute; limited mobile applicability
Stressor Type Matters (Prajod, Mahesh, & André, 2024)	WESAD + multiple datasets (cross-dataset)	ECG, EDA, ACC, RESP (varies)	Analysis of modelling factors; mixed approaches	Mixed (classical ML + DL experiments)	Accuracy, F1-score (cross-dataset evaluation)	Low (study highlights generalization issues; no mobile eval)	Poor cross-dataset generalization; stressor sensitivity

Table 1: Overview of reviewed Studies

REFERENCES

[1] Dahal, K., Bogue-Jimenez, B., & Doblas, A. (2023). *Global Stress Detection Framework Combining a Reduced Set of HRV Features and Random Forest Model.* Sensors, Basel, Switzerland. doi:https://doi.org/10.3390/s23115220

[2] Darwish, B. A., Rehman, S. U., Sadek, I., Salem, N. M., Karee,, G., & Mahmoud, L. N. (2025). *From lab to real-life: A three-stage validation of wearable technology for stress monitoring.* MethodsX. doi:https://doi.org/10.1016/j.mex.2025.103205

[3] Ghonge, M., Shukla, V. K., Pradeep, N., & Solanki, R. K. (2025). *Recent Advances in Multimodal Deep Learning for Stress Prediction: Toward Cycle-Aware and Gender-Sensitive Health Analytics.* doi:https://doi.org/10.1051/epjconf/202534101059

[4] Ghosh, S., Kim, S., Ijaz, M. F., Singh, P. K., & Mahmud, M. (2022). *Classification of Mental Stress from Wearable Physiological Sensors Using Image-Encoding-Based Deep Neural Network.* Biosensors. doi:https://doi.org/10.3390/bios12121153

[5] Hosseini, M., Sohrab, F., Gottumukkala, R., Bhupatiraju, R. T., Katragadda, S., Raitoharju, J., . . . Gabbouj, M. (2025). *A multimodal stress detection dataset with facial expressions and physiological signals.* doi:https://doi.org/10.1038/s41597-025-05812-0

[6] Islam, T., & Washington, P. (2023). *Individualized Stress Mobile Sensing Using Self-Supervised Pre-Training.* Applied Sciences. doi:https://doi.org/10.3390/app132112035

[7] Md Santo, A., Sapnil Sarker , B., Mohammod , A. M., Sumaiya, K., Manish, S., & Chowdhury, M. (2025). *TEANet: A Transpose-Enhanced Autoencoder Network for Wearable Stress Monitoring.* Von https://arxiv.org/abs/2503.12657 abgerufen

[8] Oliver, E., & Dakshit, S. (2025). *Cross-Modality Investigation on WESAD Stress Classification.* Von https://arxiv.org/abs/2502.18733 abgerufen

[9] Prajod, P., Mahesh, B., & André, E. (2024). *Stressor Type Matters! -- Exploring Factors Influencing Cross-Dataset Generalizability of Physiological Stress Detection.* doi:10.48550/arXiv.2405.09563

[10] Rashid, N., Mortlock, T., & Al Faruque, M. A. (2023). *Stress Detection using Context-Aware Sensor Fusion from Wearable Devices.* Von https://arxiv.org/abs/2303.08215 abgerufen

[11] Zhao, Z., Masood, M., & Ning, Y. (2025). *PULSE: Privileged Knowledge Transfer from Electrodermal Activity to Low-Cost Sensors for Stress Monitoring.* CA, USA. doi:10.48550/arXiv.2510.24058