ISSN 2085-4552

Multimodal Wearable-Based Stress Detection Using Machine Learning: A Systematic Review of Validation Protocols and Generalization Gaps (2021 – 2025)

Pannavira[1], Aditiya Hermawan[2]
[1]Informatics Engineering Department, Science and Technology, Buddhi Dharma University, Tangerang, Indonesia
[1]pannavira@ubd.ac.id, [2] aditiya.hermawan@ubd.ac.id

Accepted 30 December 2025
Approved 08 January 2026

***Abstract*— Stress is a major determinant of mental health and productivity, motivating growing interest in continuous and unobtrusive stress detection using wearable sensors and machine-learning (ML) techniques. This study presents a Systematic Literature Review (SLR) of 19 peer-reviewed articles published between 2021 and 2025, selected from an initial pool of 36 studies using structured inclusion and exclusion criteria. A combined quantitative and qualitative synthesis was conducted to analyze five key dimensions: sensing modalities, ML/DL algorithms, datasets, validation protocols, and deployment-related feasibility. The review identifies dominant methodological trends rather than definitive rankings. Multimodal physiological sensing—most commonly combining photoplethysmography (PPG), electrodermal activity (EDA), and accelerometer data—together with hybrid deep-learning architectures such as CNN–LSTM, is frequently associated with high reported performance on benchmark datasets. However, the analysis also reveals a pronounced lab-to-field gap. Most studies rely on intra-subject or k-fold cross-validation, while subject-independent evaluation using Leave-One-Subject-Out (LOSO) remains rarely adopted, limiting claims of real-world generalizability. In addition, fewer than 15% of the reviewed studies explicitly consider practical deployment constraints, including computational efficiency, power consumption, and data privacy. The primary contribution of this review lies in systematically quantifying the impact of validation practices and deployment considerations on reported performance. The findings highlight that, despite promising accuracy, current stress-detection models remain insufficiently validated for real-world use and point toward the need for generalizable, lightweight, and privacy-aware wearable stress-detection systems.**

Index Terms*— deep learning; machine learning; multimodal fusion; physiological sensing; stress detection; wearable computing.

I. INTRODUCTION

Stress is increasingly recognized as a global health issue affecting individual well-being and organizational productivity. Conventional detection methods, such as clinical interviews, are subjective and episodic; in contrast, the advent of wearable sensors and machine learning (ML) has enabled more objective and continuous stress monitoring [1]. Over the last decade, researchers have explored multiple physiological modalities such as heart-rate variability from *ECG/PPG*, electrodermal responses (*EDA/GSR*), brain activity (*EEG*), respiration, and facial or speech cues. These signals, when analyzed by ML or deep-learning (DL) algorithms, can classify stress with notable accuracy. This progress is foundational to the field, offering a promising pathway toward objective and continuous biomarkers for mental health [2].

Despite this progress, a critical divergence exists between laboratory results and real-world applicability. Recent studies employing Deep Learning (DL) architectures, such as CNN-LSTM hybrids, frequently report accuracies exceeding 95% on benchmark datasets [1]. However, the reliability of these results is often constrained by evaluation methodology. Several studies rely on intra-subject validation (e.g., k-fold cross-validation), which can inflate performance by mixing data from the same individuals across training and testing sets [3]. Prior work suggests that subject-independent evaluation protocols, such as Leave-One-Subject-Out (LOSO) or cross-dataset validation, provide a more realistic assessment of generalization to unseen users, yet these approaches remain relatively uncommon [4]. In addition, practical deployment introduces barriers related to societal feasibility: models are often too computationally heavy for wearable devices, leading to rapid battery drain, while sensor

comfort dictates user adherence for continuous monitoring [5][6].

Beyond these technical hurdles, a critical gap exists in the insufficient consideration of privacy and ethics. Physiological data is inherently sensitive, and continuous collection raises significant user concerns regarding data ownership, misuse, and surveillance. Integrating these systems into daily life requires frameworks that ensure user trust, such as on-device inference (Edge AI) or Federated Learning, yet these aspects are frequently overlooked in research focused purely on accuracy.

Several prior surveys have reviewed stress detection using physiological signals and machine learning, primarily focusing on traditional ML approaches and studies published before 2020 (e.g., Can et al [7]; Panicker & Gayathri [8]). While these works establish important foundations, they do not systematically address recent developments in deep learning–based multimodal models, subject-independent validation practices such as LOSO, or deployment-oriented constraints such as computational efficiency and data privacy.

In contrast, this study explicitly addresses these gaps by systematically reviewing recent (2021–2025) multimodal ML/DL studies, quantifying validation practices, and synthesizing technical performance with deployment-oriented considerations. This positioning differentiates the present SLR from prior reviews that primarily emphasize algorithmic accuracy without assessing real-world readiness.

This review systematically synthesizes published evidence to address these multifaceted gaps. We answer four research questions (RQs) designed to map the state of the art (SOTA): (RQ1) effective sensing modalities, (RQ2) reliable ML/DL algorithms, (RQ3) common validation protocols, and (RQ4) specific barriers to deployment regarding computational efficiency and data privacy. By unifying quantitative and qualitative insights, this SLR aims to guide future research toward stress-detection systems that are not only accurate but also scalable, generalizable, and ethically compliant.

The remainder of this paper is organized as follows: Section II presents theoretical foundations; Section III details the SLR methodology; Section IV discusses results and trends; and Section V concludes with research gaps and future directions.

II. METHOD

This section presents the theoretical foundations that guided our literature synthesis and the conceptual analysis framework used to extract and interpret findings from the reviewed studies.

A. Conceptual analysis framework

To ensure that the review is theory-driven rather than descriptive, we organize the theoretical discussion around four interrelated dimensions that form the analytical lens of this SLR: (1) sensing modalities, (2) modeling and representation learning, (3) validation and generalization, and (4) deployment constraints (computational efficiency and data privacy). These dimensions directly map to our research questions and the data extraction fields used in the review. Concretely, the review extracts and synthesizes evidence about which modalities are used and why (RQ1), which algorithmic paradigms prevail and how features are represented (RQ2), which validation protocols are adopted and how they affect generalization (RQ3), and which deployment-oriented considerations (e.g., on-device inference, federated learning, power/latency metrics) are addressed (RQ4).

B. Stress and Physiological Signals

Stress triggers the Autonomic Nervous System (ANS), disrupting the balance between the sympathetic ("fight or flight") and parasympathetic ("rest and digest") branches. Multimodal sensing is theoretically grounded on this systemic response [9]. Though it lacks temporal precision for short-term events, electrodermal activity (EDA), which directly reflects sympathetic arousal via sweat gland activation, is generally regarded as the most reliable indication of emotional stress [1]. At the same time, the intricate interaction between sympathetic and parasympathetic activity is measured by Heart Rate Variability (HRV), which is obtained from ECG or PPG. PPG provides a wearable-friendly substitute for ECG, however it is prone to motion artifacts [2]. ECG is still the clinical gold standard. Beyond these autonomic markers, cortical responses can be captured via EEG, while physical activity that often confounds physiological signals is monitored using accelerometers (ACC). Theoretically, multimodal fusion reduces the uncertainty associated with unimodal sensing by capturing independent stress signals, such as merging the sympathetic strength of EDA with the vagal tone of HRV [10].

C. Feature Engineering to Representation Learning

The transition from classical Machine Learning (ML) to Deep Learning (DL) represents a fundamental shift in how stress features are modeled. Traditional ML approaches relied heavily on Feature Engineering where domain experts manually extracted statistical features for classifiers like SVM or Random Forest. While interpretable, this approach is limited by the quality of handcrafted features and struggles with raw, noisy sensor data [11]. Conversely, modern architectures have shifted toward Deep Representation Learning. In particular, hybrid models such as CNN-LSTM automate feature extraction; Convolutional Neural Networks (CNN) learn spatial or spectral patterns directly from spectrograms, while Long Short-

Term Memory (LSTM) networks model the long-term temporal dependencies essential for physiological time-series analysis [1].

D. The Generalization Gap: Subject Inter-variability

Subject inter-variability is a significant theoretical difficulty in stress detection. Stress-related physiological reactions vary greatly; for example, one person's baseline heart rate may be a sign of extreme stress for another [4]. Theoretically, this has significant ramifications for validation procedures. Conventional validation techniques, such k-fold cross-validation, are predicated on the idea that the data is identically distributed and independent. This assumption is broken, though, when training and testing segments from the same subject are combined, leading the model to learn subject-specific characteristics instead of stress-specific patterns. Leave-One-Subject-Out (LOSO) validation is necessary for theoretical rigor in order to close the lab-to-field gap. In contrast to k-fold, LOSO guarantees that the model is evaluated on completely unknown users, requiring the acquisition of generalized stress features a necessary condition for reliable real-world implementation [2].

E. Constraints Efficiency and Privacy

Beyond accuracy, real-world deployment is theoretically constrained by the trade-off between model complexity and resource availability. Deep Learning models, while accurate, impose high computational costs. Furthermore, the traditional Centralized Learning paradigm where raw data is transmitted to a cloud server violates modern privacy principles. The theoretical alternative is Federated Learning (FL), a distributed optimization paradigm where models are trained locally on devices and only model updates (gradients) are shared [12]. This approach theoretically decouples learning from data centralization addressing the privacy concerns inherent in physiological sensing without compromising the model's ability to learn from population-level data.

III. Result and Discussions

A. Systematic Literature Review

This study adopts a Systematic Literature Review (SLR) approach to synthesize existing evidence on stress detection using Machine Learning. To ensure methodological rigor, transparency, and reproducibility, the review protocol adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [13]. This standard framework ensures that the selection of studies is unbiased and comprehensive, shifting focus from general descriptions to quantitative performance analysis of recent multimodal systems.

B. Research Question

The following table contain the research questions that has been carried out on this paper. Table I showed four questions that is the main focus of this paper.

Table I. Research Question

ID	Research Question	Motivation
RQ1	Which physiological or behavioral modalities are most effective for stress detection?	Identify practical and accurate sensing methods
RQ2	Which ML/DL models demonstrate robust and generalizable performance?	Determine the current state of the art
RQ3	What datasets and validation protocols are commonly used?	Evaluate reproducibility and comparability
RQ4	What societal, ethical, or deployment aspects are considered?	Assess real-world readiness

C. Work Procedure

The work procedure involves conducting a literature search, selecting relevant sources, documenting findings, analyzing the information, and drawing conclusions as visualized in the PRISMA flowchart (Fig. 1), following a structured protocol adapted from established SLR guidelines [14]. This process consists of these main steps, as detailed below.

First, the authors identified key terms relevant to the research topic. The main keywords used in the search were “stress detection,” “machine learning,” “deep learning,” “wearable sensor,” “physiological signal,” and “multimodal fusion.” Boolean combinations were applied across several databases such as IEEE Xplore, ScienceDirect, SpringerLink, and MDPI to ensure a broad coverage of recent studies [14].

Second, the authors determined the origin and source of the literature. Journals were selected from reputable international publishers that focus on artificial intelligence, biomedical engineering, and affective computing [14]. The search was conducted online and limited to peer-reviewed articles published between 2021 and 2025 to capture the most recent advances.

Third, the collected works were filtered according to strict criteria. The initial search yielded 36 papers. After removing 5 duplicates, 31 papers underwent title and abstract screening. We applied the Inclusion and Exclusion Criteria presented in Table II to filter these results. We specifically excluded qualitative surveys and studies lacking quantitative metrics (e.g., Accuracy/F1-Score). This process resulted in a final selection of 19 articles deemed relevant for detailed review.

Table II. Inclusion and Exclusion Criteria

Criterion	Inclusion	Exclusion
Publication Type	Journal Articel	Conference, Thesis, Book Chapters
Timeline	2021-2025	< 2021
Content	Quantitative ML/DL Performance	Qualitative / Theoretical only
Modality	Multimodal	Unimodal

Fourth, to ensure technical soundness, a Quality Assessment was performed on the final 19 articles. We defined four binary quality criteria focusing on reproducibility. Each paper was evaluated as Yes (1) or No (0). Only studies satisfying at least 3 out of 4 criteria were included in the final data synthesis.

TABLE III. QUALITY ASSESSMENT

ID	Assessment Criteria (QA)	Motivation
QA1	Is the dataset clearly described?	Ensures data reproducibility (participants, signals, protocols).
QA2	Are the feature extraction and ML/DL models explicitly defined?	Ensures the technical approach is replicable
QA3	Is the validation methodology clearly stated?	Critical for evaluating generalizability claims.
QA4	Are quantitative performance metrics reported?	Essential for comparative analysis.

Fifth, the extracted data were synthesized using a two-stage approach. First, a descriptive quantitative synthesis was conducted to address RQ1–RQ3 by tabulating key characteristics of the selected studies, including sensing modalities, algorithms, datasets, validation protocols, and reported performance metrics. Descriptive statistics (frequencies and ranges) were used to identify prevailing trends and state-of-the-art approaches. A formal meta-analysis was not performed due to substantial heterogeneity in datasets, experimental protocols, and evaluation metrics. A qualitative thematic synthesis was applied to address RQ4 by analyzing deployment-related considerations such as computational efficiency, validation rigor, and data privacy. This analysis also involved a critical assessment of methodological quality, particularly the use of subject-independent validation and dataset characteristics, to identify key gaps affecting real-world applicability.

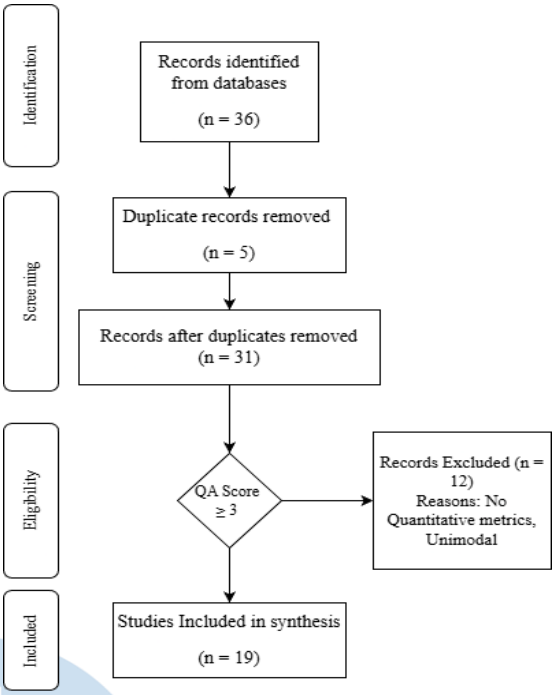

Fig. 1. PRISMA Flowchart of the Literature Selection Process

IV. RESULT AND DISCUSSIONS

A. Overview of the Studies

The systematic selection process resulted in 19 studies published between 2021 and 2025. The analysis reveals a diverse landscape of methodologies, with dataset sizes ranging from small custom cohorts (n=11) to large public benchmarks like WESAD (n=15) and SWELL-KW. Before detailing the performance metrics, it is crucial to note that direct comparison of accuracy across studies requires caution. The heterogeneity in stress-induction protocols (e.g., MIST vs. Driving Simulators) and label granularity (2-class vs. 3-class) means that a higher accuracy score does not always imply a superior model, but may reflect a simpler classification task or a less rigorous validation scheme.

Most of the reviewed works focus on physiological-signal-based stress detection, often combining more than one sensing modality. The most frequently used signals are electrodermal activity (EDA), photoplethysmography (PPG), and electrocardiography (ECG), followed by studies employing electroencephalography (EEG) or accelerometer (ACC) data [15]. These modalities are widely available in commercial wearables, which explains their popularity for daily stress-monitoring research.

In terms of methodology, traditional machine-learning classifiers such as *Support Vector Machine (SVM)*, *Random Forest (RF)*, and *XGBoost* remain common, particularly for smaller or unimodal datasets[15]. However, a clear shift toward deep learning architectures, notably *Convolutional Neural*

Networks (CNN), *Long Short-Term Memory (LSTM)* networks, and hybrid *CNN–LSTM* models, can be observed in the most recent papers. These models tend to achieve the highest accuracy, often above 90 %, especially when multiple signals are fused [16][17].

The validation methods reported vary across studies. The majority rely on k-fold cross-validation, while a smaller group applies Leave-One-Subject-Out (LOSO) or cross-dataset evaluation to test generalization [6].

Table IV provides a concise overview of each study, including its dataset, modality, algorithm, validation method, and the best reported metric.

Table IV provides a concise overview of each study, including its dataset, modality, algorithm, validation method, and the best reported metric.

TABLE IV. SUMMARY OF SELECTED STUDIES

Reference	Dataset	Modality	Algorithm	Validation Method	Best Performance Metric
[2]	Custom (40 participate, *Public Speaking*)	Multimodal (EEG, GSR, PPG)	SVM (RBF), kNN, DT, RF, MLP	Leave-One-Out (LOOCV)	Accuracy 96.25% (2 class)
[1]	ST Change DB, WESAD	EKG (Changed into Spectogram)	Ensemble CNN-LSTM		Accuracy 98.3% (2 class)
[6]	Custom (22 participate, Driving Simulator)	Multimodal (Eye Data, Vehicle, Surrounding)	Attention-based CNN-LSTM	10-fold cross-validation	Accuracy 95.5% (3 Class)
[18]	MuSE, OMG-Emotion	Text (transcript) & Acoustic (audio)	MUSER (Transformer/BERT + MLP)	Split Train/Validation/Test (Dataset)	F1-score 0.864 (2 class)
[19]	Custom (34 subject, MIST)	EKG (10 second segment)	CNN + BiLSTM	5-fold cross-validation	Accuracy 86.5% (3 class)
[20]	Custom (20 subject , MIST)	Multimodal (EKG, Sound, Face expression)	Hybrid DL (ResNet50 + I3D w/ TAM)	10-fold cross-validation	Accuracy 85.1% (2 class)
[21]	DEAP, SEED	EEG (converted into Azimuthal Projection Image)	StressNet (Hybrid 2D-CNN + LSTM)	80% Train / 20% Test	Accuracy 97.8% (2 class)
[5]	Custom (90 subject, Office Simulation)	Multimodal (Behavior: Mouse, Keyboard + Physiological: HRV)	LightGBM, SVM, RF	10-fold cross-validation (dengan SMOTE)	F1-score 0.625 (3 class , Stress)
[4]	Custom (11 subject, MBSR)	EEG	Shallow/Deep ConvNet, FBCSP+SVM	LOOCV, Mix-subject, Intra-subject	Accuracy 99.65% (Task: Meditation vs. Rest)
[22]	UBFC-Phys	rPPG (face video)	1D-CNN, LSTM, GRU	80% Train / 10% Validation / 10% Test	Accuracy 95.83% (2 class)
[23]	MultiAffectStress (MAS)	Audio-Visual (Face, Vocal, Sentiment, *Fidgeting*)	Learning-Based Late Fusion (RF)	60% Train / 20% Validation / 20% Test	F1-score 0.85 (2 class)
[3]	Custom (26 Subject, Cortisol label)	Multimodal (EKG, RESP, Electrogastrogram)	Shuffled ECA-Net (1D-CNN + Attention)	5-fold cross-validation (Intra-subject)	Accuracy 91.6% (2 class)
[11]	SWELL-KW	HRV	k-NN, Decision Tree, Logistic Regression	5-fold cross-validation	Accuracy 99.3% (3 class)
[16]	WESAD, SWELL KW, RAVDESS, EMO-DB	Physiological (EKG, EDA) & Audio	GSOA-SHBRNN (VGG-16 + PCA + Bi-RNN)	2/3 Train, 1/3 Test	Accuracy 99.52% (WESAD, 2 class)
[9]	Nursery Dataset (from Hosseini et al., 2022)	Multimodal (ACC, EDA, HR, TEMP)	MMFD-SD (Parallel CNNs Time+Frequency)	80% Train / 20% Test (Stratified Split)	Accuracy 91.00% (3 class)
[24]	WESAD	Multimodal (BVP, EDA, TEMP, ACC, RESP) converted to 2D RGB Image	CNN (Custom Architecture)		F1-score 91.67% (3 class)

[25]	LifeSnaps, PMData	Multimodal (Time Series: HR, Steps + Tabular: Demographics, Context)	Contrastive Pretraining (CLIP-style)	5-fold cross-validation	AUC 81.14% (PMData, 2 class)
[26]	Yonsei Stress Image & Speech Database (custom multimodal stress dataset)	Multimodal Facial images (RGB video frames + facial landmarks) and Speech (log mel-spectrogram)	ResNet-18 backbones, attention mechanisms + Multimodal Neglecting Mask Module (MNMM) for intermediate feature fusion	5-fold cross-validation 3/5 Train, 1/5 Validation 1/5 Test	Specificity: 88.99%
[27]	WESAD	Multimodal (ECG, EDA, EMG, Respiration, Temperature)	CNN + LSTM + Attention mechanism	Train–test split with 90% training and 10% testing	Accuracy: 92.70% (multimodal setting)

B. Modalities and Sensor Trends

Among the various physiological signals, EDA and PPG emerge as the most practical and consistent modalities for real-time, wearable-based detection. Their combination captures both sympathetic nervous system response (EDA) and cardiovascular activity (PPG), providing a comprehensive picture of physiological arousal. This multimodal fusion is shown to be highly effective, achieved 96.25% accuracy by fusing GSR (EDA) and PPG with EEG. This quantitative result supports the qualitative trend that studies integrating multiple modalities, particularly those readily available in wearables, typically outperform those relying on a single signal. [2]

ECG continues to be the reference modality in controlled laboratory environments because of its high sensitivity to subtle changes in heart-rate variability. However, it is less convenient for long-term use due to sensor placement and comfort issues[3]. EEG-based approaches, while powerful for cognitive-stress analysis, face similar challenges related to setup complexity[4].

Overall, the literature points toward wearable-friendly, multimodal sensing, often combining EDA, PPG, and ACC. This configuration balances accuracy, comfort, and cost, making it well suited for practical applications[15]. Figure 2 highlights a clear preference toward EDA and PPG as the dominant physiological modalities in recent stress-detection studies. Researchers have chosen wearable-friendly sensors over clinically accurate but invasive alternatives like ECG or EEG, which is a realistic trade-off. Analytically speaking, this distribution implies that real-world deployability concerns implicitly limit state-of-the-art research, highlighting the significance of multimodal setups that strike a balance between accuracy, comfort, and scalability.

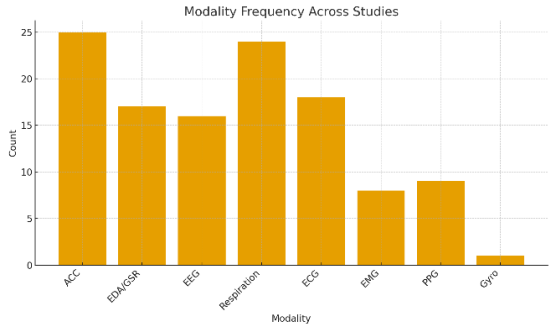

Fig. 2. Frequency of physiological modalities used

C. Validation Strategies and Datasets

A consistent observation is the dominance of k-fold cross-validation for evaluating model accuracy. While suitable for preliminary comparison, this approach often inflates results because training and testing data originate from the same participants [3]. A smaller number of studies adopt LOSO or subject-independent validation, which provides a more realistic assessment of model robustness in unseen subjects[2] [4]

Public datasets such as WESAD and DEAP appear most frequently. WESAD, in particular, serves as the primary benchmark for multimodal wearable stress detection, combining EDA, PPG, and ACC signals. Nevertheless, differences in dataset structure, participant demographics, and labeling criteria make direct comparison between studies difficult [1]. Figure 3 illustrates the distribution of validation strategies, emphasizing the need for broader adoption of cross-subject testing in future research.

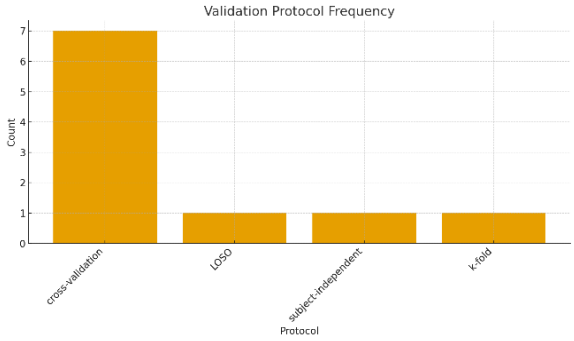

Fig. 3. Validation strategies across studies.

As illustrated in Figure 3, the dominance of k-fold cross-validation indicates that most studies prioritize performance optimization under controlled conditions rather than generalization to unseen users. Our synthesis reveals that the limited adoption of LOSO validation is not merely a methodological choice, but a key contributor to the observed lab-to-field gap. This imbalance underscores the need to reinterpret high reported accuracies with caution, particularly when claims of real-world applicability are made.

D. Reported Performance

The reported performance across all studies generally falls within the 85 %–96 % range, depending on the modality and evaluation protocol used, models cluster around 90 % accuracy or equivalent F1-scores, which is strong for physiological classification tasks [3], [6], [9], [23].

However, results obtained from LOSO or cross-dataset validation are typically 5–10 % lower, underscoring the challenge of generalizing across individuals [4]. The lower performance observed under LOSO or cross-dataset validation does not indicate inferior modeling, but rather reflects a more stringent and realistic learning objective. In intra-subject evaluation, models are exposed to physiological patterns from the same individuals during training and testing, enabling them to implicitly learn subject-specific baselines and signal idiosyncrasies. This can lead to inflated performance that reflects pattern recognition of individuals rather than genuine stress-related physiological responses. In contrast, LOSO validation enforces complete subject separation, requiring models to infer stress from physiological changes that generalize across individuals with

inherently different baselines and response dynamics. Since stress manifests as relative deviations rather than absolute signal values, LOSO-trained models are compelled to capture invariant stress-related features instead of memorizing personal signal patterns. Consequently, although LOSO evaluation yields lower numerical scores, it provides a more meaningful assessment of a model's ability to detect stress rather than merely recognizing individual-specific patterns.

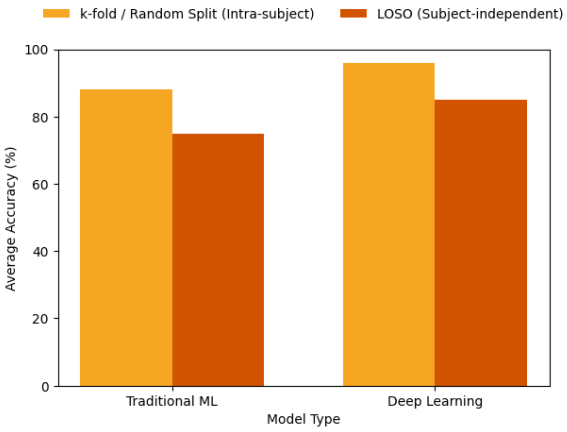

Fig. 4. Impact of validation protocol on reported stress-detection performance.

The figure summarizes average performance trends of traditional machine-learning and deep-learning models under intra-subject (k-fold) and subject-independent (LOSO) evaluation across the reviewed studies. Deep-learning approaches consistently achieve higher scores than traditional ML when trained on multimodal inputs. The combination of CNN for feature extraction and LSTM for temporal modeling remains the most successful design pattern, especially when applied to PPG and EDA data.

However, a direct comparison of these performance metrics is complicated by the significant heterogeneity across study protocols. Our analysis reveals that several factors strongly influence reported outcomes. These include the data labeling methodology (e.g., self-report vs. induced stress protocols like MIST [19], [20] ordriving simulators [6]), the signal processing details such as the length of the time segments used for analysis (e.g., 10-second segments in [19]), and the dataset characteristics, including sample size and participant diversity. For example, models validated on large, public benchmark datasets like WESAD [1][, [16], [24] may offer more generalizable insights than those trained on smaller, custom datasets [6]. These variations underscore the difficulty in establishing a single best model and highlight the critical need for standardized reporting protocols in future research.

E. Machine-Learning and Deep-Learning Approaches

The reviewed papers demonstrate two major methodological generations. Early studies typically extracted handcrafted statistical and frequency-domain features, which were then classified using SVM, RF, or logistic regression. These techniques achieved accuracies in the range of 80–90 %, proving that stress can be inferred reliably from physiological data even with simple models [11]. Moreover, classical ML methods remain attractive in scenarios involving limited data, lower computational budgets, and a need for model interpretability, which is

particularly relevant for clinical or explainable-AI contexts.

Hybrid CNN–LSTM architectures have emerged as the dominant design paradigm in recent studies. From a modeling perspective, this hybridization is well aligned with the nature of physiological stress signals: CNN components act as automated feature extractors that reduce noise and encode local patterns, while LSTM layers model the sequential evolution of these features over time. When applied to multimodal inputs such as PPG, EDA, and ACC, this architecture enables both intra-modal representation learning and temporal fusion, which explains the consistently high reported accuracies between 93 % and 96 % across multiple datasets. Extensions incorporating attention mechanisms further refine this process by dynamically weighting informative signal segments or modalities, while contrastive pre-training and teacher–student knowledge distillation aim to improve robustness and data efficiency [6].

Taken together, the reviewed literature suggests that the choice between classical machine-learning and deep-learning approaches should not be guided by accuracy alone. Instead, it should reflect the intended application context, available data, and deployment constraints. Classical ML models remain suitable as strong baselines or interpretable solutions in low-resource settings, whereas deep-learning architectures represent the current state of the art for high-performance, multimodal stress detection when sufficient data and computational capacity are available.

Table V summarizes the average performance by model type. In general, deep sequential or hybrid models outperform classical methods, though they require more computational resources.

TABLE V PERFORMANCE COMPARISON BASED ON MODEL TYPE

Model	Reference	Dataset	Performance Range (Reported)	Key Notes
ML Classic (SVM, RF, k-NN, GBM)	[2]	Custom (40 participate, *Public Speaking*)	Accuracy: 96.25% (2 class)	SVM (RBF) outperformed kNN, DT, RF, and MLP in feature fusion.
	[5]	Custom (90 subject, Office Simulation)	F1-score: 0.625 (3 class)	LightGBM outperformed SVM/RF. Behavioral data fusion (mouse/keyboard) was better than HRV.
Deep Learning (Hybrid CNN-LSTM / BiLSTM)	[1]	ST Change DB, WESAD	Accuracy 98.3% (2 class)	Time and frequency domain (spectrogram) fusion of ECG data achieves high accuracy.
	[19]	Custom (34 subject, MIST)	Accuracy 86.5% (3 class)	Effective for real-time detection (10-second segments).
	[6]	Custom (22 participate, Driving Simulator)	Accuracy 95.5% (3 class)	Non-physiological multimodal fusion using attention has proven to be highly effective.
Deep Learning (CNN Multimodal Fusion)	[9]	Nursery Dataset	Accuracy 91.00% (3 class)	The Parallel CNN architecture separates Time and Frequency domain features before fusion.
	[3]	Custom (26 Subject, Cortisol label)	Accuracy 91.6% (2 class)	Using "Shuffled ECA-Net" (Attention) for feature fusion. The stress label is validated by Cortisol.
Deep Learning (Transformer / Multi-Task)	[18]	MuSE, OMG-Emotion	F1-score: 0.864 (2 class)	Using Multi-Task Learning (MTL) where emotion recognition becomes an auxiliary task for stress detection.
	[23]	MultiAffectStress (MAS)	F1-score: 0.85 (2 class)	Using Late Fusion (Random Forest) to combine the outputs of several unimodal models (including Wav2Vec 2.0 and DistilBERT).

F. Societal Feasibility and Ethical Considerations

While high numerical accuracy remains an essential benchmark, the true success of a stress-detection model lies in its translation into everyday use. Machine-learning research is beginning to move from laboratory settings toward field deployment, yet the gap between experimental performance and societal applicability remains substantial. Several studies acknowledge that stress recognition is meaningful only when it can operate continuously, comfortably, and ethically within people's daily routines.

1) Feasibility of Deployment

Approximately one-third of the reviewed papers describe some form of prototype or pilot deployment, ranging from wrist-worn sensors to smartphone-based data collection. Wearable-centric designs particularly those relying on PPG and EDA sensors integrated in smartwatches or fitness bands emerge as the most realistic pathway for long-term stress monitoring[15].

These devices already enjoy high consumer adoption and can collect data passively without interrupting normal activity. Studies employing

multimodal fusion (EDA + PPG ± ACC/ECG) demonstrate not only technical robustness but also user comfort, battery efficiency, and signal stability during motion, all of which are prerequisites for sustainable deployment.

Conversely, models relying on EEG or ECG chest straps face usability barriers due to cumbersome electrodes and the need for skin contact. Although these sensors yield rich physiological data, their invasiveness reduces adherence outside clinical environments. A few researchers attempt to overcome these barriers through smart-textile electrodes and dry-sensor patches, indicating a promising hardware direction for future work.

2) Real-Time and Edge-AI Integration

Recent advances highlight the feasibility of running stress-recognition pipelines on resource-constrained devices. Several publications introduce lightweight CNN or LSTM architectures optimized for on-device inference, reporting inference times of less than one second on mobile processors [22]. This shift toward edge-AI brings multiple benefits: it enables immediate feedback for users, lowers network latency, and minimizes dependency on cloud connectivity factors crucial for emergency or occupational-safety contexts. Moreover, edge computing supports energy efficiency by processing only essential features locally and transmitting aggregated indicators instead of raw biosignals.

However, only a small fraction of current literature reports quantitative measurements of power consumption, model size, or latency, parameters that determine practical viability. Future publications should systematically include these metrics alongside accuracy to support reproducibility and engineering optimization.

3) Summary of Feasibility Indicators

This gap in feasibility is most critical regarding privacy and ethics. Physiological signals constitute highly sensitive personal health information, and their continuous collection raises significant user concerns over data misuse and surveillance. This review found that fewer than 15% of studies explicitly address this, often only mentioning basic anonymization. This is insufficient for real-world trust. As requested by modern data-protection laws (e.g., GDPR), the field must shift from cloud-centric processing to privacy-by-design architectures. The solution lies in the resource-efficiency models identified in this review, which enable on-device inference (Edge AI). This approach processes data locally, minimizing data transmission. For models that require continuous improvement, Federated Learning frameworks experimented with by a handful of studies offer a path forward, allowing models to be trained across distributed devices without centralizing raw data, thereby mitigating critical privacy risks.

Furthermore, our analysis highlights that technical accuracy alone is insufficient; the psychological impact and application context are paramount. Continuous stress feedback, if poorly designed, risks amplifying user anxiety rather than mitigating it. Future research must therefore bridge the gap between detection and intervention. This requires integrating psychological frameworks, such as providing Just-In-Time Adaptive Interventions (JITAI) or cognitive-behavioral prompts, transforming passive monitoring into active well-being support. In practical application contexts, such as workplace wellness programs or continuous personal health monitoring, this integration is essential. The goal is not merely to inform a user "you are stressed," but to provide an actionable, empathetic, and private pathway to improved mental resilience.

Beyond privacy and hardware, societal feasibility also involves user perception and behavioral adoption. Continuous stress feedback can empower self-awareness, yet poorly designed feedback loops risk amplifying anxiety. Few studies examine how users interpret or act upon stress predictions. Integrating psychological frameworks, such as just-in-time adaptive interventions or cognitive-behavioral prompts, could transform stress detection from passive monitoring into active well-being support.

To quantify these dimensions, each paper was scored across three observable indicators (a) use of wearable or smartphone sensors, (b) existence of a prototype or real-time system, and (c) mention of privacy or edge computing.

Overall, the evidence reveals a field that is technically sophisticated but socially nascent. To move from promising algorithms to impactful public-health tools, future research must integrate design for usability, transparency, and trust alongside continued advances in model accuracy. The ultimate benchmark for stress-detection research will not only be statistical precision but also its contribution to safer, healthier, and more empathetic human technology interaction.

This review reveals a clear methodological shift from traditional machine-learning pipelines toward deep-learning–based architectures for stress detection. As summarized in Fig. 4, deep-learning models consistently achieve higher average performance than classical approaches under both intra-subject and subject-independent evaluation. However, this advantage is accompanied by increased computational complexity, highlighting a trade-off between accuracy and deployability that must be considered in practical applications.

A key finding of this review is the substantial influence of validation strategy on reported performance. As illustrated in Fig. 4, both traditional ML and deep-learning models exhibit a consistent reduction in performance under subject-independent validation compared to intra-subject evaluation. This

pattern underscores the central role of subject inter-variability in physiological stress detection and confirms that validation methodology is a decisive factor in assessing real-world generalization capability.

Despite the observed performance trends, direct comparison across studies remains inherently limited. The reviewed literature exhibits substantial heterogeneity in datasets, stress-induction protocols, class definitions, signal preprocessing, and validation schemes. Consequently, the synthesized results should be interpreted as indicative methodological trends rather than definitive rankings of model superiority. This limitation reinforces the need for standardized reporting practices to enable more reliable comparison in future reviews.

Taken together, the findings suggest that future stress-detection research should prioritize subject-independent evaluation, multimodal sensing strategies, and deployment-aware model design. Emphasis on LOSO or cross-dataset validation, alongside transparent reporting of computational and privacy-related metrics, is essential to bridge the gap between laboratory performance and real-world applicability.

This SLR also has limitations. Our search was restricted to articles published between 2021 and 2025 to capture the most recent SOTA, which may exclude foundational papers in the field. Furthermore, due to high heterogeneity in datasets, protocols, and metrics, we performed a descriptive and thematic synthesis. A formal statistical meta-analysis was not conducted, which limits the quantitative aggregation of performance across studies.

V. CONCLUSIONS

This Systematic Literature Review analyzed 19 studies and confirmed a clear technical state-of-the-art for stress detection: multimodal sensing (PPG, EDA, ACC) combined with hybrid CNN-LSTM models consistently yields high accuracy. The review provides a structured synthesis of current methodological trends, validation practices, and deployment considerations. The main conclusions of this study are summarized as follows.

A. Main Findings

- Multimodal sensing, particularly combinations of PPG, EDA, and ACC, is the dominant and most practical configuration for wearable-based stress detection.
- Hybrid deep-learning architectures, especially CNN–LSTM models, consistently achieve higher reported performance than traditional machine-learning methods.
- Intra-subject validation (e.g., k-fold cross-validation) remains the most commonly used evaluation protocol, while subject-independent validation methods such as LOSO are still underutilized.
- Performance obtained under subject-independent validation is consistently lower but provides a more realistic estimate of real-world generalization capability.

B. Scientific Contributions

This review makes the following scientific contributions:

- It provides an up-to-date synthesis of multimodal stress-detection studies published between 2021 and 2025, capturing recent advances in deep-learning–based modeling.
- It systematically highlights the impact of validation protocols on reported performance, explicitly quantifying the lab-to-field generalization gap.
- It extends conventional performance-focused reviews by integrating deployment-oriented dimensions, including computational efficiency and data privacy considerations.

C. Research Implications

The findings of this review have several important implications for future research and practice:

- Reported accuracy alone is insufficient to assess model robustness; validation methodology must be considered a primary evaluation factor.
- Deployment feasibility, including model efficiency and privacy-preserving design, should be treated as first-class criteria alongside predictive performance.
- Without standardized validation and reporting practices, cross-study comparison will remain limited and potentially misleading.

D. Further Research Directions

Based on the identified gaps, future research should prioritize:

- The adoption of subject-independent evaluation protocols, such as LOSO or cross-dataset validation, to ensure reliable generalization.
- The development of lightweight and energy-efficient models suitable for on-device inference and edge-AI deployment.
- The integration of privacy-by-design principles, including federated learning and local processing, to address ethical and regulatory concerns.

The connection between stress detection and intervention mechanisms, such as just-in-time adaptive interventions (JITAI), to move from passive monitoring toward actionable mental well-being support

REFERENCES

[1] M. Kang, S. Shin, J. Jung, and Y. T. Kim, "Classification of Mental Stress Using CNN-LSTM Algorithms with Electrocardiogram Signals," *J Healthc Eng*, vol. 2021, 2021, doi: 10.1155/2021/9951905.

[2] A. Arsalan and M. Majid, "Human stress classification during public speaking using physiological signals," *Comput Biol Med*, vol. 133, p. 104377, Jun. 2021, doi: 10.1016/j.compbiomed.2021.104377.

[3] N. Kim, S. Lee, J. Kim, S. Y. Choi, and S. M. Park, "Shuffled ECA-Net for stress detection from multimodal wearable sensor data," *Comput Biol Med*, vol. 183, Dec. 2024, doi: 10.1016/j.compbiomed.2024.109217.

[4] B. Shang *et al.*, "EEG-based investigation of effects of mindfulness meditation training on state and trait by deep learning and traditional machine learning," *Front Hum Neurosci*, vol. 17, 2023, doi: 10.3389/fnhum.2023.1033420.

[5] M. Naegelin *et al.*, "An interpretable machine learning approach to multimodal stress detection in a simulated office environment," *J Biomed Inform*, vol. 139, Mar. 2023, doi: 10.1016/j.jbi.2023.104299.

[6] L. Mou *et al.*, "Driver stress detection via multimodal fusion using attention-based CNN-LSTM," *Expert Syst Appl*, vol. 173, Jul. 2021, doi: 10.1016/j.eswa.2021.114693.

[7] Y. S. Can, B. Arnrich, and C. Ersoy, "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey," Apr. 01, 2019, *Academic Press Inc.* doi: 10.1016/j.jbi.2019.103139.

[8] S. S. Panicker and P. Gayathri, "A survey of machine learning techniques in physiology based mental stress detection systems," Apr. 01, 2019, *Elsevier Sp. z o.o.* doi: 10.1016/j.bbe.2019.01.004

[9] J. Z. Xiang, Q. Y. Wang, Z. Bin Fang, J. A. Esquivel, and Z. X. Su, "A multi-modal deep learning approach for stress detection using physiological signals: integrating time and frequency domain features," *Front Physiol*, vol. 16, 2025, doi: 10.3389/fphys.2025.1584299.

[10] S. M. Et. al., "Mental Health Prediction Models Using Machine Learning in Higher Education Institution," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 5, pp. 1782–1792, Apr. 2021, doi: 10.17762/turcomat.v12i5.2181.

[11] D. Ghose, A. Chatterjee, I. A. M. Balapuwaduge, Y. Lin, and S. P. Dash, "Investigating lightweight and interpretable machine learning models for efficient and explainable stress detection," *Front Digit Health*, vol. 7, 2025, doi: 10.3389/fdgth.2025.1523381.

[12] A. Almadhor *et al.*, "Wrist-Based Electrodermal Activity Monitoring for Stress Detection Using Federated Learning," *Sensors*, vol. 23, no. 8, Apr. 2023, doi: 10.3390/s23083984.

[13] M. J. Page *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," Mar. 29, 2021, *BMJ Publishing Group*. doi: 10.1136/bmj.n71.

[14] D. Cabrera, L. Cabrera, and E. Cabrera, "The Steps to Doing a Systems Literature Review (SLR)," Apr. 06, 2023. doi: 10.54120/jost.pr000019.v1.

[15] S. Hosseini *et al.*, "A multimodal sensor dataset for continuous stress detection of nurses in a hospital," *Sci Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1038/s41597-022-01361-y.

[16] S. R. Kadu, S. Sadruddin, and D. Argade, "A Multimodal Fusion for Physiological Sensor and Audio Signal-based Stress Detection Using Sliding Hierarchical Bidirectional Recurrent Neural Network," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 3, pp. 569–582, 2025, doi: 10.22266/ijies2025.0430.39.

[17] S. Yang *et al.*, "A deep learning approach to stress recognition through multimodal physiological signal image transformation," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-01228-3.

[18] Y. Yao, M. Papakostas, M. Burzo, M. Abouelenien, and R. Mihalcea, "MUSER: MUltimodal Stress Detection using Emotion Recognition as an Auxiliary Task," May 2021, [Online]. Available: http://arxiv.org/abs/2105.08146

[19] P. Zhang *et al.*, "Real-time psychological stress detection according to ECG using deep learning," *Applied Sciences (Switzerland)*, vol. 11, no. 9, May 2021, doi: 10.3390/app11093838.

[20] J. Zhang, H. Yin, J. Zhang, G. Yang, J. Qin, and L. He, "Real-time mental stress detection using multimodality expressions with a deep learning framework," *Front Neurosci*, vol. 16, Aug. 2022, doi: 10.3389/fnins.2022.947168.

[21] S. A. M. Mane and A. Shinde, "StressNet: Hybrid model of LSTM and CNN for stress detection from electroencephalogram signal (EEG)," *Results in Control and Optimization*, vol. 11, Jun. 2023, doi: 10.1016/j.rico.2023.100231

[22] L. Fontes, P. Machado, D. Vinkemeier, S. Yahaya, J. J. Bird, and I. K. Ihianle, "Enhancing Stress Detection: A Comprehensive Approach through rPPG Analysis and Deep Learning Techniques," *Sensors*, vol. 24, no. 4, Feb. 2024, doi: 10.3390/s24041096.

[23] D. Ghose, O. Gitelson, and B. Scassellati, "Integrating Multimodal Affective Signals for Stress Detection from Audio-Visual Data," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Nov. 2024, pp. 22–32. doi: 10.1145/3678957.3685717.

[24] S. Yang *et al.*, "A deep learning approach to stress recognition through multimodal physiological signal image transformation," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-01228-3.

[25] Z. Yang, H. Yu, and A. Sano, "Contrastive Pretraining for Stress Detection with Multimodal Wearable Sensor Data and Surveys," 2025. [Online]. Available: https://github.com/comp-well-

[26] T. Jeon, H. Byeol Bae, and S. Lee, "Multimodal Stress Recognition Using a Multimodal Neglecting Mask Module," *IEEE Access*, vol. 12, pp. 144774–144787, 2024, doi: 10.1109/ACCESS.2024.3469575.

[27] R. Tanwar, O. C. Phukan, G. Singh, P. K. Pal, and S. Tiwari, "Attention based hybrid deep learning model for wearable based stress recognition," *Eng Appl Artif Intell*, vol. 127, Jan. 2024, doi: 10.1016/j.engappai.2023.107391