Machine learning-based automatic stress detection: Performance and generalization across datasets

Smart Health 41 (2026) 100693

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

Machine learning-based automatic stress detection: Performance and generalization across datasets

Matteo Calza-Metre [a], Luigi Borzì [a,b,*]

[a] Department of Control and Computer Engineering, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy
[b] Polito BIOMed Lab – Biomedical Engineering Laboratory, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy

ARTICLE INFO

Dataset link: https://ubi29.informatik.uni-siegen.de/usi/data_wesad.html, https://data.mendeley.com/datasets/kb42z77m2g/2, https://www.media.mit.edu/tools/affectiveroad/, https://sites.google.com/colorado.edu/verbio-dataset, https://github.com/matteo9910/StressDetectionBasedOnWearableSensorData

Keywords:
Stress detection
Wearable sensors
Physiological signals
Machine learning
Deep learning
Generalization

ABSTRACT

Psychological stress is a major risk factor for several physical and mental health conditions, motivating the development of reliable and unobtrusive monitoring solutions. Wearable devices enable continuous and ecological acquisition of physiological and physical signals, providing a promising basis for automatic stress detection in real-world settings. In this context, machine learning (ML) and deep learning (DL) methods have shown strong potential for extracting stress-related patterns from multimodal wearable data. However, most studies evaluate performance on a single dataset, limiting insights into model robustness and generalization across populations and experimental conditions. Furthermore, there is currently no comprehensive comparison between feature-based ML methods and data-driven DL methods. Finally, the use of transfer learning to address differences across various datasets has not yet been thoroughly explored.

In this work, we investigate automatic stress detection from smartwatch data by comparing feature- and data-driven models, focusing on cross-dataset generalization. Experiments were conducted on four public datasets collected with the same wearable device and sensing modalities, including photoplethysmography, acceleration, skin temperature, and electrodermal activity. We first evaluated model performance within each dataset to establish baseline results. We then assessed generalization using cross-dataset testing, where models trained on one dataset are evaluated on unseen datasets. Finally, we explored different transfer learning strategies to improve cross-dataset performance and promote generalization, especially in small datasets.

Results show that DL methods (F1-score 0.69–0.96) consistently outperform ML models (F1-score 0.69–0.95) in within-dataset evaluations, although performance varies across datasets. Cross-dataset tests reveal significant performance degradation, particularly for DL models, where the F1-score showed an average drop of −21%. Transfer learning mitigates this gap, improving robustness across datasets. These findings highlight the importance of cross-dataset evaluation for realistic assessment of wearable-based stress detection systems.

1. Introduction

Stress is a psychophysiological response to internal or external demands perceived as exceeding an individual's adaptive resources. According to Lazarus et al. (1985), stress arises when environmental demands surpass one's ability to meet, mitigate, or alter them. While acute stress triggers time-limited physiological responses to identifiable short-term stressors such as public speaking or job interviews, chronic stress stems from persistent demands including financial strain, caregiving responsibilities, or occupational

* Corresponding author.
E-mail address: luigi.borzi@polito.it (L. Borzì).

https://doi.org/10.1016/j.smhl.2026.100693
Received 23 March 2026; Received in revised form 16 June 2026; Accepted 2 July 2026
Available online 9 July 2026

Table 1
Physiological signals for stress detection: origin and typical stress-induced variations.

Biosignal	Physiological origin	Typical stress response
Electrocardiography/Photoplethysmography *(ECG/PPG)*	Cardiac electrical activity and peripheral blood volume under autonomic control	Increased heart rate, reduced heart rate variability
Electrodermal Activity *(EDA)*	Sympathetic sudomotor activity of eccrine sweat glands	Elevated baseline conductance, increased frequency and amplitude of responses
Electroencephalography *(EEG)*	Cortical electrical rhythms reflecting arousal and cognitive load	Altered brain rhythms, increased arousal patterns
Electromyography *(EMG)*	Skeletal muscle activation and tension	Elevated muscle tension, particularly in postural muscles
Skin Temperature *(SKT)*	Cutaneous perfusion modulated by thermoregulatory control	Decreased peripheral temperature due to vasoconstriction
Respiration *(RESP)*	Ventilatory control regulated by autonomic mechanisms	Faster and more irregular breathing patterns
Accelerometry *(ACC)*	Physical activity and movement patterns	Contextual movement information for artifact distinction

pressures. Prolonged exposure to chronic stress activates the hypothalamic-pituitary-adrenal axis and sympathetic-adrenal-medullary system, leading to sustained release of cortisol and catecholamines that contribute to allostatic load and cumulative biological burden (Ghasemi et al., 2024). The health consequences of chronic stress are profound, substantially increasing the risk for cardiovascular diseases, depression, immune dysfunction, and systemic inflammation, while simultaneously reducing quality of life through impaired cognitive function, disrupted sleep patterns, and diminished occupational performance (Epel et al., 2018).

Traditional stress assessment relies primarily on self-reported questionnaires such as the Perceived Stress Scale (PSS), State-Trait Anxiety Inventory (STAI), and NASA Task Load Index (TLX), which provide subjective appraisals of perceived stress levels, anxiety states, and task-related workload (de Witte et al., 2021; Harris et al., 2023). While these instruments offer valuable insights into individuals' cognitive and emotional interpretations of stressors, they suffer from critical limitations including recall bias, social desirability effects, temporal inconsistencies, and the burden of frequent reporting that reduces ecological validity (Tramel et al., 2023). Physiological markers such as salivary cortisol provide objective stress indicators but require laboratory processing and fail to capture real-time dynamics. Consequently, traditional methods are ill-suited for continuous, unobtrusive monitoring in naturalistic settings where stress unfolds over extended periods and across diverse contexts.

Wearable devices have emerged as a transformative technology for stress monitoring, enabling continuous, non-invasive acquisition of multimodal physiological signals in daily life (Kulanthaivel Lakshmanan et al., 2026). Unlike traditional laboratory equipment that constrains movement and limits ecological validity, wrist-worn wearables such as smartwatches support longitudinal observation and seamless integration into daily routines (Hickey et al., 2021; Taskasaplidis et al., 2024). Research-grade devices such as the Empatica E4 simultaneously capture electrodermal activity (EDA), photoplethysmography (PPG) for blood volume pulse (BVP) and heart rate variability (HRV), skin temperature (SKT), and tri-axial accelerometry (ACC), providing comprehensive multimodal data that increases robustness and accuracy of stress detection (D. Bolpagni et al., 2024; Milstein & Gordon, 2020). The portability and unobtrusiveness of wrist-worn form factors facilitate data collection in real-world environments, bridging the gap between controlled laboratory experiments and ecological field studies where behavioural context and motion artifacts introduce additional complexity.

Physiological signals recorded by wearable sensors reflect autonomic nervous system reactivity to stress. Electrocardiography (ECG) and PPG capture cardiac activity, with stress inducing increased heart rate (HR) and reduced HRV through sympathetic activation and parasympathetic withdrawal (Giannakakis et al., 2022). EDA measures sympathetic sudomotor function, exhibiting elevated baseline conductance and increased response frequency under stress (Giannakakis et al., 2022). Electroencephalography (EEG) reveals cortical arousal patterns associated with cognitive load (Alonso et al., 2015), while electromyography (EMG) detects elevated muscle tension in stress-responsive postural muscles. Respiration becomes faster and more irregular during stress episodes, and SKT decreases peripherally due to vasoconstriction (Herborn et al., 2015). ACC provides essential contextual information about physical activity, enabling differentiation between genuine stress responses and motion-related artifacts in free-living conditions (Gjoreski et al., 2016; Nugroho et al., 2026). The combination of these multimodal signals enables comprehensive monitoring across cardiovascular, electrodermal, thermoregulatory, muscular, and behavioural dimensions of the stress response (Abdelfattah et al., 2025; Xiang et al., 2025b), as summarized in Table 1.

Recent advances in wearable-based stress detection have been strongly driven by the availability of multimodal physiological datasets and systematic studies of real-world deployment. Recent surveys highlight a growing ecosystem of wearable stress monitoring systems, together with machine learning (ML) and deep learning (DL) methods for automated stress recognition (M. Bolpagni et al., 2024; Pinge et al., 2024). At the same time, systematic reviews emphasize that although these approaches achieve strong performance in controlled settings, their translation to real-world scenarios remains limited by dataset heterogeneity, small sample sizes, and inconsistent experimental protocols (Pataca et al., 2025; Vos et al., 2023). More recent work further shows that model performance significantly degrades under cross-dataset and cross-population evaluation, indicating that current benchmarks may overestimate generalization capability (Cui et al., 2025; Sigcha et al., 2026). These findings motivate more rigorous evaluation approaches that explicitly assess robustness across datasets, subjects, acquisition conditions, and ML methods.

2. Related work

The earliest evidence that physiological signals carry discriminative information about psychological stress comes from controlled laboratory studies conducted with standard clinical equipment. Hjortskov et al. (2004) were among the first to systematically quantify the effect of mental stress induced by computer-based tasks on cardiac autonomic regulation: HR increased significantly during the stress condition while HRV decreased, with a reduction in high frequency (HF) power and an elevated low frequency to high-frequency ratio (LF/HF) reflecting a shift towards sympathetic dominance. These findings established HRV as a robust cardiovascular marker of acute stress and provided the methodological foundation for subsequent feature-based approaches. Contemporaneously, the utility of the electrodermal channel was demonstrated by Setz et al. (2010), who showed that EDA could reliably discriminate psychological stress from cognitive load, with stress conditions eliciting significantly higher skin conductance response (SCR) frequency and amplitude than neutral cognitive tasks. Healey and Picard (2005) extended this multimodal perspective to an ecologically valid scenario, instrumenting 24 drivers with body-worn ECG, EMG, EDA, and respiration sensors during approximately 50-minute on-road sessions; supervised classifiers trained on time- and frequency-domain descriptors discriminated three driving stress levels, with multimodal fusion consistently outperforming any single-signal baseline and confirming the complementarity of cardiac and electrodermal measures. A subsequent meta-analysis by Kim et al. (2018) consolidated this body of evidence, reviewing dozens of controlled studies and confirming that temporal and spectral HRV indices constitute a reliable and replicable marker set for stress classification across experimental paradigms.

The advent of wearable sensing enabled the non-invasive, continuous, and ecologically valid acquisition of multimodal physiological signals beyond the clinical laboratory, opening new possibilities for stress monitoring in everyday life. Early affect-oriented corpora such as MAHNOB-HCI (Soleymani et al., 2012) and DEAP (Koelstra et al., 2012) combined peripheral physiology — ECG, EDA, respiration, and SKT — with controlled audiovisual elicitation and subjective ratings, catalyzing advances in feature engineering (notably HRV time–frequency analysis and tonic/phasic EDA decomposition) and introducing evaluation practices that later migrated to stress-specific research. SWELL-KW (Koldijk et al., 2014) then introduced a knowledge-work paradigm closer to realistic occupational settings: 25 participants performed document-editing tasks under time pressure and email interruptions while ECG, EDA, posture, and facial expressions were simultaneously recorded, providing a reproducible benchmark for work-related stressors validated with questionnaire-based labels. Setz et al. (2010) had already shown that wrist-worn EDA could discriminate stress in controlled settings using statistical classifiers (i.e., linear discriminant analysis — LDA), achieving approximately 78% accuracy in distinguishing stress from cognitive load, an early demonstration that miniaturized sensing and simple statistical models sufficed for binary discrimination.

The integration of ML techniques into this pipeline produced substantial performance gains and enabled systematic comparison of classifiers and feature spaces. Hovsepian et al. (2015) proposed the *cStress* methodology — covering signal quality screening, filtering, handcrafted feature computation, and support vector machine (SVM) classification. The proposed approach was validated both in a controlled setting ($n = 24$) and in daily life over one week ($n = 20$), using ecological momentary assessment as ground truth. The study reported high laboratory accuracy (around 90%) against substantially lower wild-condition performance (around 70%), and foregrounding the generalization gap that would become a central concern in the field. Gjoreski et al. (2017, 2016) addressed this gap with a two-stage architecture: a random forest (RF) detector on short windows was followed by an SVM that ingested recent detector outputs together with contextual signals (activity, posture, phone usage), improving in-lab accuracy to approximately 83% and reaching ∼92% with context over longer horizons, demonstrating that behaviour-aware features reduce false positives in unconstrained settings. The release of Wesad (2018) (Schmidt et al., 2018) marked a watershed for reproducible benchmarking, offering synchronized chest- and wrist-worn signals (ECG, EDA, BVP, respiration, ACC) from 15 participants across baseline, Trier Social Stress Test (TSST) induced social stress, and amusement. Schmidt et al. reported that an LDA classifier trained on a comprehensive handcrafted feature set achieved macro-F1 around 0.80 under leave-one-subject-out (LOSO) evaluation, with multimodal fusion consistently outperforming single-channel baselines. Wesad rapidly became the standard platform for comparing preprocessing choices, classical feature-driven models, and evaluation protocols. Benchekroun et al. (2021) further systematized the classical ML landscape by extracting temporal and spectral features from ECG, EDA, and respiration signals and benchmarking SVM, RF, and logistic regression (LR) on Wesad, reporting intra-dataset F1-scores above 0.80 under LOSO. Kaczor et al. (2020) applied the same paradigm to physicians in an emergency department, using HRV and EDA features with SVM classification and achieving approximately 84% accuracy, highlighting the clinical feasibility of occupational stress monitoring with wearables. Zhu et al. (2023) focused on the EDA signal, extracting 20 temporal and frequency-domain features from wrist-worn recordings and evaluating multiple classifiers (SVM, RF, gradient boosting — GB), reporting accuracy exceeding 85% under subject-dependent evaluation and confirming the discriminative power of EDA features even without complementary cardiac signals. Broad surveys by Can et al. (2019) and Vos et al. (2022) synthesized these developments, concluding that HRV is the dominant information source across studies, with an accuracy typically ranging from 70%–90% in laboratory environments, decreasing to 60%–80% in unconstrained daily life. The performance drop was attributed to different aspects, including small cohorts, heterogeneous stressors, and overfitting to dataset-specific artefacts.

The recent advent of DL techniques brought further performance gains by removing the need for manual feature engineering and enabling end-to-end representation learning directly from raw or minimally processed data (Patwal et al., 2023). Wang and Wu (2022) proposed a convolutional neural network (CNN)-based architecture trained on raw PPG windows, reporting F1-scores around 0.75 in intra-dataset evaluation and demonstrating that learned representations capture stress-related morphological and temporal changes without explicit HRV extraction. Ladakis et al. (2022) explored lightweight CNNs operating on spectrogram representations of multimodal biosignals, achieving competitive intra-dataset performance (F1 $\approx$ 0.82 on Wesad under LOSO) while substantially

reducing model complexity. Tanwar et al. (2023) extended this direction with an attention-based hybrid CNN-recurrent neural network (RNN) architecture for wearable EDA and BVP signals, reporting macro-F1 values above 0.87 on controlled datasets and demonstrating that temporal self-attention improves the localization of stress-relevant signal segments. Albaladejo González et al. (2022) introduced the large-scale UBFC-Phys dataset and evaluated several deep neural models — CNN, long short-term memory (LSTM), and Transformer variants — showing that richer training data and end-to-end architectures push in-domain performance beyond traditional feature-driven pipelines.

A critical dimension that cuts across the entire trajectory described above concerns the choice of validation protocol, which profoundly affects the interpretation of reported results. Subject-independent protocols — most notably LOSO cross-validation — hold out entire individuals at test time, thus reflecting the realistic scenario in which a model must generalize to previously unseen users; in contrast, random shuffling or subject-dependent k-fold protocols mix windows from the same individual across train and test splits, inflating performance estimates by allowing the model to memorize individual physiological patterns. Several studies have quantified this effect: the same classifier can exceed 90% accuracy under subject-dependent evaluation while falling below 70% under LOSO on the same dataset (Can et al., 2019; Vos et al., 2022). Cross-dataset testing — training on one corpus and evaluating on an entirely different cohort and protocol — constitutes an even stricter criterion and has emerged as the principal bottleneck in recent research. Benchekroun et al. (2023) trained HRV-based classifiers on one cohort and tested on another with different recording conditions, with a RF achieving a cross-dataset macro-F1 of approximately 0.61, clearly lagging behind within-dataset performance. Prajod et al. (2024) further showed that stressor type is a dominant determinant of cross-dataset transferability: models trained on socially induced stress generalize substantially better to similarly elicited corpora than to cognitive or arithmetic task datasets. Ladakis et al. (2025) demonstrated that feature harmonization and data pooling across multiple open datasets can raise accuracy above 80%, but performance on fully unseen participants still settles around 70%, confirming that domain shift remains a major open challenge. Lazarou and Exarchos (2024) similarly identified temporal alignment across channels, label noise, and on-device latency constraints as key obstacles to robust real-time deployment.

2.1. Research gap and study contributions

Recent survey and review studies on wearable-based stress detection consistently identify several unresolved challenges that hinder real-world applicability (Kapogianni et al., 2025; Pataca et al., 2025; Sosa et al., 2026). In particular, they highlight limited cross-dataset generalization, strong heterogeneity in data collection protocols and label definitions, and the lack of standardized evaluation frameworks. Surveys also emphasize that most studies rely on within-dataset validation, which tends to overestimate performance, while systematic comparisons between feature-driven and data-driven approaches remain scarce. Furthermore, the role of transfer learning and domain adaptation in physiological stress detection is still insufficiently explored, leaving open questions on how to achieve robust performance across heterogeneous real-world conditions.

Table 2 summarizes the studies discussed in this section, reporting for each work the dataset(s) used, wearable sensors and signals acquired, analytical approach, validation protocol, and key performance metrics.

Consistent with the gaps identified in recent survey literature, several critical limitations persist in the existing literature. Most publicly available datasets involve small participant cohorts — typically between 15 and 40 subjects — recorded predominantly in controlled laboratory settings with standardized stressors. Models trained on these datasets achieve high within-dataset accuracies that do not reflect real-world performance, where signal quality degrades due to motion artefacts and behavioural context introduces confounds (Lazarou & Exarchos, 2024). Labelling protocols and preprocessing pipelines vary widely across studies, making direct comparisons of reported results unreliable and reproducibility difficult to assess (Smets et al., 2023; Vos et al., 2022). More critically, within-dataset cross-validation — the dominant validation strategy — systematically overestimates generalization by failing to expose domain shift arising from population differences, stressor heterogeneity, and environmental variability encountered in free-living conditions (Can et al., 2019; Vos et al., 2022). Cross-dataset testing and transfer learning evaluations remain rare: a review by Vos et al. (2022) found that among 33 studies using public datasets, only a minority conducted any form of external validation. Recent studies have begun addressing this scenario (Benchekroun et al., 2023; Prajod et al., 2024), but typically evaluate zero-shot transfer without adaptation, yielding substantial performance degradation. Transfer learning strategies remain largely underexplored in physiological stress detection (Vafaei & Hosseini, 2025), and no study has systematically compared feature-driven and data-driven paradigms under both zero-shot and adaptation conditions across multiple heterogeneous corpora, leaving practitioners without evidence-based guidelines on which approach to adopt depending on target-domain data availability.

This work addresses these gaps through four main contributions: (i) a multi-dataset evaluation framework spanning four datasets (Wesad, Campanella, AffectiveRoad, VerBio) that differ in experimental protocols and ecological validity while sharing identical sensing hardware (Empatica E4), isolating domain shift from sensor heterogeneity; (ii) a unified preprocessing and feature extraction pipeline applied consistently across all datasets, ensuring that performance differences reflect modelling choices rather than dataset-specific handling; (iii) a systematic comparison between a handcrafted feature pipeline paired with classical ML models and a multi-input 1-D CNN with end-to-end representation learning, both evaluated under LOSO cross-validation; and (iv) a dual external validity assessment combining zero-shot cross-dataset testing and transfer learning, providing the first systematic evidence on when each paradigm should be preferred in real-world deployment scenarios.

Table 2
Summary of representative studies on physiological stress detection, ordered chronologically. Acc. = Accuracy; F1 = macro-averaged F1-score. SD = Subject-Dependent; LOSO = Leave-One-Subject-Out.

Study	Dataset	Sensors/Signals	Analysis	Validation	Results
Hjortskov et al. (2004)	Lab computer tasks ($n = 24$)	Standard ECG (chest)	Time/freq. HRV features; paired statistical testing	Within-subject comparison	Significant HR ↑, HRV ↓ under stress
Healey and Picard (2005)	Naturalistic driving ($n = 24$)	Body sensors (ECG, EDA, EMG, Resp.)	Time/freq. HRV & EDA features; LDA, BN classifiers; multimodal fusion	SD hold-out	Acc. ≈97% (3-class, multimodal)
Setz et al. (2010)	Lab MIST ($n = 17$)	Wearable EDA (wrist)	Tonic/phasic EDA features; LDA classifier	SD, 5-fold CV	Acc. ≈78% (stress vs. load)
Koldijk et al. (2014)	SWELL-KW ($n = 25$)	Body sensors (ECG, EDA, posture, facial)	HRV & EDA temporal/spectral features; multiple classifiers	SD, 10-fold CV	F1 ≈0.74 (best model)
Hovsepian et al. (2015)	cStress lab + field ($n = 44$)	Wrist (EDA, BVP, ACC)	HRV, EDA, motion features; SVM	SD lab; real-life EMA	Acc. ≈90% (lab), ≈70% (field)
Schmidt et al. (2018)	Wesad ($n = 15$)	RespiBAN (ECG, EDA, Resp., Temp., ACC) + Empatica E4 (EDA, BVP, Temp., ACC)	37 handcrafted temporal/spectral features; LDA, RF, AdaBoost	LOSO	F1 ≈0.80 (LDA, multimodal)
Kaczor et al. (2020)	ED physicians ($n = 12$)	Empatica E4 (EDA, BVP, Temp., ACC)	HRV & EDA temporal features; SVM	SD, RS + 5-fold CV	Acc. ≈84%
Benchekroun et al. (2021)	Wesad; AffectiveRoad	Body sensors (ECG, EDA, Resp.)	Handcrafted HRV & EDA features; SVM, RF, LR	LOSO (intra); cross-dataset	F1 >0.80 (intra); F1 ≈0.61 (cross)
Wang and Wu (2022)	Lab PPG dataset ($n = 36$)	PPG (wrist)	Raw windows → 1-D CNN; end-to-end learning	SD, 5-fold CV	F1 ≈0.75
Ladakis et al. (2022)	Wesad; custom datasets	Empatica E4 (EDA, BVP, Temp., ACC)	Spectrogram representations → lightweight CNN	LOSO (intra only)	F1 ≈0.82 (intra-Wesad)
Benchekroun et al. (2023)	Wesad; AffectiveRoad	Body sensors (ECG, EDA, Resp.)	HRV time-domain & non-linear features; RF, SVM	LOSO (intra); cross-dataset	F1 ≈0.81 (intra); F1 ≈0.61 (cross)
Tanwar et al. (2023)	Wesad; SWELL-KW	Empatica E4 (EDA, BVP, ACC)	Raw signals → attention-based CNN-RNN hybrid	LOSO	F1 ≈0.87
Zhu et al. (2023)	Lab MIST ($n = 20$)	Empatica E4 (EDA, wrist)	20 temporal/freq. EDA features; SVM, RF, GB	SD, 10-fold CV	Acc. ≈86% (RF)
Ladakis et al. (2025)	Multiple open datasets (Wesad, SWELL-KW, others)	Mixed wearables (ECG, EDA, BVP, Temp., ACC)	Harmonized features; pooled training; ensemble models	LOSO; cross-dataset	Acc. >80% (intra); ≈70% (cross)

3. Methodology

3.1. Data

Four publicly available multimodal datasets were analysed in this study, namely Wesad, Campanella, AffectiveRoad, and VerBio. The selection of these datasets was guided by the need for hardware comparability and rigorous cross-dataset generalization assessment. All these corpora include data acquired with the Empatica E4 wristband, ensuring homogeneous sensor characteristics. The recorded signals include EDA, BVP, SKT, and ACC. Despite identical hardware, datasets differ substantially in experimental protocols and stress induction procedures, spanning controlled laboratory stressors (Wesad, Campanella) to ecologically valid scenarios (AffectiveRoad, VerBio), thereby enabling the evaluation of model robustness across diverse conditions. All resources are publicly accessible, anonymized, and distributed under research-oriented licenses. Table 3 summarizes key characteristics of the four datasets.

3.1.1. Wesad

Wesad (Wearable Stress and Affect Detection) is a controlled laboratory dataset collected from 15 healthy graduate students (12 males, 3 females; mean age 27.5 ± 2.4 years) using synchronized chest-worn RespiBAN Professional and wrist-worn Empatica E4 devices. Only E4 wrist signals were retained in this study to maintain hardware homogeneity across datasets. The experimental protocol, lasting approximately two hours, aimed to elicit three affective states — baseline, stress, and amusement — with interleaved meditation periods. Protocol phases included: (a) 20-minute baseline neutral reading task, (b) 6.5-minute amusement

Table 3
Summary of dataset characteristics: sensors, experimental protocols, stress labelling schemes, and recording duration. Acronyms: TSST: Trier Social Stress Test; VR: virtual reality; RELAX: 5-min nature video baseline phase; PREP: 10-min silent article preparation phase; PPT: 5-min oral presentation phase. Duration per subject refers to the full session length; total duration is computed as duration per subject times number of subjects (or sessions for AffectiveRoad).

Dataset	Sensors	Experiments	Stress label	Duration
Wesad (Schmidt et al., 2018)	E4 (wrist) + RespiBAN (chest)	Controlled lab: baseline (20 min reading), TSST social stress (13 min), amusement (11 video clips), meditation; $n = 15$	Task-based binary: baseline vs. TSST stress	~120 min/subject; ~30 h total
Campanella et al. (2023)	E4 (wrist)	Controlled lab: alternating rest and heterogeneous stressors (Lego assembly, cognitive arithmetic, social presentation); $n = 29$	Task-based binary: rest vs. cognitive/social/combined stress	~37 min/subject; ~18 h total
AffectiveRoad (El Haouij et al., 2018)	E4 (wrist)	Naturalistic driving: 31 km route alternating parking, highway, urban traffic; $n = 9$ participants, 13 sessions	Continuous $[0, 1]$ from observer+driver consensus; binarized at threshold 0.75	~85 min/session; ~11 h total (13 sessions)
VerBio (Lab, 2023)	E4 (wrist) + Actiwave ECG (chest)	Public speaking: 10 sessions across 4 days (real + VR audiences); RELAX (5 min), PREP (10 min), PPT (5 min); $n = 29$ completed	Continuous $[0, 1]$ from 4 raters fused at 1 Hz; binarized at threshold 0.20	~20 min/session; ~97 h total (10 sessions × 29 subjects)

condition with humorous video clips, (c) modified TSST comprising 3-minute preparation, 5-minute impromptu speech before a panel, and 5-minute serial subtraction task, (d) 7-minute guided meditation after each stimulation block, and (e) 10-minute recovery. Stimulus blocks were counterbalanced across subjects (Versions A/B) to mitigate order effects, with half performing standing and half sitting.

3.1.2. Campanella

The Campanella dataset was acquired in controlled laboratory conditions from 29 healthy participants wearing a single wrist-worn Empatica E4. The experimental design reproduced a demanding work-like environment through heterogeneous stressors spanning cognitive, social, and combined modalities over approximately 37 min. After an initial 3-minute rest, participants performed a series of pre-defined tasks: 10-minute Lego assembly without instructions (problem-solving under time pressure), 2-minute recovery, 5-minute Lego assembly with instructions (reduced cognitive load), 2-minute recovery, 3-minute dual-task combining Lego assembly with backward counting from 180 (manual+arithmetic stress), 2-minute recovery, untimed cognitive arithmetic inspired by Montreal Imaging Stress Task (MIST), 2-minute recovery, and 1-minute social stressor involving oral CV self-presentation.

3.1.3. AffectiveRoad

AffectiveRoad investigates driver stress under naturalistic conditions during real traffic in the Greater Tunis area. Data were collected along a 31 km route lasting approximately 85 min, alternating city centers, highways, suburban zones, and parking areas. Each of 13 recording sessions (9 unique participants, some repeated) integrated Empatica E4 wrist signals, frontal and internal dashboard cameras, GPS, and environmental data. A trained rear-seat experimenter produced continuous real-time stress annotations on a slider interface synchronized with video feeds. The 31 km route comprised segments designed to elicit varying cognitive loads: parking and Z-zones (very low stress), highway sections at 80–100 km/h (medium stress), and dense urban traffic with frequent stops and complex interactions (high stress).

3.1.4. VerBio

VerBio targets stress detection during public speaking under real and virtual reality (VR) exposure. The corpus integrates Empatica E4 wrist signals with chest ECG (Actiwave Cardio, 512 Hz) and speech audio (16 kHz); VR scenarios were delivered via Oculus Rift using the Virtual Orator platform. The study spanned five months with 55 university students (18–30 years, near-balanced gender); 36 completed TEST series, 29 completed POST. Ten sessions per participant were organized across four days: PRE (day 1, real audience of 5 persons), TEST01-TEST08 (days 2–3, eight VR sessions), and POST (day 4, real audience). Each session followed a fixed three-phase structure: RELAX (5 min nature video baseline), PREP (10 min silent article preparation), and PPT (5 min oral presentation). VR configurations factorially varied room type (boardroom, classroom, theatre, seminar), audience size (12, 25, 54, 90), and reaction (negative, neutral, positive), with ambient classroom noise for ecological validity.

3.2. Preprocessing

The two modelling paradigms adopted in this study — the *features-driven* (FD) and the *data-driven* (DD) pipelines — share a common upstream preprocessing stage comprising signal ingestion, resampling, and denoising, but diverge after that point. In the FD pipeline, cleaned and windowed signals undergo an explicit, hand-crafted *feature extraction* phase followed by a statistical *feature selection* procedure, ultimately producing a compact tabular representation as input to classical ML classifiers. In contrast, the DD pipeline terminates preprocessing after windowing: the cleaned multi-channel time series are fed directly into a CNN, which autonomously learns discriminative representations without any manual feature engineering. This design allows a controlled comparison in which the only difference between the two branches originates from how discriminative information is extracted from the same physiologically cleaned signals.

Since the four datasets adopt heterogeneous labelling schemes, a harmonization step was applied to each corpus to produce a unified binary ground truth (0: baseline, 1: stress). For Wesad, only baseline and TSST stress segments were retained, excluding amusement and meditation periods; original labels sampled at 700 Hz were downsampled to 64 Hz to align with the wrist signals. For Campanella, the native binary taxonomy was preserved directly, mapping rest intervals to baseline and all stress-inducing tasks (cognitive, social, and combined) to stress. For AffectiveRoad, the original continuous annotations in the range $[0, 1]$ were binarized by thresholding at $s = 0.75$, corresponding to the "high stress" boundary defined in the original study (El Haouij et al., 2018), which categorized perceived stress into three levels: low ($0.0 \leq s < 0.4$), medium ($0.4 \leq s < 0.75$), and high ($s \geq 0.75$). A threshold sensitivity analysis (Supplementary Material, Table 11 and Fig. 7, Appendix A.1) confirmed that this value yields the most balanced class distribution (51% baseline/49% stress) while retaining 12 out of 13 subjects above the 5% minimum stress prevalence criterion; lower thresholds produce heavily stress-dominant distributions unsuitable for binary classification, with samples above the threshold assigned to stress and all others to baseline; the annotation trace was upsampled to 64 Hz via nearest-neighbour replication, and only the first session per participant was retained to mitigate habituation effects. For VerBio, continuous ratings from four independent raters were fused into a single trace at 1 Hz; RELAX segments were assigned to baseline, while PPT segments with fused rating $s \geq 0.20$ were assigned to stress. The 0.20 threshold was selected following a systematic sensitivity analysis (Supplementary Material, Table 12 and Fig. 7, Appendix A.1): it is the minimum value that retains all 17 subjects with more than 5% stress-labelled windows; at threshold ≥ 0.3 two subjects fall below this criterion, and at threshold ≥ 0.5 the majority of subjects retain virtually no stress labels, rendering the dataset unsuitable for classification. PREP segments were excluded as not clearly attributable to either class.

Data ingestion and resampling. All four datasets were ingested under a unified framework centered on the Empatica E4 channels: EDA, BVP, SKT, and ACC. Raw files were organized *per subject* (and *per session* when available) into a canonical tabular structure with standardized column headers and explicit identifiers (`subject`, `session`, `phase`). Dataset-specific labels — either natively provided or reconstructed from experimental protocols — were imported and linked to the corresponding temporal grid.

To enable joint multimodal analysis, all channels were resampled onto a common temporal grid at 64 Hz, matching the highest native sampling rate among the E4 modalities (BVP). This choice ensured that no high-frequency cardiovascular information (0.5–10 Hz) was lost through downsampling. Slower channels were upsampled using interpolation schemes suited to their dynamics: zero-order hold (nearest-neighbour replication) for the slowly varying EDA and SKT signals, and linear interpolation for the motion channel (ACC). Low-pass anti-alias filtering was applied prior to synchronization where necessary. Ground-truth labels sampled at lower or non-uniform frequencies were brought to the 64 Hz grid via forward-fill or discrete interpolation, ensuring a valid class label at every timestamp.

The tri-axial accelerometer components were aggregated into a single magnitude channel (Eq. (1)) to reduce dimensionality, reduce bias and noise due to orientation-specific patterns, and emphasize overall movement intensity. The pre-processing pipeline is schematically summarized in Fig. 1.

$$\mathrm{ACC}_{\mathrm{MAG}} = \sqrt{\mathrm{ACC}_x^2 + \mathrm{ACC}_y^2 + \mathrm{ACC}_z^2} \tag{1}$$

Denoising and Artifact Handling Signal cleaning was applied uniformly across all datasets and modalities using zero-phase (forward–backward) Butterworth filtering throughout, so as to prevent phase distortion and preserve the temporal alignment of physiological waveforms. The BVP signal was processed with an adaptive band-pass Butterworth filter (0.5–8 Hz) to preserve pulsatile cardiac dynamics while suppressing low-frequency drift and high-frequency noise; automatic artifact detection based on peak morphology and inter-beat interval regularity was additionally performed, and outliers and implausible peaks were corrected via local interpolation, ensuring a physiologically consistent waveform suitable for subsequent HR and HRV extraction. The EDA signal was first detrended to remove slow baseline drifts, then low-pass filtered with a Butterworth filter (cutoff: 5 Hz) to eliminate high-frequency components unrelated to electrodermal dynamics, followed by adaptive outlier correction; the cleaned signal was subsequently decomposed into its tonic component (Skin Conductance Level - SCL, reflecting baseline conductance) and phasic component (SCR, capturing transient sympathetic responses), enabling the extraction of stress-relevant electrodermal features. The aggregated accelerometry magnitude was filtered with a two-stage Butterworth scheme: a high-pass filter at 0.5 Hz to remove gravitational offset and baseline wander, followed by a low-pass filter at 20 Hz to attenuate sensor electronics noise. SKT was denoised with a single-stage Butterworth low-pass filter at 0.5 Hz, consistent with the slow thermoregulatory dynamics of peripheral skin; no high-pass stage was employed, as temperature trends evolve on the order of seconds to minutes and must be preserved for subsequent feature extraction.

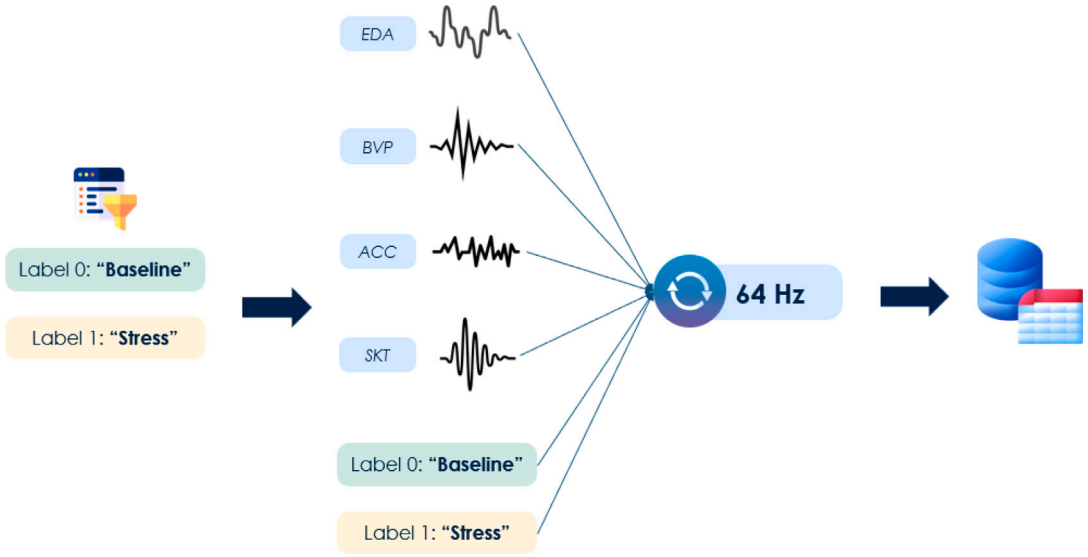


Fig. 1. Data ingestion and resampling pipeline. Each dataset is filtered to retain only baseline (label = 0) and stress (label = 1) segments; all signal modalities (EDA, BVP, ACC, SKT) and labels are resampled to a common 64 Hz grid and consolidated into a unified Dataframe.

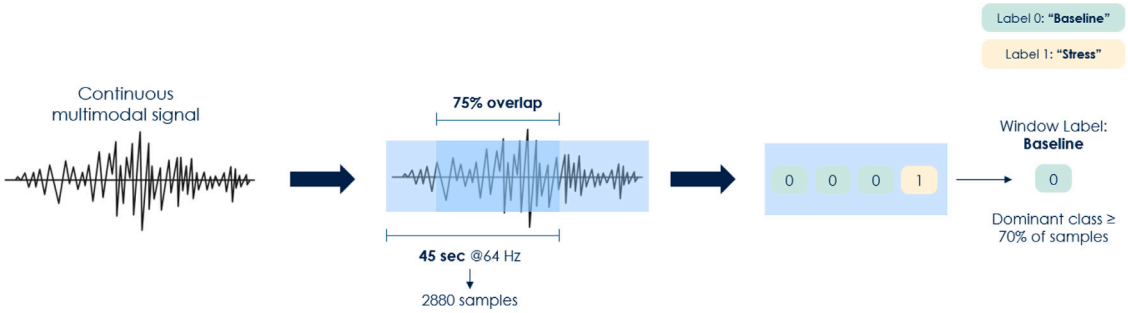


Fig. 2. Overview of the windowing and dominant-label assignment procedure.

Windowing and window-level label assignment. Following denoising, all cleaned signals were segmented into fixed-length, partially overlapping windows to produce discrete instances suitable for supervised learning.

A window duration of 45 s with 75% overlap was adopted following a systematic ablation study conducted on the Wesad benchmark, covering three combinations (window, overlap) for the FD and DD pipelines; the selected configuration emerged as the only one ranking near-optimal across both paradigms on accuracy, macro-F1, and number of features surviving selection (see Supplementary Material, Table 13, Appendix A.2). For each window, a binary label was assigned using the *dominant-label rule*: the most frequently occurring class within the window was assigned as its label. To ensure labelling reliability, only windows in which the dominant class constituted at least 70% of the samples were retained; ambiguous transition segments were discarded. This windowing and labelling strategy was applied identically across all four datasets. In the DD pipeline, preprocessing terminates at this stage: each modality's windows are stored as single-channel arrays of shape (n_windows, n_samples, 1) and fed directly to the CNN. The windowing and label assignment process is schematically represented in Fig. 2.

Feature extraction and selection. Both steps apply exclusively to the FD pipeline. From each 45s-window, a set of hand-crafted descriptors was extracted per modality. For BVP, individual pulse cycles were identified through peak detection on the cleaned waveform, from which HR and HRV indices were derived: time-domain metrics including SDNN, RMSSD, and pNN50; frequency-domain parameters including LF power, HF power, and the LF/HF ratio; nonlinear complexity measures including Sample Entropy, Fuzzy Entropy, Multiscale Entropy, Correlation Dimension, Higuchi and Katz Fractal Dimensions, Lempel–Ziv Complexity, and Multifractal DFA descriptors; and pulse-morphology descriptors including mean systolic peak amplitude, mean beat duration, and average inter-beat interval. For EDA, SCR detection was performed on the phasic component of the cleaned signal, yielding SCR count and mean SCR amplitude, together with the tonic SCL level and the phasic area under the curve (AUC). For ACC_{MAG} and SKT, descriptive statistics were computed over each window, comprising mean, standard deviation, minimum, maximum, range, median, IQR, skewness, and kurtosis; for SKT, a linear-regression slope was additionally computed to capture short-term thermal trends.

Table 4
List of extracted features retained after the selection pipeline, grouped by physiological signal.

Signal	Domain	Feature	Description
BVP	Time	MeanNN	Mean of NN (beat-to-beat) intervals
		SDNN	Standard deviation of NN intervals
		MedianNN	Median of NN intervals
		MadNN	Median absolute deviation of NN intervals (robust dispersion)
		SDRMSSD	Standard deviation of the RMSSD series
		Prc20NN	20th percentile of NN intervals
		pNN50	Percentage of successive NN pairs differing by >50 ms
		MinNN	Minimum NN interval in the window
		HTI	HRV triangular index (histogram-based dispersion)
		TINN	Triangular interpolation of the NN histogram
	Frequency	VHF	Spectral power in the very-high-frequency band (>0.40 Hz)
		S	Total spectral power of the NN-interval series
		PIP	Parasympathetic index (HF/total power)
		PI	Parasympathetic index variant (HF/(LF + HF))
	Nonlinear	SampEn	Sample entropy of the NN-interval series (pattern irregularity)
		FuzzyEn	Fuzzy entropy of the NN-interval series (noise-robust complexity)
		MSEn	Multiscale entropy of the NN-interval series
		CD	Correlation dimension of the NN-interval series
		HFD	Higuchi fractal dimension of the NN-interval series
		KFD	Katz fractal dimension of the NN-interval series
		LZC	Lempel–Ziv complexity of the NN-interval series
		MFDFA_alpha1_Max	Maximum of the singularity spectrum from multifractal DFA
		MFDFA_alpha1_Fluct	Fluctuation range of the multifractal DFA singularity spectrum
		MFDFA_alpha1_Peak	Peak (mode) of the multifractal DFA singularity spectrum
	Morphology	Amplitude	Mean systolic peak amplitude across beats in the window
		Duration	Mean beat duration (onset-to-onset) across beats
EDA	Phasic	Peaks_N	Number of skin conductance responses (SCRs) detected in the window
		Peaks_Amplitude_Mean	Mean amplitude of detected SCR peaks
ACC	Statistical	Mean	Mean of ACC_{MAG} in the window
		Std	Standard deviation of ACC_{MAG}
		Max	Maximum ACC_{MAG} value in the window
		Median	Median of ACC_{MAG}
		IQR	Interquartile range of ACC_{MAG}
		Skew	Skewness of ACC_{MAG} (asymmetry of the distribution)
		Kurtosis	Kurtosis of ACC_{MAG} (tail heaviness)
SKT	Statistical	Mean	Mean skin temperature in the window
		Std	Standard deviation of skin temperature
		Slope	Linear-regression slope (short-term thermal trend)

The extracted features were then filtered through a four-stage selection pipeline applied uniformly to each dataset. First, features with more than 50% missing values were removed and remaining gaps were filled by median imputation. Second, collinearity was reduced by discarding one member of every feature pair with $|\rho| > 0.8$ (Pearson). Third, a Shapiro–Wilk test confirmed widespread non-normality, and a three-way ANOVA model

$$\text{Feature} \sim C(\text{Label}) + C(\text{Subject}) + C(\text{Label}) : C(\text{Subject})$$

quantified the subject–label interaction. Fourth, the Mann–Whitney U test was applied subject-wise (stress vs. baseline); a feature was retained only if it reached significance ($p < 0.05$) in a minimum proportion of subjects, set proportionally to the dataset size (5 subjects for Wesad, VerBio and AffectiveRoad; 10 for Campanella). After selection, the four signal-specific matrices were concatenated into a single multimodal feature table per dataset. The complete set of unique features that survived the full extraction and selection pipeline is listed in Table 4.

3.3. Classification pipeline

Feature-driven machine learning. The FD paradigm treats stress classification as a supervised tabular-data problem. The multimodal feature vector assembled during preprocessing is fed directly to classical ML classifiers, and every discriminative pattern must already be encoded in those hand-crafted descriptors. This design makes the pipeline fully interpretable — each model's decision can be traced back to named physiological features — and requires comparatively little labelled data, at the cost of relying on the completeness and relevance of the upstream feature-engineering stage.

Six classifiers are compared within this paradigm. LR with ℓ_2 regularization ($C = 1.0$) serves as a linear baseline whose signed coefficients directly quantify the monotonic association of each feature with the log-odds of stress. RF aggregates 100 decorrelated

Table 5
Feature-driven classifiers: hyperparameter configurations and imbalance-handling strategy adopted for each algorithm.

Algorithm	Key hyperparameters	Imbalance handling
Logistic Regression	ℓ_2 penalty, $C = 1.0$, `max_iter`=1000	`class_weight=balanced`
Random Forest	100 trees, Gini, `max_depth`=None, $\sqrt{p}$ features	`class_weight=balanced`
Gradient Boosting	100 estimators, $\eta = 0.1$, `max_depth`=3	None
XGBoost	100 estimators, $\eta = 0.1$, `max_depth`=3, `subsample`=0.8, `colsample`=0.8	`scale_pos_weight`=$N_{\text{neg}}/N_{\text{pos}}$
SVM (RBF)	C=1.0, γ = scale, prob. calibration	`class_weight=balanced`

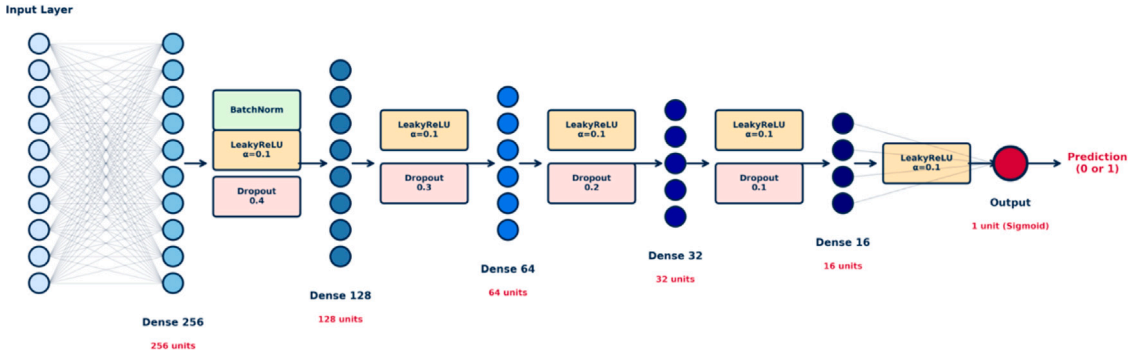


Fig. 3. Architecture of the feed-forward MLP used in the feature-driven branch. The network receives the concatenated multimodal feature vector and progressively reduces dimensionality through five hidden layers with LeakyReLU activations and decreasing dropout, ending with a single sigmoid output unit for binary stress classification.

trees (Gini impurity, $\sqrt{p}$ candidate features per split) to capture nonlinear thresholds while limiting variance. GB and extreme GB (XGB) both build sequential ensembles of shallow trees (depth 3, learning rate 0.1, 100 estimators); XGB adds row- and column-subsampling (both 0.8) and explicit regularization for improved generalization. SVM with a radial basis funzion (RBF) kernel ($C = 1.0$, γ = scale) maximizes the margin in a nonlinearly transformed space, proving robust in noisy, heterogeneous settings. Finally, a shallow multi-layer perceptron (MLP) (Fig. 3) closes the gap towards representation learning while still operating on the same tabular input: five hidden layers of decreasing width (256–128–64–32–16 units) with batch normalization, LeakyReLU activations ($\alpha = 0.1$), and progressively decreasing dropout (0.4–0.3–0.2–0.1) are optimized with Adam and binary cross-entropy loss. A summary of the hyperparameter settings adopted for each classifier is provided in Table 5.

Data-driven deep learning. The DD paradigm consists of a 1D multi-input CNN that receives the cleaned, windowed raw signals and learns discriminative temporal representations end-to-end. It captures subtle waveform motifs — such as pulse morphology, micro-conductance transients, and acceleration patterns — that are difficult to encode manually, at the cost of higher data requirements and reduced interpretability. The architecture (Fig. 4) follows a multi-stream, late-fusion design. Each of the four physiological modalities (BVP, EDA, SKT, ACC) is processed by an independent temporal stream comprising three successive convolutional blocks that perform hierarchical feature extraction while progressively reducing temporal resolution. Block 1 applies 32 filters with a wide kernel to capture coarse, slow dynamics, followed by batch normalization, ReLU, max pooling, and spatial dropout (rate 0.2). Block 2 narrows the kernel and doubles the filter count to 64 for intermediate-scale patterns, with the same structure and dropout rate of 0.3. Block 3 applies the narrowest kernel with 128 filters to resolve fine-grained temporal detail, again followed by BatchNorm, ReLU, max pooling, and dropout 0.3. All convolutional layers use He initialization and ℓ_2 weight decay ($\lambda = 10^{-3}$). Each stream ends with global average pooling, which collapses the temporal axis into a fixed-length descriptor.

The four pooled outputs are concatenated (late fusion) and fed to a two-block classification head. The first block comprises a fully connected layer with 64 units, batch normalization, ReLU, and dropout 0.4 (with ℓ_2 regularization $\lambda = 10^{-3}$); the second mirrors this design with 32 units and dropout 0.3. A final sigmoid unit outputs the stress probability. Training uses focal loss ($\gamma = 2.0$, $\alpha = 0.25$) to down-weight easy negatives, per-fold class weights to counter imbalance, and Adam optimization ($\eta = 10^{-3}$) with early stopping (patience 5) and ReduceLROnPlateau scheduling (factor 0.7, patience 3).

Training and validation. Generalization across individuals is assessed through a LOSO cross-validation protocol. In each fold, all windows belonging to a single held-out subject form the test set while the remaining subjects constitute the training set. All preprocessing transformations — feature scaling, normalization, and (for the DD branch) per-signal min–max calibration — are fitted exclusively on the training partition and applied unchanged to the held-out subject, thereby preventing any information leakage. Predictions across folds are concatenated to produce global performance summaries.

Class imbalance is addressed without resampling. In the FD branch, learners that support native class weights use `class_weight=balanced`; XGB compensates via `scale_pos_weight` = $N_{\text{neg}}/N_{\text{pos}}$ on each training split. In the DD branch, per-fold class weights are passed to the focal-loss objective. A subject-level class-balance screening is applied prior to evaluation: held-out folds in which the positive-class (stress) prevalence falls below a threshold $\tau = 5\%$ are excluded to avoid degenerate, near-single-class

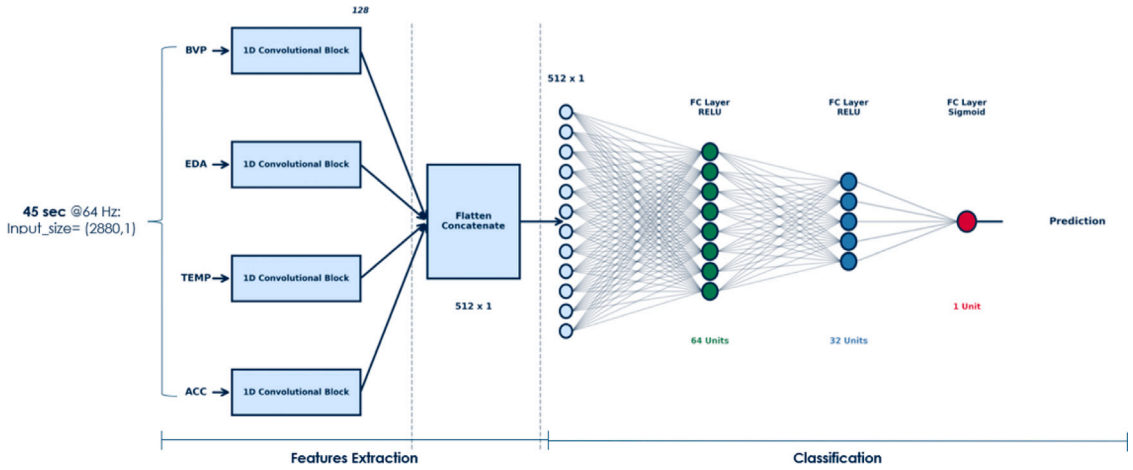


Fig. 4. Architecture of the 1D multi-input CNN used in the data-driven branch. Four independent per-modality streams (BVP, EDA, SKT, ACC) each contain three convolutional blocks with progressively narrower kernels and increasing filter counts (32–64–128). Global average pooling produces a fixed-length descriptor per stream; late concatenation feeds a two-layer classifier head ending with a sigmoid output for binary stress prediction. Input windows are 45 s at 64 Hz, yielding tensors of shape (2880, 1) per modality.

test conditions. Decision thresholds for discrete predictions are selected per fold by maximizing the Youden's J statistic on the ROC curve derived from validation scores within that fold.

Evaluation strategies. Model performance is assessed through a three-stage experimental protocol designed to evaluate within-dataset accuracy, cross-dataset generalizability, and the benefit of transfer learning. In the first stage, for each of the four datasets, all classifiers are trained and validated under the LOSO protocol described previously. Performance is quantified via macro-averaged F1-score, precision, recall, and AUC, establishing an in-domain baseline for each model on each corpus. The second stage assesses out-of-domain generalization by applying models trained on a source dataset $D^{(s)}$ directly to a distinct target dataset $D^{(t)}$ without any retraining. This process requires identical feature spaces across corpora. For the FD branch, we adopt the feature subset produced by the Wesad preprocessing pipeline as the common reference, re-extracting all four datasets by retaining, per modality, exactly the features that survived Wesad's selection stage, and retraining all FD models on this harmonized feature space. For each dataset and each algorithm, the LOSO fold with the highest macro-F1 is selected as the source checkpoint, together with its fitted preprocessing transformers (scalers, normalizers). This best-performing model is then applied directly to the target dataset: input windows from $D^{(t)}$ are transformed using the source preprocessing, and predictions are evaluated with the same metrics used in-domain (macro-F1 as primary; precision, recall, ROC-AUC as secondary). The performance drop quantifies the model's ability to generalize across protocol heterogeneity, stressor mismatch, and label noise. For the DD branch, the same zero-shot protocol is applied: the CNN checkpoint with the highest LOSO F1-score on $D^{(s)}$ is evaluated on $D^{(t)}$ using the source per-signal scalers, with a fixed inference threshold $\tau = 0.5$ to maintain comparability. The third stage determines whether limited target-domain supervision can mitigate the cross-dataset performance drop by fine-tuning the best source models on $D^{(t)}$ under an adaptation protocol whose layer-freezing depth was selected through a dedicated sensitivity analysis (Supplementary Material, Table 14, Appendix A.3). For the FD branch, *full fine-tuning* is adopted: all 6 layers of the MLP are left trainable; inputs are transformed with the source scaler without refitting; optimization uses Adam with gradient clipping (`clipnorm`=1.0), binary cross-entropy loss, learning rate $\eta = 10^{-4}$, mini-batches of size 16, and at most 10 epochs; early stopping monitors validation loss with patience = 5, minimum improvement $\Delta = 10^{-3}$, and `restore_best_weights`=True; a `ReduceLROnPlateau` scheduler halves the rate on stagnation (factor 0.5, patience = 3, floor 10^{-7}). For the DD branch, *Block 3 + Head* is adopted: the first two convolutional blocks of every per-modality stream (8 Conv1D frozen layers) are held fixed, while the third convolutional block (4 Conv1D layers, one per stream) and the classification head (2 dense hidden layers and 1 output layer) are re-optimized; the same optimizer, learning rate $\eta = 10^{-4}$, epoch cap, early stopping, and scheduler as the FD branch are used, with gradient clipping `clipnorm` = 0.5; per-stream scalers from the source are applied unchanged; decision thresholds are selected per fold by maximizing the Youden's J statistic on the validation ROC curve. In both cases, the LOSO partition of $D^{(t)}$ is always respected: each fold trains on target-training windows and evaluates on the held-out target subject. Balanced class weights are applied when both classes are present. This stage measures whether representation learning (CNN) or decision-boundary adaptation (MLP) can recover performance lost in zero-shot transfer. All three stages use a consistent set of metrics to enable direct comparison. The macro-averaged F1-score is the primary metric, chosen because it weights both classes (stress and no-stress) equally and remains robust under class imbalance, which varies substantially across datasets and subjects. Macro-averaged precision and recall are reported as secondary metrics to characterize the precision–recall trade-off. ROC curves and their summary statistic AUC are computed for all models to assess discrimination independently of the operating-point choice. Confusion matrices provide a detailed breakdown of true/false positives and negatives, and classification reports aggregate per-class metrics. All metrics are expressed as mean ± standard deviation across the eligible LOSO folds (after subject-level class-balance

screening at $\tau = 5\%$). This unified evaluation framework allows systematic comparison of within-dataset performance, zero-shot generalization, and transfer-learning gain across both the FD and DD paradigms.

All experiments were conducted on Google Colaboratory (Colab), running on a Python 3 Google Compute Engine backend (RAM: 12.7 GB, Disk storage: 107.7 GB). Data ingestion, manipulation, and tabular operations were handled with `pandas` and `NumPy`. Signal filtering, resampling, spectral analysis, and statistical computations were performed using `SciPy`. Physiological feature extraction relied on `NeuroKit2` (Makowski et al., 2021). Statistical tests were implemented via `statsmodels`. Feature-driven classification algorithms were implemented using `scikit-learn`, while neural networks were implemented and trained using `TensorFlow/Keras`. All random seeds were set to 42 to ensure reproducibility. Result visualization was produced with `Matplotlib` and `Seaborn` libraries.

4. Results and discussion

4.1. Within-dataset performance

Table 6 reports LOSO performance after subject-level class-balance screening at $\tau = 5\%$. Two consistent patterns emerge. First, Wesad and Campanella achieve substantially higher performance (best macro-F1 0.82–0.96) than VerBio and AffectiveRoad (0.63–0.74), reflecting differences between laboratory controlled protocols with phase-aligned labels versus ecological/rating-based data collection. Despite the 5% prevalence threshold, AffectiveRoad and VerBio retain substantial inter-subject heterogeneity and residual imbalance, translating into larger standard deviations. Fig. 5 provides a visual comparison of all classifiers across the four datasets, confirming that no single algorithm dominates uniformly. The CNN exhibits the most asymmetric profile — excelling on Wesad and VerBio while falling below ensemble methods on Campanella — whereas tree-based models form a compact cluster with more balanced cross-dataset behaviour. Second, nonlinear learners consistently outperform linear baselines. Among FD models, gradient-boosted ensembles dominate in controlled settings: XGBoost achieves the highest macro-F1 on Wesad (0.866 ± 0.114), while MLP reaches the FD peak on Campanella (0.819 ± 0.039) with high recall (0.900 ± 0.079). The RBF-SVM proves particularly robust in noisier domains, attaining the best FD macro-F1 on AffectiveRoad (0.626 ± 0.083) and VerBio (0.627 ± 0.248), where maximizing the margin mitigates label noise and residual imbalance. The CNN extends these trends by learning temporal representations end-to-end. On Wesad, it achieves near-ceiling performance (macro-F1 0.958 ± 0.084, AUC 0.98), with precision close to 0.99, indicating that temporal filters capture discriminative waveform structure. The CNN improves operating points on all other datasets (VerBio 0.742, AffectiveRoad 0.693, Campanella 0.743), though with precision-favouring asymmetry on ecological data. Overall, tree ensembles excel with clean labels, SVM provides robustness under noise, and the CNN offers the largest gains where temporal regularities are reliable. Detailed ROC curves for FD and DD approaches are reported in Appendix A.4, separately for each dataset. Per-subject F1-score distributions are provided in Fig. 16 (Appendix A.4), confirming the high inter-subject variability on VerBio and AffectiveRoad highlighted by the bootstrap confidence intervals in Table 6.

4.2. Cross-dataset performance

Table 7 presents unified zero-shot results. Rows are targets, columns are sources, and the diagonal shows in-domain performance. Fig. 6 provides a condensed heatmap view of the same results, reporting the best FD classifier and the CNN per source–target pair to highlight cross-dataset degradation patterns at a glance. Within each target block, the best off-diagonal score is highlighted in bold. Zero-shot cross-test exposes sizeable domain gaps across the board, with most off-diagonal cells showing substantial drops relative to the diagonal. Nevertheless, certain laboratory-to-laboratory directions prove relatively resilient: Wesad↔Campanella transfers maintain moderate performance (LR achieves 0.719 and 0.768), suggesting that controlled protocols with phase-aligned labels yield more portable representations. Within the FD family, Campanella emerges as the most transferable source, repeatedly yielding the highest off-diagonal scores towards Wesad (XGBoost 0.630), VerBio (XGBoost 0.589), and AffectiveRoad (RF 0.668). Tree ensembles (GB, XGBoost, RF) deliver the best zero-shot performance on most targets, confirming their robustness to distribution shift when operating on hand-crafted features. LR and MLP both exhibit complete collapses (macro-F1 = 0.000) exclusively when VerBio acts as the source dataset. Three concurring factors explain this pattern. First, VerBio is the most class-imbalanced corpus in this study (71% No Stress/29% Stress at the windowed feature level), inducing a strong majority-class bias in both classifiers. Second, VerBio stress labels are derived from a continuous multi-rater fused score thresholded at 0.2 on a $[0, 1]$ scale. At this low value, labelled stress episodes correspond to mild perceived arousal rather than strong and discrete autonomic responses — the physiological difference between baseline and stress phases is subtle and gradual, in stark contrast to the abrupt sympathetic activation produced by laboratory stressors such as TSST (Wesad) and cognitive tasks (Campanella). Third, models trained on VerBio therefore learn a weakly separating decision boundary calibrated on subtle physiological patterns that fails to recognize the strong, discrete stress responses of laboratory-protocol targets, producing the observed collapse. These collapses are fully resolved by Full Fine-tune transfer learning. The CNN, evaluated strictly out-of-domain without fine-tuning, shows larger drops on pairs combining social-evaluative or rating-derived labels with motion-rich acquisition. For example, the CNN achieves only 0.361 macro-F1 when trained on Wesad and tested on VerBio, and 0.268 on Wesad→AffectiveRoad. This reflects the CNN's greater reliance on temporal pattern alignment, with distributional shifts in waveform morphology and noise characteristics degrading end-to-end performance more severely than hand-crafted feature-based models. In practical terms, when no target labels are available, the FD pipeline — especially tree ensembles trained on Campanella or Wesad — provides safer cross-dataset deployment.

Table 6
Within-dataset LOSO performance. Primary metric: macro-F1 (mean ± SD [95% bootstrap CI]; B=10,000). Best score per dataset in bold.

Dataset	Algorithm	F1-score	Precision	Recall
AffectiveRoad	Logistic Regression	0.618 ± 0.130 [0.542, 0.682]	0.657 ± 0.164 [0.570, 0.744]	0.637 ± 0.185 [0.527, 0.727]
	Random Forest	0.620 ± 0.067 [0.584, 0.657]	0.638 ± 0.180 [0.585, 0.778]	0.611 ± 0.111 [0.551, 0.671]
	Gradient Boosting	0.577 ± 0.126 [0.501, 0.638]	0.628 ± 0.194 [0.521, 0.734]	0.577 ± 0.170 [0.480, 0.664]
	XGBoost	0.604 ± 0.090 [0.556, 0.653]	0.651 ± 0.178 [0.555, 0.749]	0.611 ± 0.144 [0.531, 0.686]
	SVM (RBF)	0.626 ± 0.083 [0.580, 0.670]	0.685 ± 0.150 [0.604, 0.766]	0.602 ± 0.115 [0.541, 0.664]
	MLP	0.601 ± 0.122 [0.585, 0.684]	0.703 ± 0.145 [0.622, 0.757]	0.551 ± 0.160 [0.541, 0.687]
	CNN	**0.693 ± 0.109** [0.632, 0.749]	0.769 ± 0.103 [0.716, 0.827]	0.658 ± 0.178 [0.562, 0.753]
Campanella	Logistic Regression	0.747 ± 0.120 [0.701, 0.787]	0.816 ± 0.112 [0.777, 0.857]	0.729 ± 0.191 [0.659, 0.795]
	Random Forest	0.806 ± 0.049 [0.788, 0.823]	0.769 ± 0.072 [0.744, 0.796]	0.862 ± 0.106 [0.823, 0.899]
	Gradient Boosting	0.805 ± 0.052 [0.787, 0.824]	0.776 ± 0.082 [0.746, 0.805]	0.854 ± 0.110 [0.814, 0.893]
	XGBoost	0.812 ± 0.055 [0.791, 0.830]	0.781 ± 0.078 [0.754, 0.809]	0.861 ± 0.111 [0.820, 0.901]
	SVM (RBF)	0.778 ± 0.069 [0.754, 0.803]	0.818 ± 0.100 [0.782, 0.854]	0.764 ± 0.132 [0.715, 0.810]
	MLP	**0.819 ± 0.039** [0.812, 0.835]	0.759 ± 0.071 [0.745, 0.797]	0.900 ± 0.079 [0.868, 0.925]
	CNN	0.743 ± 0.191 [0.667, 0.803]	0.866 ± 0.080 [0.838, 0.896]	0.701 ± 0.216 [0.617, 0.772]
VerBio	Logistic Regression	0.610 ± 0.246 [0.492, 0.730]	0.575 ± 0.305 [0.425, 0.726]	0.773 ± 0.233 [0.650, 0.873]
	Random Forest	0.530 ± 0.200 [0.432, 0.629]	0.661 ± 0.329 [0.499, 0.818]	0.532 ± 0.215 [0.429, 0.637]
	Gradient Boosting	0.588 ± 0.243 [0.466, 0.703]	0.661 ± 0.335 [0.493, 0.821]	0.617 ± 0.228 [0.499, 0.723]
	XGBoost	0.589 ± 0.221 [0.477, 0.694]	0.647 ± 0.322 [0.486, 0.805]	0.637 ± 0.221 [0.524, 0.739]
	SVM (RBF)	0.627 ± 0.248 [0.502, 0.742]	0.565 ± 0.295 [0.419, 0.707]	0.815 ± 0.233 [0.694, 0.917]
	MLP	0.588 ± 0.255 [0.468, 0.705]	0.608 ± 0.315 [0.467, 0.774]	0.655 ± 0.259 [0.523, 0.768]
	CNN	**0.742 ± 0.179** [0.612, 0.856]	0.767 ± 0.146 [0.672, 0.868]	0.777 ± 0.250 [0.588, 0.920]
Wesad	Logistic Regression	0.806 ± 0.113 [0.749, 0.859]	0.771 ± 0.133 [0.704, 0.835]	0.883 ± 0.167 [0.794, 0.954]
	Random Forest	0.820 ± 0.182 [0.724, 0.898]	0.872 ± 0.144 [0.798, 0.937]	0.823 ± 0.223 [0.701, 0.916]
	Gradient Boosting	0.847 ± 0.118 [0.788, 0.904]	0.845 ± 0.161 [0.762, 0.917]	0.885 ± 0.151 [0.807, 0.952]
	XGBoost	0.866 ± 0.114 [0.810, 0.918]	0.868 ± 0.157 [0.787, 0.939]	0.892 ± 0.133 [0.821, 0.950]
	SVM (RBF)	0.824 ± 0.115 [0.767, 0.879]	0.781 ± 0.157 [0.705, 0.857]	0.913 ± 0.149 [0.829, 0.973]
	MLP	0.846 ± 0.136 [0.794, 0.919]	0.867 ± 0.113 [0.804, 0.924]	0.859 ± 0.192 [0.791, 0.946]
	CNN	**0.958 ± 0.084** [0.911, 0.987]	0.988 ± 0.024 [0.975, 0.998]	0.939 ± 0.126 [0.867, 0.981]

Table 7
Unified cross-dataset results (macro-F1 ± SD). Rows are targets, columns are sources. Diagonal shows in-domain performance. Best off-diagonal score per target in bold.

Target	Algorithm	Wesad	Campanella	VerBio	AffectiveRoad
Wesad	Logistic Regression	0.806 ± 0.113	**0.719 ± 0.156**	0.000 ± 0.000	0.247 ± 0.267
	Random Forest	0.820 ± 0.182	0.546 ± 0.092	0.272 ± 0.196	0.164 ± 0.130
	Gradient Boosting	0.847 ± 0.118	0.573 ± 0.209	0.558 ± 0.258	0.248 ± 0.210
	XGBoost	0.866 ± 0.114	0.630 ± 0.142	0.553 ± 0.217	0.197 ± 0.194
	MLP	0.862 ± 0.130	0.573 ± 0.096	0.000 ± 0.000	0.303 ± 0.205
	CNN	0.958 ± 0.084	0.537 ± 0.189	0.501 ± 0.321	0.459 ± 0.237
Campanella	Logistic Regression	**0.768 ± 0.078**	0.759 ± 0.127	0.000 ± 0.000	0.537 ± 0.214
	Random Forest	0.520 ± 0.161	0.813 ± 0.046	0.281 ± 0.167	0.369 ± 0.202
	Gradient Boosting	0.591 ± 0.151	0.812 ± 0.055	0.563 ± 0.130	0.366 ± 0.134
	XGBoost	0.520 ± 0.167	0.810 ± 0.057	0.590 ± 0.120	0.284 ± 0.179
	MLP	0.694 ± 0.132	0.828 ± 0.048	0.000 ± 0.000	0.604 ± 0.163
	CNN	0.576 ± 0.255	0.743 ± 0.191	0.569 ± 0.269	0.588 ± 0.246
VerBio	Logistic Regression	0.477 ± 0.153	0.477 ± 0.153	0.590 ± 0.261	0.477 ± 0.153
	Random Forest	0.439 ± 0.242	0.485 ± 0.161	0.540 ± 0.245	0.280 ± 0.166
	Gradient Boosting	0.469 ± 0.228	0.326 ± 0.247	0.555 ± 0.253	0.241 ± 0.162
	XGBoost	0.459 ± 0.204	**0.585 ± 0.184**	0.537 ± 0.261	0.309 ± 0.217
	MLP	0.478 ± 0.152	0.477 ± 0.153	0.539 ± 0.270	0.478 ± 0.153
	CNN	0.361 ± 0.324	0.494 ± 0.279	0.742 ± 0.179	0.459 ± 0.316
AffectiveRoad	Logistic Regression	0.591 ± 0.184	0.548 ± 0.176	0.000 ± 0.000	0.620 ± 0.130
	Random Forest	0.327 ± 0.291	**0.668 ± 0.121**	0.154 ± 0.128	0.579 ± 0.100
	Gradient Boosting	0.413 ± 0.311	0.566 ± 0.153	0.215 ± 0.178	0.588 ± 0.131
	XGBoost	0.383 ± 0.323	0.584 ± 0.152	0.291 ± 0.191	0.603 ± 0.111
	MLP	0.476 ± 0.270	0.660 ± 0.114	0.000 ± 0.000	0.611 ± 0.125
	CNN	0.268 ± 0.317	0.279 ± 0.224	0.093 ± 0.137	0.693 ± 0.109

4.3. Transfer learning results

The initial transfer learning strategy, selected on an empirical basis, adopted a Last-Only configuration in which only the last layer was left trainable for both the MLP and the CNN. Table 8 reports the results obtained under this setting. Several criticalities

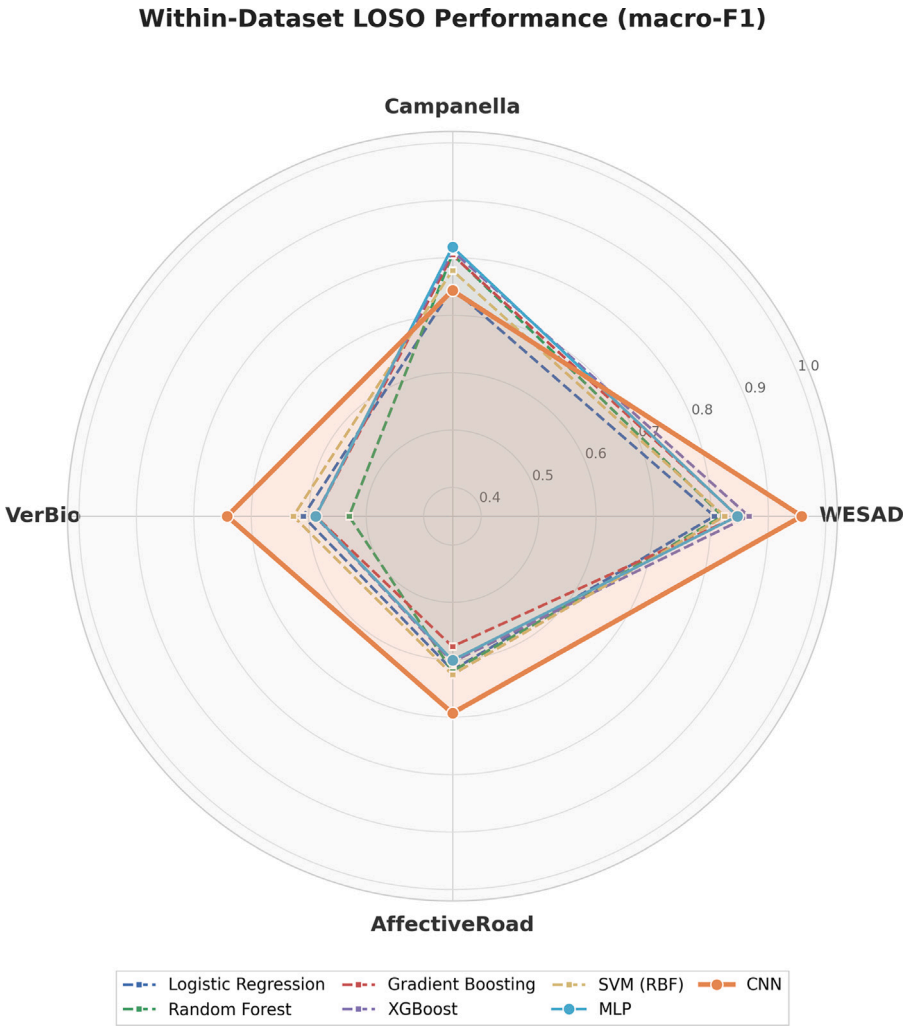
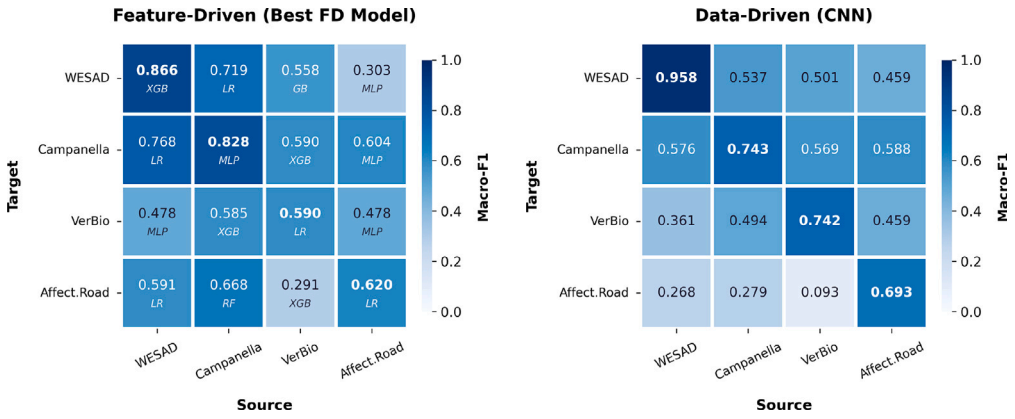


Fig. 5. Radar chart of within-dataset LOSO macro-F1 across all classifiers and datasets. Each axis corresponds to a dataset; each polygon represents one algorithm. The CNN exhibits the widest reach on Wesad and VerBio but contracts on Campanella, whereas ensemble methods (XGB, GB) maintain more uniform profiles.

emerge. First, the MLP retains the complete class collapses (macro-F1 = 0.000) already observed in the zero-shot evaluation on all three VerBio-sourced directions (VerBio→Wesad, VerBio→Campanella, VerBio→AffectiveRoad), indicating that adapting the last layer alone is insufficient to overcome the majority-class bias encoded in the frozen hidden representations. Second, for the MLP the gains over the cross-dataset zero-shot baseline (Table 7) are inconsistent: several pairs show negligible improvement (e.g., Campanella→Wesad: from 0.573 to 0.593, +0.020) or slight degradation (e.g., Wesad→Campanella: from 0.694 to 0.674). Third, the CNN achieves more substantial recovery on selected directions — most notably VerBio→Wesad, which rises from 0.501 to 0.803 (+0.302) — yet the improvement is not systematic across all targets. When comparing the best Last-Only result per target against the best zero-shot cross-dataset model regardless of algorithm, the Last-Only protocol surpasses the cross-dataset best only on Wesad (0.803 vs. 0.719, +0.084); on Campanella the gain is marginal (0.771 vs. 0.768, +0.003), on VerBio modest (0.646 vs. 0.585, +0.061), and on AffectiveRoad the best Last-Only result (0.655) falls below the zero-shot RF baseline (0.668, −0.013). The additional computational cost of Last-Only transfer learning is therefore not fully justified across targets.

Following the sensitivity analysis described in Section 3.3 (full results in Supplementary Material, Table 14, Appendix A.3), two optimized configurations were identified: Full Fine-tune for the MLP (all 6 layers trainable) and Block 3 + Head for the CNN (7 trainable layers out of 15). Table 9 presents the results obtained under these configurations. The improvement over Last-Only is substantial across both paradigms. For the MLP, Full Fine-tune eliminates all class collapses on VerBio as source (VerBio→Wesad: from 0.000 to 0.805; VerBio→Campanella: from 0.000 to 0.814; VerBio→AffectiveRoad: from 0.000 to 0.562) and produces systematic gains across all 12 source–target pairs, with an average off-diagonal improvement of +0.34 macro-F1 relative to Last-Only, at a marginal computational overhead (~9 s vs. ~5 s per fold on CPU). For the CNN, Block 3 + Head yields consistent



(a) Feature-driven (Best FD model). (b) Data-driven (CNN).

Fig. 6. Heatmap of zero-shot cross-dataset macro-F1. Rows are targets, columns are sources; the diagonal shows within-dataset performance. (a) Best feature-driven classifier per source–target pair, with the winning algorithm in each cell. (b) Data-driven CNN. Darker blue indicates higher macro-F1.

Table 8
Unified transfer learning results (macro-F1 ± SD). Rows are targets with the *Algorithm* column distinguishing feature-driven MLP vs. data-driven CNN. The light-green diagonal shows in-domain performance; within each target, the best *cross-dataset* score is in **bold**.

Target	Algorithm	Wesad	Campanella	VerBio	AffectiveRoad
Wesad	MLP	0,8616 ± 0,1296	0,5934 ± 0,0934	0,0000 ± 0,0000	0,3022 ± 0,2044
	CNN	0,9575 ± 0,0841	0,7854 ± 0,1357	**0,8030 ± 0,2035**	0,6780 ± 0,1558
Campanella	MLP	0,6742 ± 0,1425	0,8276 ± 0,0477	0,0000 ± 0,0000	0,5764 ± 0,1731
	CNN	0,7681 ± 0,1330	0,7434 ± 0,1910	0,7664 ± 0,1907	**0,7713 ± 0,1516**
VerBio	MLP	0,4766 ± 0,1505	0,4766 ± 0,1534	0,5385 ± 0,2696	0,4785 ± 0,1530
	CNN	0,5942 ± 0,2517	0,6368 ± 0,2533	0,7418 ± 0,1791	**0,6458 ± 0,2321**
AffectiveRoad	MLP	0,5060 ± 0,2406	0,6295 ± 0,1351	0,0000 ± 0,0000	0,6110 ± 0,1251
	CNN	0,4970 ± 0,2754	**0,6550 ± 0,1373**	0,5088 ± 0,2992	0,6934 ± 0,1089

Table 9
Transfer learning results with optimized configurations (macro-F1 ± SD). MLP: Full Fine-tune (all 6 layers trainable); CNN: Block 3 + Head (7 trainable layers out of 15). The light-green diagonal shows in-domain performance; within each target, the best off-diagonal score is in bold.

Target	Algorithm	Wesad	Campanella	VerBio	AffectiveRoad
Wesad	MLP	0,862 ± 0,130	0,779 ± 0,122	**0,805 ± 0,118**	0,794 ± 0,103
	CNN	0,958 ± 0,084	0,929 ± 0,088	**0,935 ± 0,098**	0,892 ± 0,115
Campanella	MLP	0,775 ± 0,115	0,828 ± 0,048	**0,814 ± 0,041**	0,749 ± 0,135
	CNN	0,746 ± 0,192	0,743 ± 0,191	0,647 ± 0,245	**0,784 ± 0,136**
VerBio	MLP	**0,638 ± 0,219**	0,614 ± 0,264	0,539 ± 0,270	0,491 ± 0,338
	CNN	0,636 ± 0,259	**0,736 ± 0,186**	0,742 ± 0,179	0,671 ± 0,289
AffectiveRoad	MLP	0,616 ± 0,120	**0,667 ± 0,110**	0,562 ± 0,289	0,611 ± 0,125
	CNN	0,597 ± 0,186	**0,653 ± 0,139**	0,599 ± 0,143	0,693 ± 0,109

gains, with the most impressive improvements towards Wesad as target: Campanella→Wesad rises from 0.785 to 0.929 (+18%), VerBio→Wesad from 0.803 to 0.935 (+16%), and AffectiveRoad→Wesad from 0.678 to 0.892 (+32%); the average off-diagonal improvement is +0.29 macro-F1. Two directions towards Campanella as target show a slight decrease (VerBio→Campanella: −0.119; Wesad→Campanella: −0.022), suggesting that for this target the Classifier Head was already sufficient and unfreezing the third convolutional block introduces mild overfitting.

Comparing the optimized transfer learning against the best zero-shot cross-dataset model per target — regardless of algorithm, including tree ensembles and LR which are not amenable to fine-tuning — the adapted models achieve substantial gains on most targets: WESAD rises from 0.719 (LR, Campanella source) to 0.935 (CNN, VerBio source; +30%); VerBio from 0.585 (XGBoost, Campanella source) to 0.736 (CNN, Campanella source; +26%); and Campanella from 0.768 (LR, WESAD source) to 0.814 (MLP,

VerBio source; +6%). On AffectiveRoad, however, the best transfer learning result (MLP, Campanella source: 0.667) is essentially equivalent to the zero-shot RF baseline (0.668), indicating that on this particular target the additional computational cost of adaptation does not yield a practical advantage over a simple zero-shot deployment with tree-based models. In several cases, the optimized transfer learning not only recovers the domain-shift drop but surpasses the within-dataset performance reported on the diagonal of Table 9. For the MLP, this occurs on VerBio as target — WESAD→VerBio reaches 0.638 vs. a within-dataset baseline of 0.539 (+0.099), Campanella→VerBio 0.614 (+0.075) — and on AffectiveRoad as target, where Campanella→AffectiveRoad achieves 0.667 vs. 0.611 (+0.056). For the CNN, AffectiveRoad→Campanella reaches 0.784 against a within-dataset baseline of 0.743 (+0.041). These surpluses concentrate on targets characterized by high inter-subject variability and weaker within-dataset baselines, where a model pre-trained on a source corpus with stronger and more consistent physiological stress patterns provides a better initialization than training from scratch on the noisy target alone. Directly comparing the feature-driven (MLP) and data-driven (CNN) paradigms under transfer learning reveals a marked shift in relative competitiveness depending on the freezing configuration. Under Last-Only, the CNN dominates on 11 out of 12 off-diagonal directions with an average gap of +0.283 macro-F1, making it the only viable paradigm for adaptation; the MLP, constrained to a single trainable layer, cannot reshape its internal representations and retains all class collapses. With the optimized configurations, Full Fine-tune narrows this gap dramatically: the CNN still prevails on 7 out of 12 directions, but the average difference shrinks to +0.043—an 85% reduction. The CNN retains a clear advantage on laboratory targets, particularly Wesad (average off-diagonal CNN 0.919 vs. MLP 0.793, +0.126), where its convolutional features adapt effectively to strong temporal stress patterns. The MLP becomes competitive or superior on Campanella as target (average off-diagonal MLP 0.779 vs. CNN 0.726, +0.053—most notably VerBio→Campanella: MLP 0.814 vs. CNN 0.647) and on AffectiveRoad (MLP 0.615 vs. CNN 0.616, essentially tied). On VerBio as target, the CNN leads (0.681 vs. 0.581, +0.100). In summary, CNN Block 3 + Head provides the highest average cross-dataset performance (0.735 vs. 0.692) and is the preferred choice when target supervision and computational resources are available, particularly for laboratory targets where the gap is largest. MLP Full Fine-tune offers a credible alternative at substantially lower computational cost (∼9 s vs. ∼340 s per fold) and is the only paradigm that consistently surpasses within-dataset performance on ecological targets.

4.4. Comparison with similar studies

Table 10 highlights that, while prior work has reported strong within-dataset performance (often F1 > 0.85), cross-dataset generalization remains substantially weaker, with typical F1 scores dropping to the 0.55–0.75 range and transfer learning rarely evaluated. In this context, the proposed approach achieves competitive in-domain performance (F1 = 0.96) while also reaching F1 = 0.77 in the cross-dataset setting, placing it among the best-performing methods under zero-shot conditions. Importantly, the observed performance drop between intra- and cross-dataset evaluations (≈0.19) is smaller than that reported in most prior studies, indicating improved robustness to distribution shifts. In the transfer learning scenario, our method attains F1 = 0.80, matching or exceeding the limited number of works that report comparable results. Beyond performance metrics, model efficiency further distinguishes the proposed approach. None of the compared deep learning studies reports explicit parameter counts or computational cost, limiting direct complexity comparisons. Based on the available architectural descriptions, prior CNN-based models range from approximately 6M parameters (Gil-Martín et al., 2022) to lightweight designs of around 34K parameters (Almadhor et al., 2025). In contrast, our MLP (≈53K parameters) and multi-stream CNN (∼180K parameters) remain in a compact regime, while still achieving competitive or superior generalization. In particular, the proposed CNN attains comparable cross-dataset performance with roughly 30× fewer parameters than larger CNN architectures, demonstrating that strong generalization does not require high model complexity and supporting suitability for deployment on resource-constrained wearable devices.

In addition, unlike most prior studies that focus on either within-dataset or limited cross-domain evaluation, this work provides a unified and systematic assessment across intra-dataset, cross-dataset, and transfer learning settings using four heterogeneous corpora under a consistent subject-independent protocol. Only a small subset of existing works evaluate transfer learning at all, and even fewer combine zero-shot and fine-tuning analyses within the same framework. The proposed evaluation considers both paradigms, offering a more complete and reliable picture of model generalizability under real-world distribution shifts.

4.5. Limitations

Several limitations qualify these findings. The evaluation covers four datasets which, despite their heterogeneity, do not exhaust the full spectrum of real-world stressors and ecological contexts. Label definitions differ across corpora — continuous rating-based annotations in VerBio and AffectiveRoad versus phase-discrete labels in Wesad and Campanella — and residual label mismatch can confound cross-dataset comparisons. Subject cohorts are modest in size (13–29 participants per dataset) and unbalanced across conditions, potentially inflating LOSO variance estimates. Evaluation emphasized macro-F1; calibration quality, decision-cost curves, and temporal detection latency were not assessed. The CNN family lacked self-supervised pretraining or domain generalization objectives; alternative architectures might yield different trade-offs. Finally, motion artefacts were addressed only implicitly through accelerometry features, with no explicit activity-recognition stage to disentangle physical movement from physiological stress responses. This limitation is particularly relevant for AffectiveRoad, where wrist movements during driving (steering, gear changes) can produce artefacts on BVP and ACC signals that overlap with stress-related autonomic responses. The absence of a dedicated motion-rejection or activity-recognition module may therefore attenuate the ecological validity of the performance estimates on driving-scenario data and, more broadly, on any deployment context involving sustained physical activity. Future systems targeting real-world deployment should integrate an explicit artefact-detection stage upstream of the stress classifier. Real-time

Table 10
Overview of selected studies on wearable stress detection. Reported values are the best achieved macro-F1 for each evaluation setting. Acc* indicates that only accuracy was reported. A dash (–) means the evaluation was not performed.

Article	Signals	Datasets	Method	F1 Intra	F1 Cross	F1 TL
Benchekroun et al. (2023)	ECG, EDA, RESP	Wesad, AffectiveRoad	Classical ML	>0.80	0.61	–
Prajod et al. (2024)	ECG, BVP	Wesad, SWELL-KW, VerBIO, ForDigitStress	Classical ML	0.86	0.55	–
Vos et al. (2022)	EDA, HR	SWELL, NEURO, Wesad, UBFC-Phys	Ensemble (GB + ANN)	0.85	0.75	–
Albaladejo González et al. (2022)	HR	Wesad, SWELL-KW	ML (MLP, LOF)	0.99	<0.60	0.55
Gil-Martín et al. (2022)	ECG, EDA, EMG, RESP, TEMP, ACC	Wesad	DL (CNN)	0.90	–	–
Abdelfattah et al. (2025)	ECG, EDA, EMG, RESP, TEMP, ACC	Wesad	DL (DNN, CNN, RNN)	0.93	–	–
Xiang et al. (2025a)	ACC, EDA, HR, TEMP	Custom (nurses)	DL (CNN dual-branch)	0.91	–	–
Moser et al. (2024)	EDA, TEMP	Custom (Empatica E4)	DL (LSTM + GAN)	0.85	–	–
Almadhor et al. (2025)	ECG, EDA, EMG, RESP, TEMP, ACC	Wesad, ScientISST, DREAMER	DL (1D-CNN, Conformer)	0.98*	–	0.82*
This Study	BVP, EDA, TEMP, ACC	Wesad, AffectiveRoad, Campanella, VerBio	Feature-based ML, CNN	**0.96**	**0.77**	**0.80**

deployment was not investigated in this study. However, eventual online stress detection approaches do not pose strict real-time requirements, as the temporal resolution in stress prediction can be in the order from tens of seconds up to minutes, also considering delayed physiological responses. From the perspective of edge deployment, requirements are more related to memory usage and computational complexity. In this context, the implemented CNN contains approximately 180K trainable parameters, corresponding to a memory footprint of about 0.7 MB in 32-bit floating-point precision. When deployed, the model requires less than 20 MB of RAM to store intermediate feature maps during inference. Due to its relatively small size and moderate memory requirements, the model is well-suited for edge deployment. The application of additional optimization techniques such as 8-bit quantization can further reduce the storage requirement, making it compatible with resource-constrained embedded systems and low-power devices.

5. Conclusion and future work

This study aimed to evaluate the performance of ML algorithms for automatic stress detection based on physiological signals collected through wearable sensors. To provide a comprehensive performance and generalization estimate, multiple datasets were investigated, which differed in the subject cohort, experiments, tasks, and data collection setting. To mitigate confounding factors possibly generated by different sensing devices, we analysed only data collected using the same wearable sensor (i.e. Empatica E4). Similarly, we extracted the same feature set and applied the same pre-processing and classification algorithms to all the datasets. Diverse ML models were evaluated using a consistent subject-independent validation scheme, differentiating between FD and DD ML methods. Overall, this work addressed the complexity of automatic stress detection from multiple aspects, evaluating how different datasets, learning strategies, ML models, and evaluation schemes affect classification performance.

The analysis performed on each dataset separately revealed significant heterogeneity in classification performance, with differences in F-score in the range from 5 to 20%. This reflects the different complexity of the classification tasks. The best results were obtained in datasets collected under controlled conditions in laboratory settings, while significant lower performance was observed over less supervised, continuous recordings. DL methods consistently outperformed ML models in three out of 4 datasets, with improvements in F-score in the range from 6% to 11%. This suggests that automatically extracted features better model signal patterns at the level of each single dataset.

When evaluating generalization capability over cross-dataset settings, all classification approaches showed a significant performance drop. Moreover, DL methods were more affected than ML models, with a difference in F-score in the range from −8% to −40% between the two approaches. This suggests that feature-based learning better generalizes to diverse conditions and environments, while data-driven models may model dataset-specific patterns. Transfer learning produced substantial improvements over zero-shot tests, with improvements in F-score of 17%–40% in MLP and 33%–134% for the CNN.

Beyond computational efficiency, practical deployment of wearable stress detection systems requires consideration of signal quality, sensor reliability, and user variability. Physiological signals such as EDA and PPG are highly sensitive to motion artefacts, environmental conditions, and individual baseline differences. In real-world scenarios, robust pre-processing pipelines and artefact detection mechanisms are essential to ensure reliable predictions. Moreover, inter-subject variability suggests that lightweight personalization or calibration phases may be necessary to achieve stable performance across users. Battery consumption and continuous sensing constraints also play a key role, favouring models that minimize both computation and data transmission.

The unified experimental protocol provides robust evidence for model selection in wearable stress detection. The data-driven CNN offers state-of-the-art accuracy in controlled settings and substantial gains through adaptation, while the FD pipeline provides robustness under domain shift and superior interpretability. Building on these results, future work may construct a pooled multi-source training corpus by harmonizing the four datasets under a unified label scheme, enabling domain-robust models transferable

to unseen targets with minimal fine-tuning. Hybrid architectures that append recurrent layers (LSTM or GRU) after the convolutional encoder while jointly incorporating the handcrafted HRV and EDA descriptors could combine accuracy with cross-domain robustness. Evaluation should shift from window-level classification to continuous session-level assessment, measuring temporal detection latency and decision calibration under realistic conditions. Cost-sensitive or focal losses may further improve recall in skewed datasets, and self-supervised pretraining or foundation models could enhance out-of-domain generalization. Finally, collecting a dedicated in-the-wild validation set with ground truth obtained through ecological momentary assessment would provide the ultimate test of the generalization claims established here in controlled and semi-controlled settings.

CRediT authorship contribution statement

Matteo Calza-Metre: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation. **Luigi Borzì:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Formal analysis, Conceptualization.

Funding

This work received no external funding.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Supplementary material

A.1. Threshold sensitivity analysis for continuous-annotation datasets

See Tables 11 and 12 and Fig. 7.

A.2. Ablation study on windowing parameters

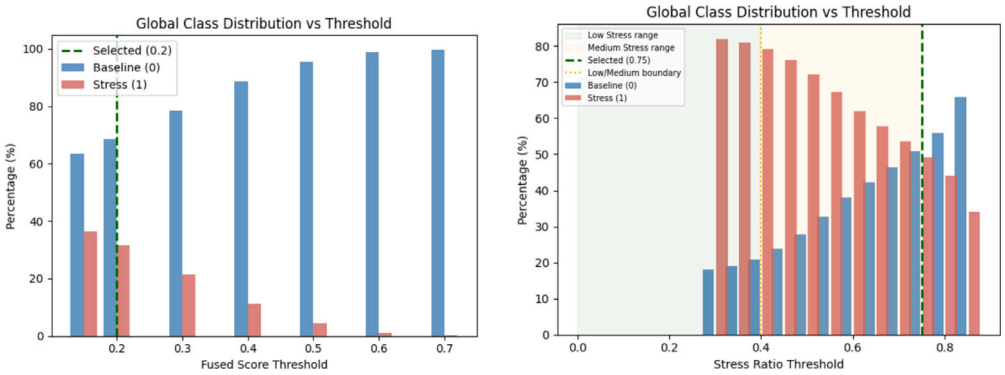
See Table 13.

Table 11
Threshold sensitivity analysis for AffectiveRoad. The original study defined three stress levels: low ($s < 0.4$), medium ($0.4 \leq s < 0.75$), and high ($s \geq 0.75$). The selected threshold (0.75) is shaded.

Threshold	Baseline	Stress	Stress %	Subj. >5%	Subj. 0%	Min %	Max %
0.30	1 027 539	4 666 113	81.95	13/13	0	53.39	97.65
0.40	1 189 535	4 504 117	79.11	13/13	0	50.62	93.48
0.50	1 582 252	4 111 400	72.21	13/13	0	38.57	87.93
0.60	2 162 106	3 531 546	62.03	13/13	0	26.49	80.86
0.70	2 646 267	3 047 385	53.52	12/13	0	0.07	75.13
0.75	**2 886 508**	**2 807 144**	**49.30**	**12/13**	**1**	**0.00**	**72.60**
0.80	3 186 642	2 507 010	44.03	12/13	1	0.00	72.07
0.85	3 754 822	1 938 830	34.05	12/13	1	0.00	65.30

Table 12
Threshold sensitivity analysis for VerBio. For each binarization threshold on the fused rater score, the table reports the global class distribution, the number of subjects with >5% stress windows, and the per-subject stress prevalence range. The selected threshold (0.20) is shaded.

Threshold	Baseline	Stress	Stress %	Subj. >5%	Subj. 0%	Min %	Max %
0.15	1 372 711	786 901	36.44	17/17	0	25.47	98.99
0.20	**1 479 091**	**680 521**	**31.51**	**17/17**	**0**	**7.31**	**97.93**
0.30	1 696 778	462 834	21.43	15/17	0	0.10	94.18
0.40	1 918 181	241 431	11.18	11/17	3	0.00	81.45
0.50	2 063 241	96 371	4.46	7/17	6	0.00	51.26
0.60	2 137 777	21 835	1.01	3/17	9	0.00	12.83
0.70	2 153 550	6 062	0.28	0/17	13	0.00	3.63



(a) VerBio. (b) AffectiveRoad.

Fig. 7. Threshold sensitivity analysis: global class distribution vs. binarization threshold for (a) VerBio (fused rater score, selected threshold 0.20, dashed green line) and (b) AffectiveRoad (observer-driver consensus score, selected threshold 0.75, dashed green line; orange dotted line marks the low/medium boundary at 0.40 from the original study).

Table 13
Ablation study on windowing parameters, conducted on the WESAD benchmark under LOSO cross-validation. For the FD pipeline, accuracy and macro-F1 correspond to the best-performing classifier per row (MLP in all cases). The selected configuration (45 s, 75%) is shaded.

Window/Overlap	Accuracy	Macro-F1	# Windows	# Features
Feature-driven pipeline				
20 s/70%	0.8931	0.8713	4554	30
45 s/75%	**0.9082**	0.8638	2399	**33**
60 s/50%	0.8977	0.8526	897	25
Data-driven pipeline (multi-stream 1-D CNN)				
20 s/70%	0.9500	0.9398	4554	–
45 s/75%	**0.9633**	**0.9575**	2399	–
60 s/50%	0.9253	0.9185	897	–

Table 14
Sensitivity analysis of fine-tuning depth: macro-F1 (mean ± SD) under LOSO cross-validation and average per-fold CPU adaptation time. Benchmark target: WESAD; sources: Campanella, VerBio, AffectiveRoad. The selected configurations are shaded.

Configuration	Trainable	Campanella	VerBio	AffectiveRoad	Time/fold
Feature-driven MLP (6 layers)					
Last-Only	1	0.688 ± 0.145	0.000 ± 0.000[a]	0.302 ± 0.204	~5 s
Last + 1H	2	0.695 ± 0.141	0.000 ± 0.000[a]	0.405 ± 0.202	~6 s
Last + 2H	3	0.712 ± 0.151	0.126 ± 0.182	0.492 ± 0.188	~7 s
Last + 3H	4	0.732 ± 0.154	0.070 ± 0.185	0.601 ± 0.218	~7 s
Last + 4H	5	0.769 ± 0.142	0.199 ± 0.302	0.676 ± 0.243	~8 s
Full Fine-tune	**6**	**0.789 ± 0.116**	**0.773 ± 0.144**	**0.787 ± 0.105**	**~9 s**
Data-driven CNN (15 layers)					
Last-Only	1	0.529 ± 0.267	0.475 ± 0.356	0.446 ± 0.252	~165 s
Classifier Head	3	0.584 ± 0.340	0.484 ± 0.330	0.598 ± 0.309	~178 s
Block 3 + Head	**7**	**0.631 ± 0.263**	**0.772 ± 0.294**	**0.719 ± 0.235**	**~340 s**
Block 2-3 + Head	11	0.842 ± 0.146	0.812 ± 0.227	0.836 ± 0.222	~560 s

[a] Denotes complete class collapse across all 15 folds.

A.3. Sensitivity analysis of freezing depth

See Table 14.

A.4. Supplementary figures

Figs. 8–11 report the Receiver Operating Characteristic (ROC) curves obtained under Leave-One-Subject-Out (LOSO) cross-validation for the feature-driven models, while Figs. 12–15 show the corresponding curves for the data-driven CNN. Fig. 16 shows per-subject macro-F1 distributions for all classifiers on each dataset, complementing the summary statistics and bootstrap confidence intervals reported in Table 6.

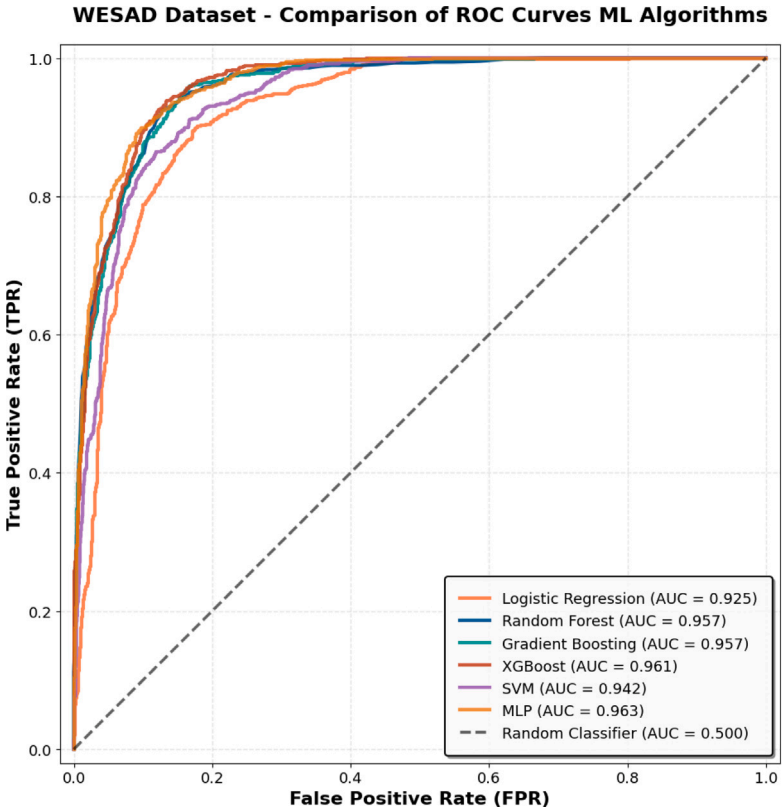


Fig. 8. ROC curves for feature-driven models on the Wesad dataset (LOSO).

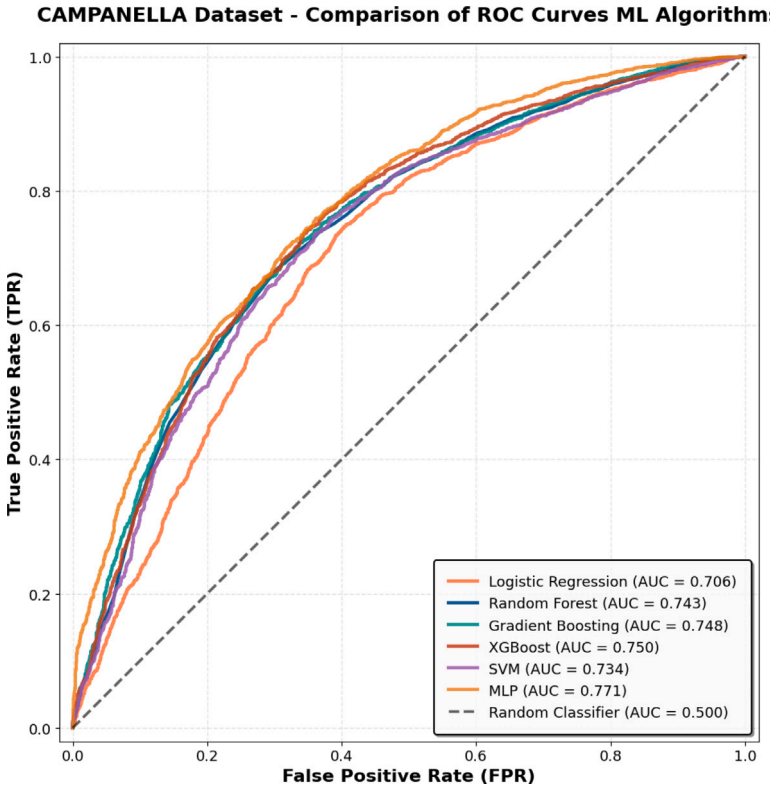


Fig. 9. ROC curves for feature-driven models on the Campanella dataset (LOSO).

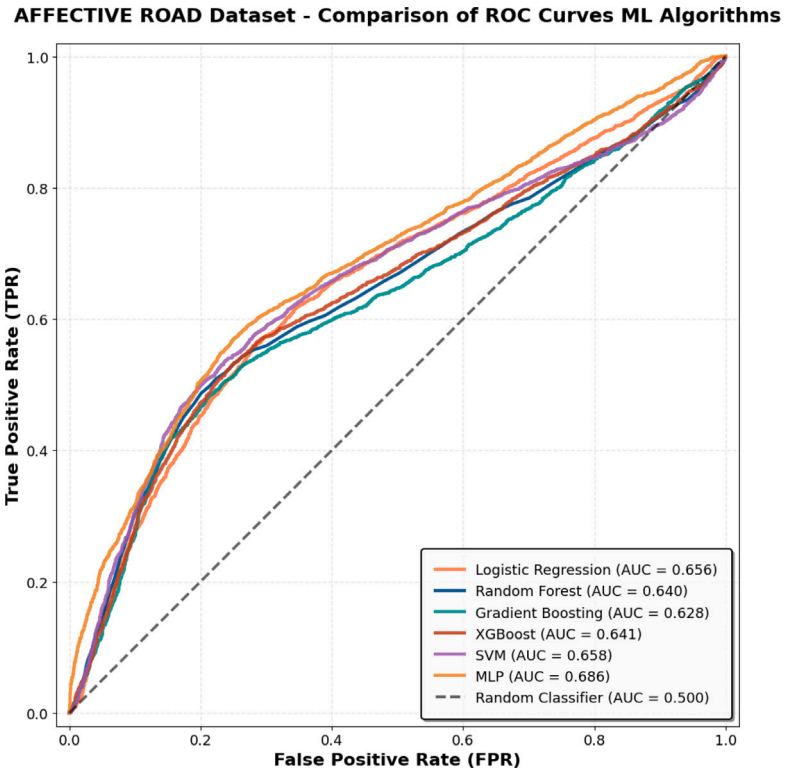


Fig. 10. ROC curves for feature-driven models on the AffectiveRoad dataset (LOSO).

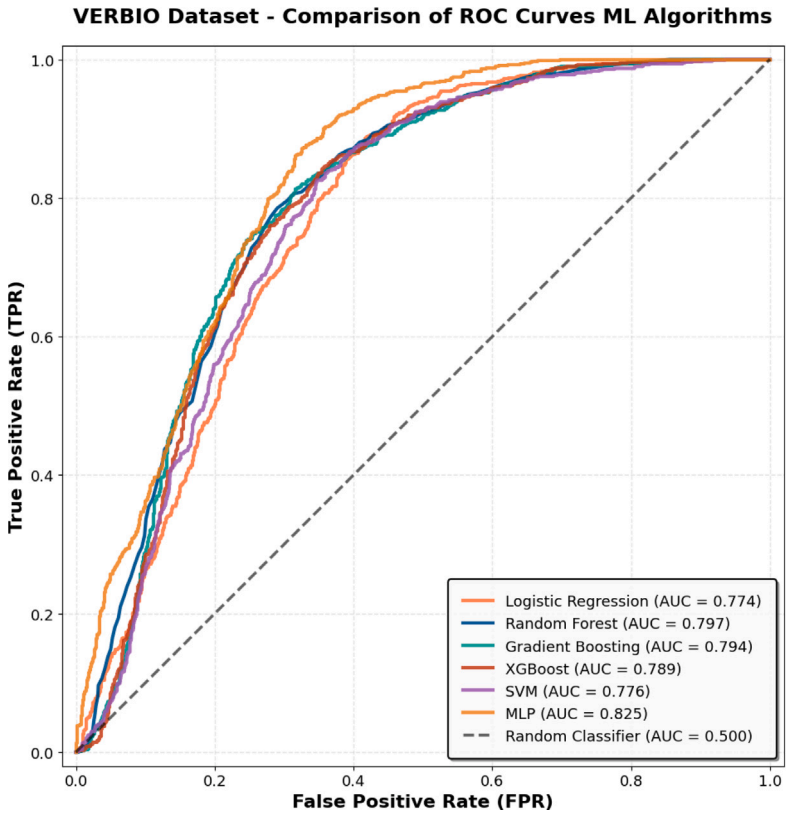


Fig. 11. ROC curves for feature-driven models on the VerBio dataset (LOSO).

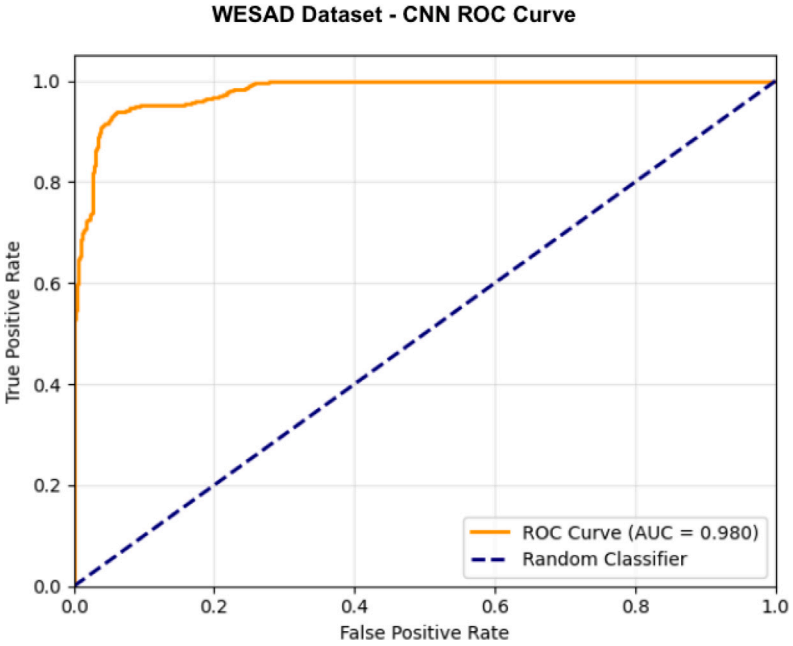


Fig. 12. ROC curve for the data-driven CNN on the Wesad dataset (LOSO).

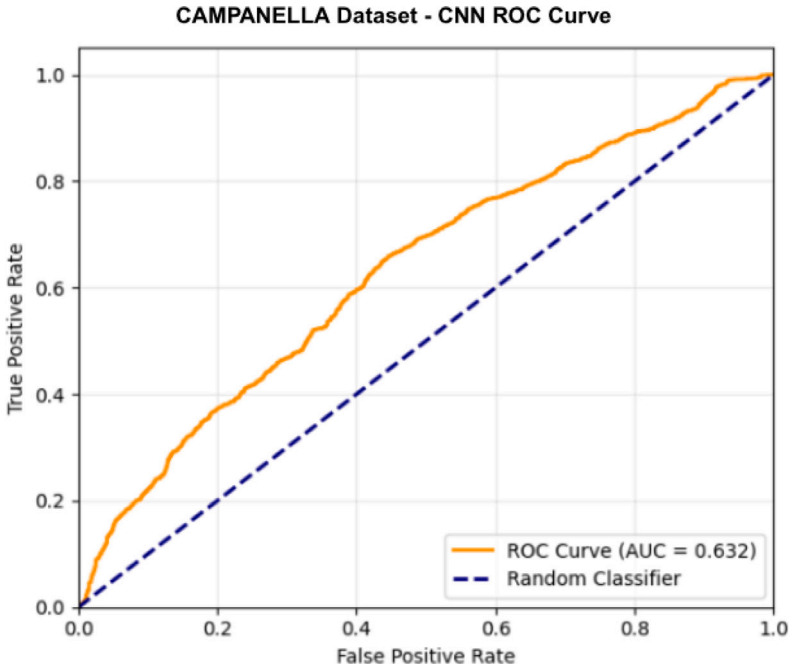


Fig. 13. ROC curve for the data-driven CNN on the Campanella dataset (LOSO).

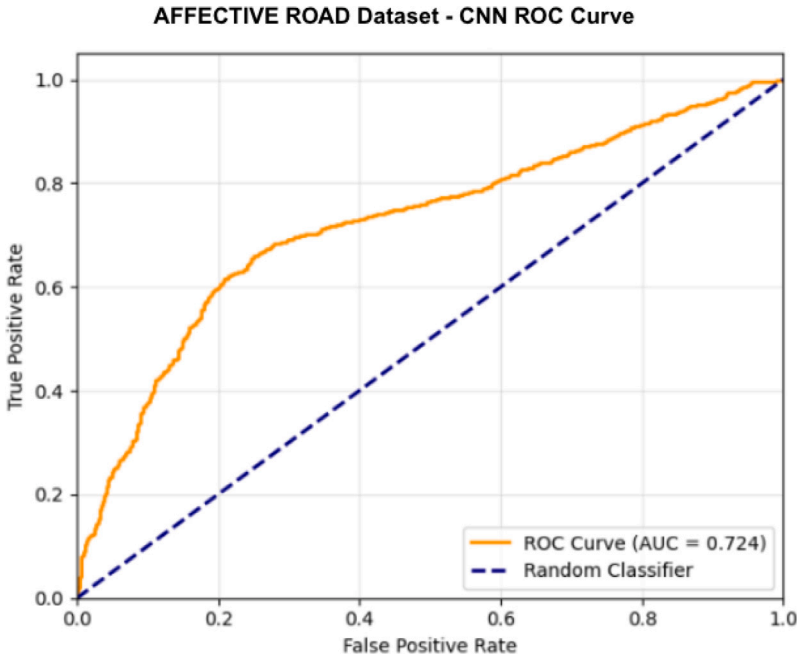


Fig. 14. ROC curve for the data-driven CNN on the AffectiveRoad dataset (LOSO).

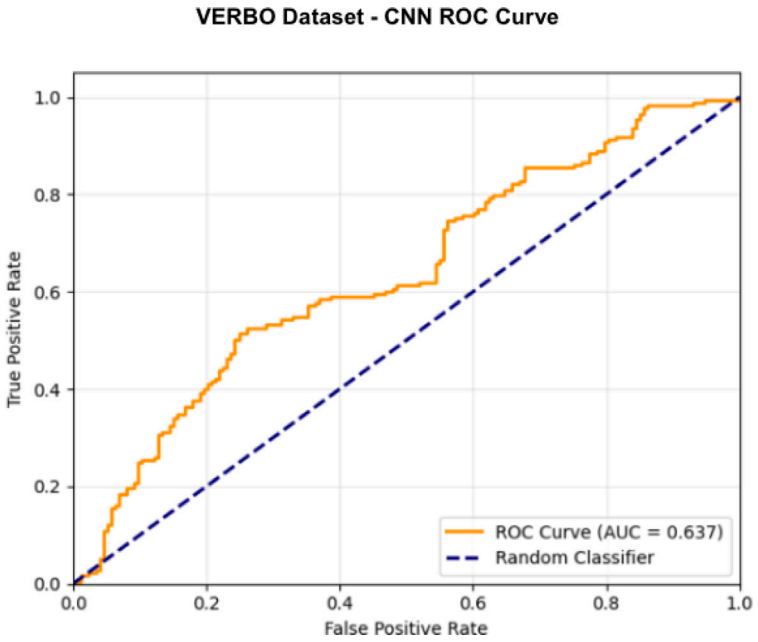


Fig. 15. ROC curve for the data-driven CNN on the VerBio dataset (LOSO).

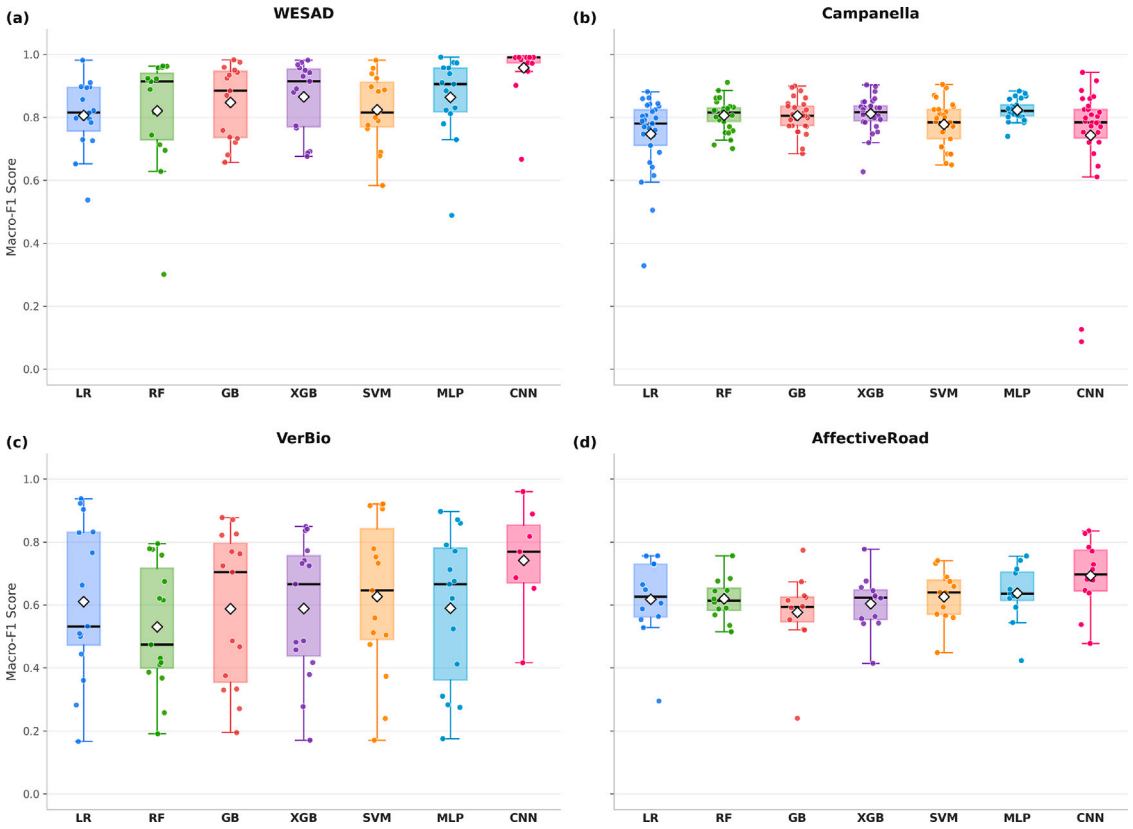


Fig. 16. Per-subject macro-F1 distributions under LOSO cross-validation. (a) Wesad, (b) Campanella, (c) VerBio, (d) AffectiveRoad. Each box shows the interquartile range with the median (black line) and mean (diamond); individual subjects are overlaid as coloured dots. Wesad and Campanella exhibit compact distributions centred above 0.80, while VerBio and AffectiveRoad show markedly wider spread, reflecting ecological variability and smaller cohort sizes.

Data availability

All datasets used in this study are publicly available. The Wesad dataset can be accessed at https://ubi29.informatik.uni-siegen.de/usi/data_wesad.html. The Campanella et al. dataset is available on Mendeley Data at https://data.mendeley.com/datasets/kb42z77m2g/2. The AffectiveRoad dataset can be downloaded from the MIT Media Lab at https://www.media.mit.edu/tools/affectiveroad/. The VerBio dataset is accessible at https://sites.google.com/colorado.edu/verbio-dataset. The code developed for this study is publicly available at https://github.com/matteo9910/StressDetectionBasedOnWearableSensorData.

References

Abdelfattah, E., Joshi, S., & Tiwari, S. (2025). Machine and deep learning models for stress detection using multimodal physiological data. *IEEE Access*, *13*, 4597–4608. http://dx.doi.org/10.1109/ACCESS.2024.3525459.

Albaladejo González, M., Ruipérez-Valiente, J. A., & Gomez Marmol, F. (2022). Evaluating different configurations of machine learning models and their transfer learning capabilities for stress detection using heart rate. *Journal of Ambient Intelligence and Humanized Computing*, *14*, http://dx.doi.org/10.1007/s12652-022-04365-z.

Almadhor, A., Ojo, S., Nathaniel, T. I., Ukpong, K., Alsubai, S., & Al Hejaili, A. (2025). A Cross-Domain framework for emotion and stress detection using WESAD, ScientISST-MOVE, and DREAMER datasets. *Frontiers in Bioengineering and Biotechnology*, *13*, Article 1659002. http://dx.doi.org/10.3389/fbioe.2025.1659002.

Alonso, J., Romero, S., Ballester, M., Antonijoan, R., & Mañanas, M. (2015). Stress assessment based on EEG univariate features and functional connectivity measures. *Physiological Measurement*, *36*(7), 1351–1365. http://dx.doi.org/10.1088/0967-3334/36/7/1351.

Benchekroun, Y., Abdelaziz, M., & Messaoud, S. (2021). Stress detection using wearable physiological sensors: A comparative study. In *2021 IEEE international conference on smart healthcare* (pp. 1–6). IEEE.

Benchekroun, M., Istrate, D., Zalc, V., & Lenne, D. (2023). Cross dataset analysis for generalizability of HRV-Based stress detection. *Sensors*, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9960690/.

Bolpagni, M., Pardini, S., Dianti, M., & Gabrielli, S. (2024). Personalized stress detection using biosignals from wearables: A scoping review. *Sensors*, *24*(10), http://dx.doi.org/10.3390/s24103221.

Bolpagni, D., et al. (2024). Personalized stress detection using biosignals from wearable devices: A scoping review. *Sensors*, *24*(10), 3221. http://dx.doi.org/10.3390/s24103221, URL https://www.mdpi.com/1424-8220/24/10/3221.

Campanella, S., Altaleb, A., Belli, A., Pierleoni, P., & Palma, L. (2023). A method for stress detection using empatica E4 bracelet and machine-learning techniques. *Sensors*, *23*(7), 3565. http://dx.doi.org/10.3390/s23073565, URL https://www.mdpi.com/article/10.3390/s23073565.

Can, Y. S., Chalabianloo, N., Ekiz, D., & Ersoy, C. (2019). Continuous stress detection using wearable sensors in real life: A review. *Sensors*, *19*(19), 4415. http://dx.doi.org/10.3390/s19194415, URL https://www.mdpi.com/1424-8220/19/19/4415.

Cui, X., Sun, H., Chen, Z., & Peng, C.-K. (2025). Enhanced PPG-based stress detection: a multivariate cross-dataset analysis across devices and tasks. *Biomedical Signal Processing and Control*, [ISSN: 1746-8094] *110*, Article 108149. http://dx.doi.org/10.1016/j.bspc.2025.108149.

de Witte, M., et al. (2021). Self-Report stress measures to assess stress in adults. *Frontiers in Psychology*, *12*, http://dx.doi.org/10.3389/fpsyg.2021.742566, URL https://www.frontiersin.org/articles/10.3389/fpsyg.2021.742566/full.

El Haouij, N., Poggi, J.-M., Sevestre-Ghalila, S., Ghozi, R., & Jaïdane, M. (2018). AffectiveROAD system and database to assess driver's attention. In *Proceedings of the 33rd annual ACM symposium on applied computing* (pp. 800–803). http://dx.doi.org/10.1145/3167132.3167280.

Epel, E. S., Crosswell, A. D., Mayer, S. E., Prather, A. A., Slavich, G. M., Puterman, E., & Mendes, W. B. (2018). More than a feeling: A unified view of stress measurement for population science. *Frontiers in Neuroendocrinology*, URL https://www.sciencedirect.com/science/article/pii/S0091302218300219.

Ghasemi, F., Beversdorf, D. Q., & Herman, K. C. (2024). Stress and stress responses: A narrative literature review from physiological mechanisms to intervention approaches. *Journal of Pacific Rim Psychology*, URL https://doi.org/10.1177/18344909241289222.

Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., & Tsiknakis, M. (2022). Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*.

Gil-Martín, M., San-Segundo, R., Mateos, A., & Ferreiros-López, J. (2022). Human stress detection with wearable sensors using convolutional neural networks. *IEEE Aerospace and Electronic Systems Magazine*, *37*(1), 60–70. http://dx.doi.org/10.1109/MAES.2021.3115198.

Gjoreski, M., Čvetković, B., Gjoreski, H., & Luštrek, M. (2017). Monitoring stress with a wrist device using context. *Journal of Biomedical Informatics*, *73*, 159–170. http://dx.doi.org/10.1016/j.jbi.2017.08.006, URL https://dis.ijs.si/mitjal/documents/GjoreskiM-Monitoring_stress_with_a_wrist_device_using_context-JBI-17.pdf.

Gjoreski, M., Gjoreski, H., Luštrek, M., & Gams, M. (2016). Continuous stress detection using a wrist device—in laboratory and real life. In *Proceedings of the ACM international joint conference on pervasive and ubiquitous computing adjunct* (pp. 1185–1193). http://dx.doi.org/10.1145/2968219.2968306, URL https://ubicomp-mental-health.github.io/papers/gjoreski-stress.pdf.

Harris, K. M., et al. (2023). The perceived stress scale as a measure of stress. *Frontiers in Psychology*, *14*, Article 10498818. http://dx.doi.org/10.3389/fpsyg.2023.10498818, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10498818/.

Healey, J. A., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, *6*(2), 156–166. http://dx.doi.org/10.1109/TITS.2005.848368, URL https://www.cs.cmu.edu/~cga/behavior/Healey.pdf.

Herborn, K. A., Graves, J. L., Jerem, P., Evans, N. P., Nager, R., McCafferty, D. J., & McKeegan, D. E. (2015). Skin temperature reveals the intensity of acute stress. *Physiology & Behavior*, *152*, 225–230. http://dx.doi.org/10.1016/j.physbeh.2015.09.032, Epub 2015 Oct 3.

Hickey, B. A., Chalmers, T., Newton, P., Lin, C.-T., Sibbritt, D., McLachlan, C. S., Clifton-Bligh, R., Morley, J., & Lal, S. (2021). Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review. *Sensors*, *21*(10), 3461. http://dx.doi.org/10.3390/s21103461, URL https://www.mdpi.com/1424-8220/21/10/3461.

Hjortskov, N., Rissén, D., Blangsted, A. K., Fallentin, N., Lundberg, U., & Søgaard, K. (2004). The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal of Applied Physiology*, *92*(1–2), 84–89. http://dx.doi.org/10.1007/s00421-004-1055-z, Epub 2004 Feb 27. PMID: 14991326.

Hovsepian, K., al'Absi, M., Ertin, E., Kamarck, T., Nakajima, M., & Kumar, S. (2015). Cstress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing* (pp. 493–504). ISBN: 9781450335744, http://dx.doi.org/10.1145/2750858.2807526.

Kaczor, E. E., Carreiro, S., Stapp, J., Chapman, B., & Indic, P. (2020). Objective measurement of physician stress in the emergency department using a wearable sensor. In *Proceedings of the annual hawaii international conference on system sciences* (pp. 3729–3738). URL https://hdl.handle.net/10125/64061.

Kapogianni, N.-A., Sideraki, A., & Anagnostopoulos, C.-N. (2025). Using smartwatches in stress management, mental health, and well-being: A systematic review. *Algorithms*, *18*(7), http://dx.doi.org/10.3390/a18070419.

Kim, H.-G., Cheon, E.-J., Bai, D.-S., Lee, Y. H., & Koo, B.-H. (2018). Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry Investigation*, *15*(3), 235–245. http://dx.doi.org/10.30773/pi.2017.08.17, Epub 2018 Feb 28.

Koelstra, S., Mühl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & Patras, I. (2012). DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, *3*(1), 18–31. http://dx.doi.org/10.1109/T-AFFC.2011.15, URL https://ieeexplore.ieee.org/document/6119062.

Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A., & Kraaij, W. (2014). The SWELL knowledge work dataset for stress and user modeling research. In *Proceedings of the ACM international conference on multimodal interaction* (pp. 291–298). http://dx.doi.org/10.1145/2663204.2663257, URL https://cs.ru.nl/~skoldijk/Papers/ICMI%202014%20paper_final_cr.pdf.

Kulanthaivel Lakshmanan, L., Leelasankar, K., & Subbiyan, B. (2026). Stress detection using machine learning and deep learning techniques: A systematic review and Meta-Analysis. *Archives of Computational Methods in Engineering*, *33*, 4183–4215. http://dx.doi.org/10.1007/s11831-025-10429-y.

Lab, C. (2023). VERBIO dataset documentation (README). Supplementary documentation distributed with the public VERBIO dataset release, Includes folder structure, signal formats, and continuous fused annotation specification at 1 Hz.

Ladakis, E., Athanasopoulos, G., & Papadopoulos, C. (2022). A lightweight CNN for multimodal stress detection using biosignal spectrograms. In *Proceedings of the 16th international joint conference on biomedical engineering systems and technologies*.

Ladakis, I., Fotopoulos, D., & Chouvarda, I. (2025). Integrative analysis of open datasets for stress prediction. *Journal of Medical and Biological Engineering*, URL https://link.springer.com/article/10.1007/s40846-025-00958-z.

Lazarou, E., & Exarchos, T. P. (2024). Predicting stress levels using physiological data: Real-time stress prediction models utilizing wearable devices. *PMC*, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11230864/, Open access review.

Lazarus, R. S., DeLongis, A., Folkman, S., & Gruen, R. (1985). Stress and adaptational outcomes: The problem of confounded measures.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. H. A. (2021). NeuroKit2: A python toolbox for neurophysiological signal processing. Accessed 08 October 2025.

Milstein, N., & Gordon, I. (2020). Validating measures of electrodermal activity and heart rate variability derived from the empatica E4 utilized in research settings that involve interactive dyadic states. *Frontiers in Behavioral Neuroscience*, *14*, 148. http://dx.doi.org/10.3389/fnbeh.2020.00148, URL https://www.frontiersin.org/articles/10.3389/fnbeh.2020.00148/full.

Moser, M. K., Ehrhart, M., & Resch, B. (2024). An explainable deep learning approach for stress detection in wearable sensor measurements. *Sensors*, *24*(16), 5085. http://dx.doi.org/10.3390/s24165085.

Nugroho, T., Rahmaniar, W., & Ma'arif, A. (2026). A narrative review of AI frameworks for chronic stress detection using physiological sensing: Resting, longitudinal, and reactivity perspectives. *Sensors*, *26*(8), http://dx.doi.org/10.3390/s26082345.

Pataca, A. O., Zdravevski, E., Coelho, P. J., Garcia, N. M., Deryuck, M., Albuquerque, C., & Pires, I. M. (2025). Use of machine learning for predicting stress episodes based on wearable sensor data: A systematic review. *Computers in Biology and Medicine*, [ISSN: 0010-4825] *198*, Article 111166. http://dx.doi.org/10.1016/j.compbiomed.2025.111166.

Patwal, A., Diwakar, M., Tripathi, V., & Singh, P. (2023). An investigation of videos for abnormal behavior detection. *Procedia Computer Science*, [ISSN: 1877-0509] *218*, 2264–2272. http://dx.doi.org/10.1016/j.procs.2023.01.202, International Conference on Machine Learning and Data Engineering.

Pinge, A., Gad, V., Jaisighani, D., Ghosh, S., & Sen, S. (2024). Detection and monitoring of stress using wearables: a systematic review. *Frontiers in Computer Science*, *6*, http://dx.doi.org/10.3389/fcomp.2024.1478851.

Prajod, P., Mahesh, B., & André, E. (2024). Stressor type matters! exploring factors influencing cross-dataset generalizability of physiological stress detection. arXiv preprint arXiv:2405.09563 URL https://arxiv.org/abs/2405.09563.

Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., & Van Laerhoven, K. (2018). Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction* (pp. 400–408). http://dx.doi.org/10.1145/3242969.3242985, URL https://ubi29.informatik.uni-siegen.de/usi/pdf/ubi_icmi2018.pdf.

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., & Ehlert, U. (2010). Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on Information Technology in Biomedicine*, *14*(2), 410–417. http://dx.doi.org/10.1109/TITB.2009.2036164.

Sigcha, L., Pereira, E., Borzì, L., Gachet, D., Cardoso, P., & Costa, N. (2026). Enhancing cross-dataset mental workload detection using electrodermal activity and domain adaptation. *Applied Sciences*, *16*(10), http://dx.doi.org/10.3390/app16104673.

Smets, E., et al. (2023). Ensemble machine learning model trained on a new synthesized dataset generalizes well for stress prediction using wearable devices. *Journal of Biomedical Informatics*, *148*, Article 104540. http://dx.doi.org/10.1016/j.jbi.2023.104540.

Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). MAHNOB-HCI: A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, *3*(1), 42–55. http://dx.doi.org/10.1109/T-AFFC.2011.15, URL https://ieeexplore.ieee.org/document/6004928.

Sosa, S., Fontecchio, A. K., Chrysikou, E. G., & Atchison, J. S. (2026). Beyond EDA: A systematic review of multimodal sympathetic nervous system arousal classification for stress detection. *Sensors*, *26*(5), http://dx.doi.org/10.3390/s26051584.

Tanwar, S., et al. (2023). Attention based hybrid deep learning model for wearable based stress recognition. *Engineering Applications of Artificial Intelligence*, *126*, Article 107139. http://dx.doi.org/10.1016/j.engappai.2023.107139.

Taskasaplidis, G., Fotiadis, D. A., & Bamidis, P. D. (2024). Review of stress detection methods using wearable sensors. *IEEE Access*, *12*, 38219–38239. http://dx.doi.org/10.1109/ACCESS.2024.3373010.

Tramel, W., Schram, B., Canetti, E., & Orr, R. (2023). An examination of subjective and objective measures of stress in tactical populations: A scoping review. *Healthcare*, *11*(18), 2515. http://dx.doi.org/10.3390/healthcare11182515, URL https://www.mdpi.com/2227-9032/11/18/2515.

Vafaei, N., & Hosseini, S. (2025). A cross-domain framework for emotion and stress detection using WESAD, SCIENTISST-MOVE, and DREAMER datasets. *Frontiers in Bioengineering and Biotechnology*, *13*, http://dx.doi.org/10.3389/fbioe.2025.1659002.

Vos, G., Trinh, K., Sarnyai, Z., & Rahimi Azghadi, M. (2022). Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review. arXiv preprint arXiv:2209.15137 URL https://arxiv.org/abs/2209.15137.

Vos, G., Trinh, K., Sarnyai, Z., & Rahimi Azghadi, M. (2023). Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review. *International Journal of Medical Informatics*, [ISSN: 1386-5056] *173*, Article 105026. http://dx.doi.org/10.1016/j.ijmedinf.2023.105026.

Wang, Z.-H., & Wu, Y.-C. (2022). A novel rapid assessment of mental stress by using PPG signals based on deep learning. *IEEE Sensors Journal*, *22*, 21232–21239. http://dx.doi.org/10.1109/JSEN.2022.3208427.

Xiang, J.-Z., Wang, Q.-Y., Fang, Z.-B., Esquivel, J. A., & Su, Z.-X. (2025a). A Multi-Modal deep learning approach for stress detection in nursing staff. *Frontiers in Physiology*, *16*, Article 1584299, URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11997569/.

Xiang, J.-Z., Wang, Q.-Y., Fang, Z.-B., Esquivel, J. A., & Su, Z.-X. (2025b). A multi-modal deep learning approach for stress detection using physiological signals: integrating time and frequency domain features. *Frontiers in Physiology*, *16*, http://dx.doi.org/10.3389/fphys.2025.1584299.

Zhu, L., Spachos, P., Ng, P. C., Yu, Y., Wang, Y., Plataniotis, K. N., & Hatzinakos, D. (2023). Stress detection through wrist-based electrodermal activity monitoring and machine learning. *IEEE Journal of Biomedical and Health Informatics*, *27*(5), 2155–2165. http://dx.doi.org/10.1109/JBHI.2023.3239305, Epub 2023 May 4.