Cross-dataset generalizability analysis of multimodal self-supervised learning for stress recognition across lab and daily contexts

Yekta Said Can, Mohamed Benouis, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Can, Yekta Said, Mohamed Benouis, and Elisabeth André. 2026. "Cross-dataset generalizability analysis of multimodal self-supervised learning for stress recognition across lab and daily contexts." *IEEE Access* 14: 35930–43. https://doi.org/10.1109/access.2026.3670764.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 February 2026, accepted 2 March 2026, date of publication 5 March 2026, date of current version 10 March 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3670764

RESEARCH ARTICLE

Cross-Dataset Generalizability Analysis of Multimodal Self-Supervised Learning for Stress Recognition Across Lab and Daily Contexts

YEKTA SAID CAN, MOHAMED BENOUIS, AND ELISABETH ANDRÉ, (Senior Member, IEEE)
Chair for Human-Centered Artificial Intelligence, Institute of Computer Science, Universität Augsburg, 86159 Augsburg, Germany
Corresponding author: Yekta Said Can (yekta.can@uni-a.de)

ABSTRACT Stress recognition is a key component of affect-aware systems to improve mental and physical well-being. While multimodal affect recognition systems based on physiological signals have shown promise, achieving robust generalization across different datasets remains a major challenge due to variations in stress induction protocols and labeling practices. For instance, ''stress'' labels can vary widely between datasets: WESAD uses the Trier Social Stress Test to induce social-evaluative stress, while SWELL-KW relies on cognitive workload tasks. Such differences in the nature and intensity of stressors, as well as inconsistencies in how labels are defined (e.g., ''social stress'' vs. ''mental stress''), create major challenges for generalization. Prior work has explored deep transfer learning and unimodal self-supervised methods, but cross-dataset generalizability is still limited. To address this gap, we propose a multimodal self-supervised learning (SSL) framework based on contrastive objectives that learns transferable representations from unlabeled physiological signals. Unlike conventional deep transfer learning approaches, our framework does not rely on stress labels during the pretraining stage and is evaluated under a strict leave-one-subject-out (LOSO) protocol to ensure realistic cross-subject generalization. We systematically study the impact of SSL across multiple encoder architectures, including Convolutional Neural Networks (CNN), Temporal Convolutional Networks (TCN), ResNet34-1D, and a CNN–Transformer hybrid, enabling a systematic analysis of how different encoder architectures affect representation transferability. Experiments are conducted across three laboratory datasets (WESAD, VERBIO, AffectHRI) and two daily life datasets (SWEET, LD), covering lab-to-lab, lab-to-daily, and daily-to-lab transfer scenarios. Overall, our findings highlight multimodal self-supervised learning as an effective and label-efficient framework for improving cross-dataset generalization, particularly under realistic cross-subject and cross-context evaluation settings.

INDEX TERMS Wearable, affective computing, emotion recognition, deep learning, transfer learning, self-supervised learning, physiological signals.

I. INTRODUCTION

Stress is a pervasive issue in modern society, significantly impacting both physical and mental health. Chronic stress has been linked to a range of adverse health outcomes, including cardiovascular diseases, weakened immune function, and mental health disorders such as anxiety and depression [1]. Early detection and effective management of stress are crucial to mitigate these negative effects and improve overall well-being.

Affective computing, a field dedicated to enabling machines to recognize and respond to human emotions, has significantly advanced stress detection by integrating various modalities [2]. Advancements in wearable technology have enabled continuous monitoring of physiological signals, offering new avenues for stress detection and intervention. Early research primarily focused on analyzing physiological signals such as heart rate variability, galvanic skin response, and

The associate editor coordinating the review of this manuscript and approving it for publication was Norbert Herencsar.

electroencephalography (EEG) to identify stress indicators. Over time, the scope expanded to include behavioral cues like facial expressions, speech patterns, and body movements, recognizing that stress manifests through multiple channels [3]. A multimodal approach enhances the accuracy and reliability of stress detection systems, as it captures a comprehensive picture of an individual's emotional state. Multimodal machine learning models have been developed to analyze multimodal data streams in publicly available datasets, aiming to identify stress patterns and predict stress events.

However, a significant challenge in this domain is the variability across different datasets, particularly due to differences in stress induction methods and labeling protocols. Models trained on one dataset often perform poorly when applied to another, as the specific stressors and labeling criteria can significantly influence outcomes. The difficulty of transferring pretrained models from public datasets hinders the real-time affect monitoring and intervention system performance and raises questions regarding the generalizability of the proposed models. Researchers tried deep transfer learning approaches for training models in one dataset and testing in the other. If the transfer of knowledge is successful, the real-time intervention systems will require less amount of data collection for deployment since some knowledge from the public datasets can be transferred. However, the supervised deep transfer learning approaches cross dataset accuracies are around 70% [4] which is far from a robust performance.

Lately, self-supervised learning (SSL) methods have been explored for their potential to enhance model robustness in multimodal stress detection tasks. SSL leverages large amounts of unlabeled data to learn meaningful representations, reducing reliance on labeled data and mitigating the impact of dataset-specific stressors and labels. Specifically, Matton et al. [5] apply unimodal (only Electrodermal Activity-based) contrastive SSL on laboratory-collected datasets (WESAD and VERBIO), demonstrating higher generalizability than supervised learning, with up to 79.8% accuracy across datasets. Meanwhile, Eom et al. [6] present SIM-CNN, a self-supervised framework evaluated on a real-world multimodal dataset collected from nurses, comprising over 1,250 hours of biosignals (83 hours labeled with stress levels). While they focus on real-world deployment without cross-dataset validation, Matton et al. [5] emphasize architectural innovations in SSL.

By integrating SSL techniques, stress detection models can achieve greater generalizability across different datasets, reducing the dependency on specific stress induction methods and labels, and thereby enhancing their applicability in real-world scenarios. This advancement holds promise for developing more reliable and personalized stress monitoring systems, ultimately contributing to better health outcomes.

In this study, we proposed a multimodal self-supervised affect recognition system and tested it with multimodal physiological datasets collected in the laboratory and in the wild. We tested various architectures for recognizing emotions and stress. We evaluate whether contrastive self-supervised learning, previously shown to be effective on unimodal laboratory datasets, can generalize across real-world daily life settings. To the best of our knowledge, this is the first work to apply multimodal SSL to physiological affect recognition in a cross-dataset setting that spans diverse stress stimuli (e.g., cognitive workload, social evaluation) and environments (e.g., lab vs. daily life), offering a scalable and robust framework for stress modeling in everyday life contexts. The main contributions of this work can be summarized as follows:

- We propose a multimodal self-supervised learning (SSL) framework for physiological stress recognition that learns transferable representations from unlabeled biosignals using contrastive objectives. Unlike conventional supervised deep transfer learning (DTL), the proposed framework removes the dependency on stress labels during the pretraining stage, addressing label noise and protocol inconsistencies across datasets.
- We conduct a systematic cross-dataset evaluation under a strict leave-one-subject-out (LOSO) protocol, ensuring realistic subject-independent generalization.
- We perform an architecture-aware comparison by evaluating SSL and supervised transfer learning across multiple encoder designs, including CNN, Temporal Convolutional Networks (TCN), ResNet34-1D, and CNN–Transformer hybrids. This analysis reveals how architectural inductive biases influence representation transferability and robustness under domain shift.
- We systematically investigate cross-domain transfer directions, including lab-to-lab, lab-to-daily, and daily-to-lab scenarios.

The rest of the paper is organized as follows: In Section II, the related work for automatic affect recognition systems that use physiological signals is presented. In Section III, the used datasets are explained. In Section IV, different self-supervised affect recognition architectures are explained. In Section V, the experimental results of the proposed systems are discussed. In Section VI, we summarize and discuss the findings and future work of the current research is presented.

II. BACKGROUND AND RELATED WORK

A. PSYCHOMETRIC BACKGROUND: STRESS AS A LATENT CONSTRUCT

In psychological measurement, stress is not directly observable but is conceptualized as a latent construct inferred from systematic covariation among multiple indicators under an explicit measurement model. Modern validity theory emphasizes that the meaning of a stress score depends on the theoretical construct it is intended to represent and on accumulated evidence supporting the interpretation and use of that score, rather than on the label itself or face validity alone [10], [11], [12]. From this perspective, stress is defined operationally through psychometric models that

TABLE 1. Summary of related works on stress and emotion recognition using wearable signals. ST is for skin temperature.

Study	Signals	Context	Modality	Method	Label Type	Generalizability Analysis	Notes
Akbulut et al. [7]	ECG, GSR, ST, SpO_2	Lab	Multimodal	LSTM, CNN, CNN-RF	Emotions	No	High lab accuracy
Eom et al. [6]	Multimodal	Lab	Multimodal	SSL (SIM-CNN)	Stress	No	SSL pretraining; lab-only
Prajod et al. [8]	HRV	Lab	Single	Supervised Deep Transfer Learning	Stress	Cross-dataset (HRV)	Stressor type critical; low accuracies (60-70%)
Benchekroun et al. [4]	HRV	Lab	Single	Supervised Deep Transfer Learning	Stress	Cross-dataset (HRV)	Traditional supervised TL
Prajod et al. [9]	ECG	Lab	Single	Supervised Deep Transfer Learning	Stress	Cross-dataset (ECG)	Moderate generalization
Matton et al. [5]	EDA	Lab	Single	SSL (contrastive)	Stress	Cross-dataset (EDA)	SSL better than DTL; single modality
Our study	EDA, TEMP, PPG	Lab + Daily Life	Multimodal	SSL (contrastive)	Stress, Arousal	Cross-dataset (multimodal)	**First multimodal SSL cross-dataset study validated in daily life**

distinguish true-score variance from measurement error and method-related effects.

Most contemporary stress research is grounded in appraisal-based theory, where stress reflects perceived imbalance between environmental demands and coping resources rather than objective exposure to stressors [13]. This conceptualization underlies widely used self-report instruments such as the Perceived Stress Scale (PSS), which operationalizes stress as a global appraisal characterized by perceived unpredictability, uncontrollability, and overload [14]. Psychometric analyses of such instruments frequently reveal structured dimensionality, often requiring confirmatory factor or bifactor models to separate a general stress factor from wording- or valence-related method effects [15]. These findings indicate that treating stress as a single unmodeled sum score can obscure construct meaning and compromise comparability.

These psychometric considerations become particularly salient in cross-dataset and algorithmic settings, where stress labels are often inherited from dataset-specific protocols (e.g., social-evaluative stress, cognitive workload, time pressure) and heterogeneous annotation practices. Without construct-aware alignment, models risk learning protocol-specific proxies rather than transferable representations of stress. Measurement invariance testing and related latent-variable methods provide a principled framework for assessing whether stress measures function equivalently across populations, contexts, and datasets, thereby supporting meaningful cross-dataset comparison and generalization [16], [17]. Incorporating such construct-level considerations strengthens both psychological interpretability and algorithmic robustness in wearable stress recognition research.

B. LITERATURE REVIEW

In order to deploy real-time affect recognition systems is the need for a huge amount of data for training the initial models. One way to alleviate this problem is to transfer knowledge from the existing public datasets and require less amount of local data for the initial deployment. Thus, advancements in affect recognition have increasingly focused on leveraging multimodal physiological signals combined with deep transfer learning to enhance model generalizability across diverse datasets. In this way, a model is trained with one dataset, and by freezing or fine-tuning with the second dataset, the cross-dataset performance is measured. We created a table (Table 1) for studies that conduct cross-dataset evaluation for affect recognition using physiological signals. Note that (SIM-CNN) [6] was included due to its application of self-supervised learning on a real-world multimodal dataset collected from nurses, representing one of the earliest SSL attempts in real-life stress detection, even though it does not involve cross-dataset evaluation. For instance, a study by Can et al. explored the feasibility of using laboratory experiments to monitor stress in real-life settings, highlighting the challenges of transferring models trained in controlled environments to daily life scenarios [18]. Their findings underscore the necessity for models that can adapt to varying contexts and individual differences. Traditional supervised DTL is applied for single modalities such as PPG [4] or ECG [9] and the cross-dataset accuracies are around 60-70%. To properly contextualize the reported cross-dataset accuracies, we also include within-dataset performance as a baseline. For WESAD, leave-one-subject-out (LOSO) evaluations using HRV-based models (e.g., SVM, RF, MLP) typically yield around 83% accuracy and for SWELL-KW around 70% [8]. In contrast, cross-dataset tests from SWELL-KW to WESAD show marked drops: RF 67.6%, MLP 57.4%, SVM 48.2%. Similarly, WESAD-trained models evaluated on SWELL-KW fall below 50% accuracy (MLP 47.9%, RF 45.0%, SVM 43.9%). These results highlight the substantial performance decline when models are applied across datasets, emphasizing the necessity of including within-dataset baselines for meaningful comparisons. A key

challenge in this field is the inconsistency across datasets, especially due to variations in stress induction techniques and labeling standards. Models developed on one dataset frequently underperform on others, as differing stressors and labeling practices can substantially impact the results.

Prajod et al. [8] systematically investigated cross-dataset generalizability by categorizing primary stressors as either social-evaluative (e.g., ForDigitStress: job interview, VERBIO: public speaking, WESAD: Trier Social Stress Test) or mental effort (e.g., SWELL-KW: cognitive workload). They demonstrated that mismatches in stressor types caused substantial performance drops. By applying supervised DTL techniques across four datasets (WESAD, SWELL-KW, ForDigitStress, and VERBIO), they demonstrated that differences in stressor types between datasets have the greatest negative impact on transferability, more so than sensor brands or stress severity variations. Despite utilizing deep learning models with labeled data, their reported cross-dataset accuracies remained modest, typically around 60–70%, highlighting the severe limitations caused by dependence on labels and inconsistent stress induction protocols. This work clearly illustrates the inherent challenges of supervised transfer learning for real-world stress recognition systems.

In an effort to address these limitations, more recent studies have explored self-supervised learning (SSL) techniques to reduce reliance on labeled data. Matton et al. [5] proposed a contrastive SSL framework using electrodermal activity (EDA) signals to enhance generalization across datasets. Their results showed that SSL approaches could outperform traditional supervised transfer learning in certain conditions, emphasizing the potential of label-free representation learning. However, their study was restricted to a single physiological modality and laboratory-collected datasets, limiting the broader applicability of their findings.

Existing SSL-based approaches are either restricted to a single modality or evaluated with a fixed encoder architecture, leaving the interaction between architectural inductive bias and cross-dataset generalization largely unexplored. In contrast to prior work, our study proposes a multimodal self-supervised learning framework for affect recognition using physiological signals. By leveraging multiple modalities (EDA, temperature, PPG) and employing contrastive learning objectives, we capture richer, more transferable representations. Moreover, unlike prior work such as SIM-CNN [6], which evaluated SSL models on a single real-world dataset without assessing generalizability, we validate our models across both laboratory and daily life datasets in a cross-dataset setup, highlighting their robustness across heterogeneous sources, systematically addressing the limitations caused by labeling inconsistencies and stress induction variability. To the best of our knowledge, this is the first work to demonstrate robust cross-dataset generalizability in multimodal daily life stress recognition using self-supervised learning.

III. PREPROCESSING AND DESCRIPTION OF DATASETS

Reliable evaluation of cross-dataset generalizability in stress recognition requires access to diverse datasets collected under varying conditions. In this study, we selected a combination of laboratory-based and daily life datasets that cover a wide range of stress induction methods, sensor modalities, recording environments, and labeling strategies. This selection enables a comprehensive analysis of how self-supervised models can generalize across controlled experimental settings and the more complex, noisy conditions encountered in real-world applications. In the following subsections, we provide detailed descriptions of the datasets used and the preprocessing steps applied to the raw physiological signals.

A. WESAD

The Wearable Stress and Affect Detection (WESAD) dataset [19], based on the Trier Social Stress Test (TSST), contains data from 15 participants and was collected in a controlled laboratory setting. It encompassed conditions such as amusement, stress, meditation, and recovery. Self-reports included assessments from the Positive and Negative Affect Schedule (PANAS), State-Trait Anxiety Inventory (STAI), and Likert scale questions (stress, frustration, happy, and sad). The recorded physiological signals included ECG, EDA, EMG, PPG, respiration, accelerometer, and skin temperature, spanning a duration of 2 hours. Specifically for this study, we focused on three primary modalities: skin temperature, electrodermal activity, and blood volume pressure. Employing established preprocessing techniques [19], raw EDA and TEMP signals underwent low-pass Butterworth filtering (cutoff frequency: 0.5 Hz), followed by standard deviation normalization and downsampling to 4 Hz to normalize and expedite computation. The four states, baseline, amusement, stress, and meditation, were condensed into two classes: stress and non-stress.

B. VERBIO

The VERBIO dataset [20] comprises bio-behavioral responses and self-reported data collected during public speaking presentations, both in real-life and virtual settings. Participants delivered 10 presentations each, lasting approximately 5 minutes, across three segments of the study conducted over four days: PRE (1 session, Day 1), TEST (8 sessions, Days 2-3), and POST (1 session, Day 4). The PRE and POST segments involved real-life audiences, while the TEST segment involved various virtual audiences, distributed across two days to prevent participant overexertion. The study resulted in 10,800 minutes of acoustic and physiological data from 82 real and 216 virtual reality (VR) presentations. Only physiological data collected using the Empatica E4 device, with identical sampling rates and modalities as the WESAD dataset, were utilized.

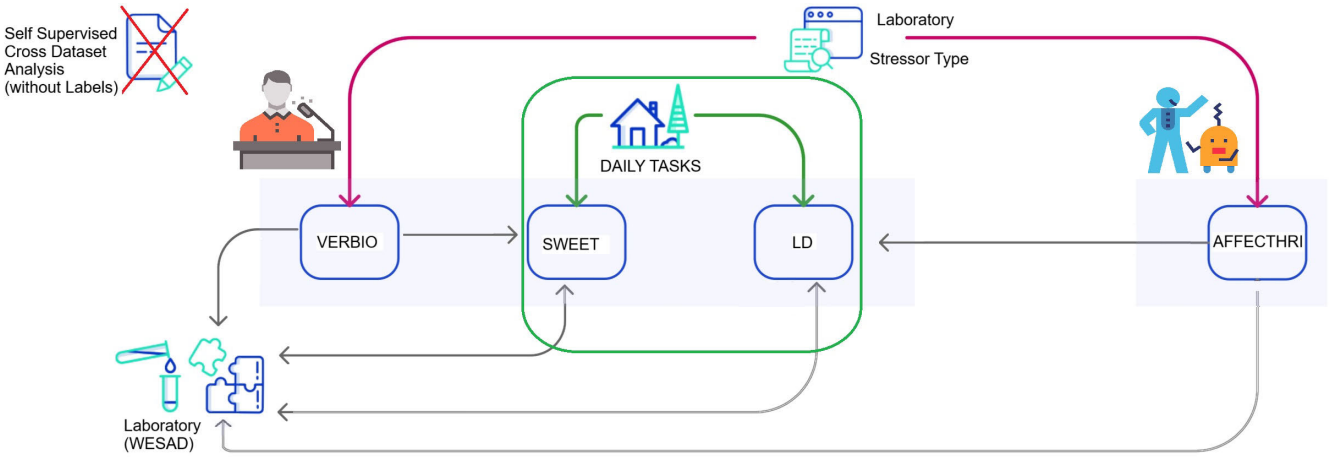


FIGURE 1. Multimodal self supervised cross dataset generalisability analysis methodology. LD is an abbreviation of the LabtoDaily dataset.

C. AffectHRI

The AffectHRI dataset is a comprehensive collection of data gathered from a study on human-robot interaction (HRI). This dataset includes physiological data labeled with human affect (emotions and mood), along with questionnaire ratings regarding affects provided by 146 participants. The physiological signals are recorded with the Empatica E4 wristband (Photoplethysmography (PPG), Electrodermal Activity (EDA), Skin Temperature and Acceleration). The physiological data was labeled with human affect (emotions and mood) based on the Self-Assessment Manikins (SAM) scale. We used the binary Arousal label in this study. For more details, please refer to the dataset description paper [21].

D. SWEET

Imec's SWEET (Stress in the Work Environment) dataset [22] is the largest of its kind, utilizing wearable technology to explore the relationship between stress and physiological factors. It was collected in Leuven, Belgium and consists of data from over 1,000 participants and provided researchers with a specific subset of 179 participants for focused analysis. Participants wore clinical-grade wristbands and wireless ECG patches continuously for five days, capturing comprehensive physiological data including heart rate, heart rate variability, skin conductance, skin temperature, and movement. ECG data is collected in one sample in one minute and we downsampled acceleration, electrodermal activity, and skin temperature signals to this sampling rate. This creates a difference from the other datasets, where we downsampled to four samples per second. However, we want to include ECG data and its sample rate forced us to downsample other signals for this purpose. The physiological data were supplemented by contextual information from smartphones, such as GPS data, phone activity, noise levels, and self-reported stress levels and daily activities. The dataset, enriched with physiological and contextual information, aims to facilitate the development of personalized, context-sensitive feedback systems.

E. LabToDaily (LD)

Can et al. [18] conducted a daily life experiment involving 14 university students aged between 20 and 25. Each participant wore Empatica E4 smart bands for one week and was instructed to wear them for twelve hours a day in their daily routine. Participants completed an online version of the Perceived Stress Scale (PSS-5) questionnaire every three hours, assessing five emotions on a 6-point Likert scale. The total stress score ranged from 0 to 30, divided into low (0-15) and high (15-30) perceived stress categories. A total of 989 hours of physiological data and 332 self-reports were obtained, with some sessions containing missing Ecological Momentary Assessments (EMAs), resulting in the exclusion of their corresponding physiological data. The dataset exhibited an imbalance in the number of samples between stress and relaxation classes, with 73% of the data labeled as relaxed and 27% as stressed.

F. PREPROCESSING

Following established preprocessing methods [19], raw EDA, BVP, ACC, and TEMP signals underwent low-pass Butterworth filtering (cutoff frequency: 0.5 Hz). Standard deviation normalization and downsampling to 4 Hz were applied to normalize and facilitate faster computation. Furthermore, based on previous work [23], we segmented the signal recordings of all datasets into windows of length 60 s with around 99.5% overlap, which corresponds to one sample shift. After that, these processed physiological signals are fed into the self-supervised architectures for classification purposes. We chose not to use handcrafted features since raw data is commonly used with self-supervised deep learning architectures [23] and using raw data directly with them provides better performance when compared to handcrafted features [24].

IV. METHODOLOGY

In this study, we propose a unified experimental framework to evaluate the cross-dataset generalizability of self-supervised

learning (SSL) approaches using multimodal physiological signals (see Figure 1). Unlike previous works that focus on maximizing performance within a single dataset, our core objective is to analyze whether representations learned via contrastive learning can transfer effectively across datasets that differ substantially in collection protocols, environments, sensor configurations, and subject populations.

Importantly, our goal is not to introduce or optimize novel deep learning architectures per se, but rather to investigate how different commonly used time-series encoders behave under identical self-supervised and transfer learning conditions. To this end, we evaluate multiple encoder families under the same training protocol, augmentation strategy, and evaluation setting, enabling a controlled and architecture-aware analysis of cross-dataset generalization.

A. SELF-SUPERVISED CONTRASTIVE LEARNING FOR PHYSIOLOGICAL SIGNALS

Self-supervised contrastive learning aims to learn informative representations from unlabeled data by encouraging agreement between different augmented views of the same input while pushing apart representations of different samples. Given a physiological signal segment, multiple transformations are applied to generate positive pairs, whereas segments originating from different temporal windows or subjects form negative pairs. The model is optimized using a contrastive objective to minimize the distance between positive pairs and maximize the distance between negatives.

In this work, we adopt a contrastive learning formulation based on the NT-Xent loss with cosine similarity and a fixed temperature of 0.4. During pretraining, only unlabeled physiological signals are used, and no task-specific supervision is introduced. This setup allows the learned representations to remain agnostic to dataset-specific stress labels and protocols.

B. APPLIED TRANSFORMATION TYPES

To construct diverse yet semantically consistent views of physiological signals, we apply a set of time-series–specific augmentations that have been shown to be effective for contrastive learning in biosignal analysis. The applied transformations include:

1) NOISING

Random Gaussian noise is added to the input signal to encourage robustness against sensor noise and real-world measurement artifacts.

2) NEGATING

The signal polarity is inverted, forcing the encoder to learn polarity-invariant representations rather than relying on absolute signal direction.

3) SCALING

Signal amplitudes are scaled by random factors to promote invariance to subject-dependent signal magnitude differences.

4) PERMUTING

Temporal segments of the signal are randomly permuted, encouraging the model to focus on local temporal patterns rather than absolute ordering.

5) TIME SHIFTING

Signals are shifted along the temporal axis to increase robustness against misalignment and windowing effects.

6) TS-TCC

In addition to basic augmentations, we incorporate temporal and contextual contrasting (TS-TCC) as described in [25], which explicitly contrasts temporal dynamics across different contextual views of the same signal.

Unless otherwise stated, the Noise–Permute transformation pair is used for all SSL experiments, as it consistently yielded the most stable validation behavior across datasets.

C. ENCODER ARCHITECTURES

To study the effect of architectural bias on representation transfer, we evaluate four encoder families that are commonly used for time-series and physiological signal modeling. All encoders are configured to produce a fixed 64-dimensional embedding to ensure fair comparison across architectures.

1) CNN ENCODER

The baseline CNN encoder follows the architecture originally proposed by Cheng et al. [27] and later adapted for multimodal contrastive learning by Can et al. [26]. It consists of a stack of convolutional and residual blocks with progressively increasing receptive fields, enabling hierarchical feature extraction from raw signals. The architecture was extended to support four-channel multimodal input (BVP, EDA, TEMP, ACC) while preserving the original representation dimensionality.

2) TEMPORAL CONVOLUTIONAL NETWORK (TCN)

To explicitly model temporal dependencies with long receptive fields, we include a Temporal Convolutional Network (TCN) encoder. The TCN consists of dilated causal convolutions arranged in multiple layers, allowing it to capture long-range temporal patterns while maintaining computational efficiency. The TCN encoder is configured with two residual blocks and outputs a 64-dimensional embedding.

3) ResNet34-1D

We further evaluate a 1D adaptation of the ResNet34 architecture, which introduces deeper residual connections to facilitate learning of complex temporal hierarchies. The ResNet34-1D encoder uses stacked residual blocks with increasing channel widths and kernel sizes tailored for physiological signals. Compared to the CNN baseline, this architecture provides greater depth and representational capacity.

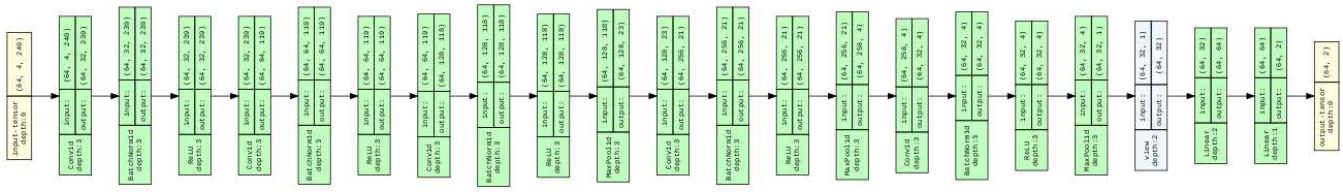


FIGURE 2. Architecture diagram for the used CNN architecture [26].

4) CNN–TRANSFORMER ENCODER

Finally, we include a hybrid CNN–Transformer encoder that combines local feature extraction via convolutional layers with global temporal modeling via multi-head self-attention. The CNN front-end extracts short-term temporal patterns, which are then segmented and processed by a Transformer encoder with four attention heads. This design allows the model to capture both local and long-range dependencies within physiological signals.

D. PRETRAINING STRATEGY (UPSTREAM TRAINING)

All encoders are pretrained using self-supervised contrastive learning on the source dataset. The upstream training phase is conducted for a maximum of 25 epochs using the Adam optimizer with a learning rate of 10^{-3} and weight decay of 10^{-4}. Early stopping with a patience of five epochs is applied based on validation contrastive loss to prevent overfitting and unnecessary computation.

During upstream training, only the encoder and projection head are optimized. The projection head is discarded after pretraining, and only the encoder weights are transferred to downstream tasks. For reproducibility and transparency, the best-performing model checkpoint and full training curves are stored for each encoder–dataset combination.

E. DOWNSTREAM FINE-TUNING AND EVALUATION

For downstream evaluation, pretrained encoders are transferred to the WESAD dataset and combined with a lightweight linear classification head. In all reported experiments, downstream fine-tuning is deliberately limited to a single epoch. This design choice emphasizes representation quality rather than task-specific overfitting and ensures that performance gains stem primarily from the pretrained embeddings.

All downstream evaluations are conducted using a leave-one-subject-out (LOSO) protocol, enforcing strict subject-level separation between training and testing data. In each fold, the encoder is either frozen or minimally fine-tuned, and only the linear classifier is optimized using cross-entropy loss. Performance is reported using weighted F1-score to account for class imbalance.

This unified training and evaluation protocol enables a fair comparison across learning paradigms (supervised vs. self-supervised), datasets, modalities, and architectures, while closely reflecting realistic deployment scenarios for wearable stress recognition systems.

V. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present an evaluation of our self-supervised contrastive learning framework for cross-dataset stress recognition using multimodal physiological signals. All downstream evaluations in this study are conducted using a leave-one-subject-out (LOSO) cross-validation protocol. This choice was made deliberately to ensure strict subject-level separation between training and testing data and to avoid optimistic bias that may arise from random data splits. Under LOSO, each subject is held out once as the test set while the remaining subjects are used for training, providing a realistic assessment of cross-subject generalization performance.

Performance is reported using the F1-score, which balances precision and recall and is particularly suitable for stress recognition tasks where class imbalance is common. Unless otherwise stated, all reported results correspond to LOSO evaluation on the WESAD dataset.

A. HYPERPARAMETER OPTIMIZATION

We adopted the key hyperparameter values—namely the contrastive loss temperature and the augmentation transformations—from our previous study on unimodal physiological stress recognition using self-supervised learning [26]. In that study, an extensive grid search was conducted on the WESAD dataset to evaluate the impact of different augmentation combinations and temperature values on the quality of learned representations.

Among all transformation pairs, the combination of Gaussian noise injection and permutation (denoted as Noise-Permute) yielded the most robust results. This pair was effective in introducing both temporal and amplitude variability, encouraging the model to learn invariant features across these dimensions. Noise injection simulates sensor-level fluctuations, while permutation disrupts local temporal dependencies—forcing the encoder to capture more global and stable physiological patterns.

Additionally, the temperature parameter of the NT-Xent contrastive loss was tuned. A value of 0.4 was found to yield optimal balance between hard-negative separation (see Table 2) and gradient stability. Lower values (e.g., 0.1) led to vanishing gradients, while higher values (e.g., 0.8) reduced the discriminative power of the learned representations. As a result, we fixed the temperature at 0.4 and used the Noise-Permute augmentation pair for all contrastive learning experiments in this study.

By keeping these hyperparameters fixed across all datasets and modalities, we aimed to eliminate potential tuning bias and

TABLE 2. Hyperparameter optimization results.

Parameter	Best Value
Temperature	0.4
Transformation	Noise-Permute

TABLE 3. VERBIO SSL modality ablation results under WESAD LOSO (CNN-Transformer encoder).

Modality	Accelerometer	PPG	EDA	Skin Temperature	All
F1-score (%)	81.39	72.96	81.82	71.87	**96.38**

TABLE 4. AffectHRI SSL modality ablation results under WESAD LOSO (CNN-Transformer encoder).

Modality	Accelerometer	PPG	EDA	Skin Temperature	All
F1-score (%)	73.36	72.06	75.36	80.27	**94.79**

TABLE 5. Supervised VERBIO modality ablation under WESAD LOSO (CNN-Transformer encoder).

Modality	Accelerometer	PPG	EDA	Skin Temperature	All
F1-score (%)	80.27	72.32	91.01	73.54	**97.14**

ensure a fair evaluation of generalizability. Our goal was not to tailor each configuration for peak accuracy, but to test whether a robust configuration derived from one dataset can generalize well across others—a realistic requirement for scalable, real-world stress recognition systems.

B. COMPARISON OF DEEP TRANSFER LEARNING AND SELF-SUPERVISED LEARNING

This section provides a comparison between supervised deep transfer learning (DTL) and the proposed self-supervised contrastive learning (SSL) framework for cross-dataset stress recognition under a strict LOSO evaluation protocol. To ensure a fair comparison, both learning paradigms employ identical encoder architectures and downstream classifiers. The only difference lies in the pretraining strategy—supervised DTL relies on labeled source data, whereas SSL leverages unlabeled data through contrastive objectives. This setup allows us to isolate the effect of representation learning strategy without confounding architectural factors.

Despite not using any stress labels during pretraining, VERBIO-SSL achieves a strong multimodal F1-score of 96.38% under LOSO, in contrast, supervised VERBIO pretraining, which relies on fully labeled source data, reaches a slightly higher multimodal performance of 97.14% (as reported in Tables 3 and 5). The difference between the two approaches is therefore marginal (0.76 percentage points), indicating that self-supervised learning can recover most of the performance of supervised transfer even without access to labels.

For reference, an intra-dataset baseline using the same CNN–Transformer architecture trained and evaluated on WESAD under LOSO yields an F1-score of 97.63% (see Table 7). This result represents a near upper bound under full alignment of source and target distributions. That VERBIO-SSL approaches this reference performance despite substantial differences in stress induction protocols and without label supervision highlights the effectiveness of contrastive self-supervised representations.

C. MODALITY ABLATION ANALYSIS

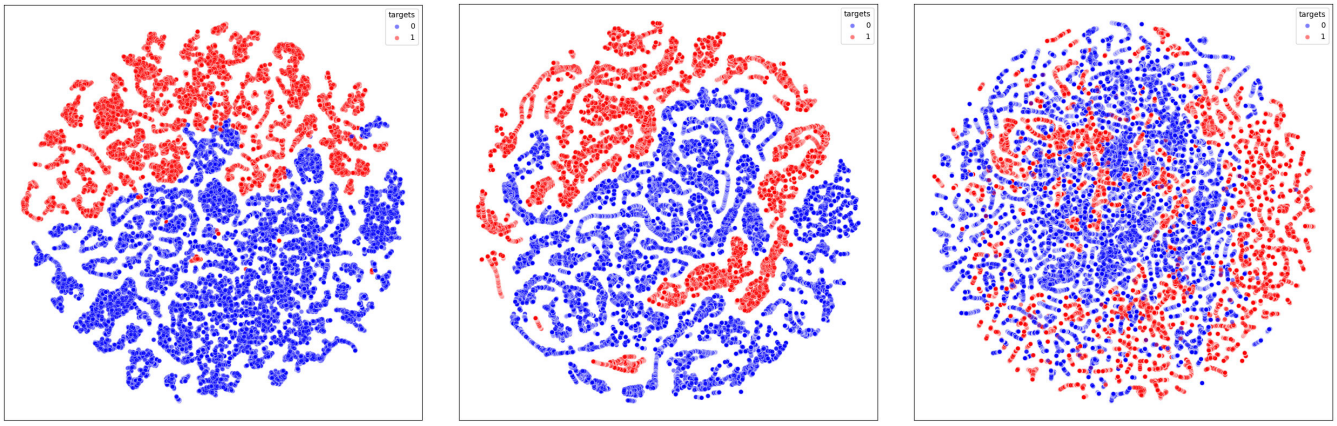
We also conducted a modality ablation analysis under a strict leave-one-subject-out (LOSO) evaluation protocol. In these experiments, models were pretrained using a single modality at a time and subsequently evaluated on WESAD, allowing direct comparison with their corresponding multimodal counterparts. This setup enables a clear separation between the effect of individual sensors and the benefits arising from multimodal representation learning.

The results for SSL pretraining on VERBIO are summarized in Table 3. When only a single modality is used, SSL achieves moderate performance, with F1-scores ranging between 71.87% and 81.82% across modalities. In contrast, combining all four modalities results in a substantial performance increase, reaching 96.38% under the same LOSO protocol. Importantly, this multimodal performance is markedly higher than the best unimodal SSL result, indicating that the gains achieved by SSL cannot be attributed to a single dominant physiological signal but instead emerge from the complementary integration of multiple modalities.

A similar pattern is observed for AffectHRI-based SSL pretraining (Table 4). Although AffectHRI is collected in a laboratory environment, unimodal SSL performance remains relatively limited, with F1-scores between 72.06% and 80.27%. This behavior is likely influenced by two factors. First, the dataset is collected in a human–robot interaction setting, where stressors differ in nature from the social-evaluative (WESAD) and public-speaking (VERBIO) paradigms, introducing an additional form of domain shift at the modality level. Second, AffectHRI is comparatively smaller, which can limit the effectiveness of unimodal representation learning. Nevertheless, when all modalities are combined, SSL achieves a strong multimodal performance of 94.79%, again substantially exceeding all unimodal results and underscoring the benefit of multimodal fusion.

To further contextualize these findings, we compare SSL-based modality ablation results with a supervised VERBIO pretraining baseline under the same LOSO evaluation setting (Table 5). Supervised pretraining achieves a slightly higher multimodal F1-score of 97.14%, but its unimodal behavior is more uneven. In particular, supervised learning attains a notably high unimodal performance on EDA (modality 5), reaching 91.01%, outperforming the corresponding SSL unimodal configurations. This suggests that when a single modality is highly informative and well-aligned between source and target datasets, supervised learning can more effectively exploit modality-specific label information. However, this advantage does not generalize uniformly across modalities, as other supervised unimodal configurations show substantial performance drops.

Across all settings, a consistent trend emerges: multimodal configurations significantly outperform even the best unimodal results, regardless of whether the model is pretrained



((a)) a. WESAD → WESAD (LOSO) ((b)) b. VERBIO → WESAD ((c)) c. AffectHRI → WESAD

FIGURE 3. T-SNE projections of learned representations from the SSL-based CNN model. Subfigure (a) shows intra-dataset embedding clusters using LOSO on WESAD. Subfigures (b) and (c) visualize cross-dataset transfer scenarios from VERBIO and AffectHRI to WESAD, respectively.

TABLE 6. WESAD LOSO transfer performance (F1-score %) using different SSL pretraining sources and encoder architectures.

Pretraining Source	CNN	TCN	ResNet34-1D	CNN-Transformer
LD-SSL	98.43	96.41	98.98	93.89
VERBIO-SSL (All)	99.39	98.58	99.16	96.38
AffectHRI-SSL (All)	98.61	98.43	98.07	94.79

TABLE 7. Cross-dataset transfer results between laboratory datasets. CNN_Transformer architecture was used.

From → To	WESAD
WESAD	97.63%
VERBIO	96.38%
AffectHRI	94.79%

TABLE 8. Cross-dataset transfer results from laboratory to daily life datasets. CNN architecture was used.

From → To	SWEET	LD
VERBIO →	90.06%	76.13%
WESAD →	89.52%	75.77%
AffectHRI →	87.04%	75.10%

in a supervised or self-supervised manner. This indicates that multimodal fusion provides information that is not recoverable from any single physiological channel alone. While supervised learning may retain an advantage in specific, well-aligned unimodal cases (such as EDA), self-supervised learning offers more balanced behavior across modalities and achieves strong multimodal performance without reliance on labels.

Overall, these findings confirm that the primary source of performance gains in cross-dataset stress recognition lies in synergistic multimodal representation learning rather than modality-specific dominance. They also highlight a complementary relationship between supervised and self-supervised paradigms: supervised learning can excel when a dominant modality is available and well aligned, whereas self-supervised learning provides a more robust and label-efficient foundation for multimodal and cross-dataset scenarios.

D. ARCHITECTURE-WISE COMPARISON UNDER LOSO EVALUATION

Beyond the learning paradigm itself, the choice of encoder architecture influences how effectively learned representations generalize across subjects and datasets. To examine this effect in a controlled manner, we evaluated four commonly used encoder families under the same self-supervised learning (SSL) framework and leave-one-subject-out (LOSO) evaluation protocol on WESAD: a conventional CNN, a Temporal Convolutional Network (TCN), a ResNet34-1D, and a CNN–Transformer hybrid. Importantly, all architectures were pretrained and transferred under identical conditions, allowing differences in performance to be attributed primarily to architectural inductive bias rather than training procedure.

The results in Table 6 show that all encoder architectures benefit from SSL pretraining, achieving consistently high F1-scores across different source datasets. Among them, ResNet34-1D attains the highest or near-highest performance in most settings, reaching up to 99.16% when pretrained on VERBIO and 98.98% when pretrained on LD. This suggests that deeper residual architectures provide a favorable inductive bias for cross-subject generalization, potentially due to improved optimization stability and their ability to capture hierarchical temporal patterns in physiological signals.

The standard CNN encoder also demonstrates strong and competitive performance, particularly when pretrained on large-scale or heterogeneous datasets. For instance, CNN-based models achieve F1-scores of 99.39% with VERBIO-SSL and 98.43% with LD-SSL. These results indicate that relatively shallow convolutional architectures can generalize effectively when paired with diverse self-supervised pretraining data. Compared to ResNet-based models, however, CNNs exhibit slightly greater sensitivity to the choice of pretraining source, suggesting a stronger dependence on data diversity to achieve optimal transfer.

TABLE 9. Cross-dataset transfer results from daily life to laboratory datasets. TCN architecture was used.

From → To	WESAD
SWEET → WESAD	94.36%
LD → WESAD	96.41%

TCN-based encoders achieve robust but moderately lower performance overall, with F1-scores ranging from 96.41% to 98.58%. While dilated temporal convolutions are designed to model long-range dependencies, their fixed receptive field structure may offer less flexibility under strong subject-level variability, which is a defining characteristic of LOSO evaluation in physiological data.

The CNN–Transformer hybrid exhibits a more mixed behavior. Although it achieves competitive performance in several configurations (e.g., 96.38% with VERBIO-SSL), it consistently underperforms compared to CNN and ResNet34-1D under LOSO transfer. This suggests that self-attention mechanisms, while powerful, may require either larger datasets or more stable temporal alignment to fully realize their advantages in physiological signal modeling. In subject-wise transfer settings with limited per-subject data, the additional model complexity may increase sensitivity to noise and inter-subject variability.

Overall, these results indicate that architectural inductive bias plays an important role in cross-subject generalization, even when the learning paradigm and training protocol are held constant. Deep residual architectures emerge as the most reliable choice under LOSO evaluation, while simpler convolutional models remain strong and efficient baselines when coupled with sufficiently diverse SSL pretraining data. More complex hybrid architectures, such as CNN–Transformer models, appear more sensitive to data scale and distribution, underscoring the importance of matching architectural complexity to the characteristics of physiological datasets.

E. CROSS-DATASET PERFORMANCE: LAB-TO-LAB

To analyze cross-dataset generalization under controlled conditions while eliminating architectural confounds, we focus on laboratory-to-laboratory transfer results obtained with a fixed CNN–Transformer encoder under a strict leave-one-subject-out (LOSO) evaluation protocol. This comparison, summarized in Table 7, includes both cross-dataset transfers (VERBIO → WESAD, AffectHRI → WESAD) and the intra-dataset reference (WESAD → WESAD), enabling a direct assessment of representation transferability across laboratory settings.

Despite differences in experimental design, self-supervised pretraining on VERBIO achieves a high F1-score of 96.38% when transferred to WESAD, closely approaching the intra-dataset WESAD baseline of 97.63%. Pretraining on AffectHRI yields a slightly lower but still robust performance of 94.79%. These results demonstrate that SSL representations learned from one laboratory dataset can generalize effectively to another, even when stress elicitation protocols are not identical.

The particularly strong transfer performance from VERBIO to WESAD can be attributed to two complementary factors. First, VERBIO is substantially larger and more diverse, providing richer representation learning during self-supervised pretraining. Second, the public speaking paradigm employed in VERBIO constitutes a core component of the Trier Social Stress Test used in WESAD. This partial overlap in stressor type likely facilitates alignment at the representational level, enabling more effective transfer despite differences in overall protocol structure.

In contrast, AffectHRI relies on stress and affect elicitation within human–robot interaction scenarios, which differ conceptually from both social-evaluative and public speaking stressors. Combined with the smaller scale of the dataset, this divergence likely contributes to the larger performance gap observed relative to the intra-dataset baseline.

These quantitative trends are further supported by the qualitative visualization results shown in Fig. 3. The t-SNE projections reveal that representations learned and evaluated within WESAD exhibit the clearest class separation, whereas VERBIO-to-WESAD transfer shows moderately increased overlap between classes. In contrast, AffectHRI-to-WESAD representations appear noticeably more entangled, reflecting the greater mismatch in stressor characteristics and dataset structure. This progressive degradation in cluster separability closely mirrors the performance differences observed in Table 7.

Taken together, these lab-to-lab results indicate that transfer performance in self-supervised learning is governed not only by dataset scale and diversity, but also by the degree of conceptual overlap in stressor type between source and target datasets. Importantly, the ability of VERBIO-pretrained models to approach the WESAD intra-dataset upper bound suggests that contrastive SSL can effectively exploit both representation diversity and stressor alignment without relying on explicit label supervision. This controlled comparison provides a meaningful reference for interpreting the larger domain shifts examined in subsequent laboratory-to-daily-life and daily-life-to-laboratory analyses.

F. CROSS-DATASET PERFORMANCE: LAB-TO-DAILY

While laboratory-to-laboratory transfer provides insight under controlled conditions, real-world deployment of stress recognition systems requires robustness to the variability inherent in daily life data. To evaluate this scenario, models pretrained on laboratory datasets of varying scale and diversity (VERBIO, WESAD, and AffectHRI) were transferred to two daily life datasets, SWEET and LD.

As expected, performance decreases compared to lab-to-lab transfer due to increased noise, contextual variability, and subject diversity. Nevertheless, SSL-based models maintain competitive performance. Models pretrained on VERBIO achieve the highest transfer performance, reaching 90.06% F1-score on SWEET and 76.13% on LD, followed

by WESAD-pretrained models with 89.52% and 75.77%, respectively (see Table 8). AffectHRI-based pretraining yields slightly lower but still robust performance on both targets.

This consistent ranking—VERBIO outperforming WESAD, followed by AffectHRI—mirrors the relative scale and diversity of the laboratory source datasets. VERBIO, being the largest and most varied laboratory dataset, provides a richer pool of physiological patterns during self-supervised pretraining, which translates into more transferable representations. WESAD, while well-structured and widely used, is smaller in scale and exhibits more limited variability, whereas AffectHRI is both smaller and centered around a more specific interaction-driven context. These factors likely constrain the breadth of representations learned during pretraining and, consequently, their effectiveness under strong domain shift.

The performance gap between SWEET and LD further reflects differences in target data characteristics. SWEET, although collected in naturalistic settings, benefits from higher signal quality and semi-structured reporting protocols. In contrast, LD represents a highly unconstrained daily life scenario with substantial variability in activities, sensor placement, and contextual factors. That SSL models achieve F1-scores above 75% on LD nevertheless highlights their ability to extract domain-robust physiological representations despite severe distribution shift.

Importantly, all reported results are obtained without using any labels during pretraining, underscoring the suitability of contrastive self-supervised learning for real-world deployment scenarios where annotation is expensive, noisy, or infeasible. Overall, the lab-to-daily transfer results emphasize the importance of dataset scale and representational diversity in self-supervised learning and demonstrate that, while daily life generalization remains challenging, SSL provides a scalable and effective foundation for stress recognition in uncontrolled environments.

G. CROSS-DATASET PERFORMANCE: DAILY-TO-LAB

We further evaluate cross-dataset generalization in the reverse direction by pretraining models on daily life datasets (SWEET and LD) and transferring them to the laboratory dataset WESAD. Contrary to the common assumption that controlled laboratory data provide the most effective pretraining signal, this direction yields the strongest overall transfer performance.

Models pretrained in a self-supervised manner on LD and SWEET achieve F1-scores of 96.41% and 94.36% on WESAD under LOSO evaluation, respectively (see Table 9). Notably, LD-based pretraining slightly outperforms VERBIO-based lab pretraining, despite LD originating from a substantially different and less controlled daily life environment. This result is particularly striking given that LD labels are subjective and noisy by nature; however, since no labels are used during SSL pretraining, these ambiguities do not hinder representation learning.

The superior performance of LD can be attributed primarily to its scale and diversity. LD contains long-duration recordings collected across unconstrained daily activities, capturing a wide range of physiological states and contextual variations. This richness enables SSL to learn highly generalizable and invariant representations that transfer effectively even to controlled laboratory settings such as WESAD. In this sense, representation diversity appears to outweigh protocol mismatch when labels are removed from the learning objective.

In contrast, SWEET-based pretraining yields lower transfer performance despite also being a daily life dataset. This behavior can be explained by the comparatively shorter recording durations, lower diversity, and more limited variability in SWEET, which restrict the richness of representations learned during self-supervised training. As a result, SWEET pretraining is less effective when transferred to WESAD and falls below several laboratory-based pretraining configurations.

Overall, these daily-to-lab results highlight that the effectiveness of self-supervised pretraining is driven not simply by whether data originate from laboratory or daily life settings, but by the scale, duration, and diversity of the source dataset. Large and heterogeneous daily life datasets such as LD can provide even stronger pretraining signals than laboratory datasets, whereas smaller or less diverse daily datasets may not fully realize this advantage. These findings reinforce the importance of leveraging large-scale, unconstrained data for robust cross-context generalization in self-supervised physiological stress recognition.

H. COMPUTATIONAL COST AND EFFICIENCY ANALYSIS

In addition to recognition performance, we analyze the computational characteristics of the evaluated encoder architectures to assess their suitability for real-world and wearable deployments. Specifically, we report encoder parameter counts and total upstream (pretraining) time for self-supervised learning (SSL) across the laboratory datasets. Since the upstream SSL stage optimizes both the encoder and a projection head but transfers only the encoder weights to downstream tasks, all reported parameter counts correspond to the encoder alone.

Furthermore, we provide a relative inference latency comparison across architectures based on model complexity and parameter scale. While absolute latency and energy consumption depend on the target hardware and software stack, these proxy metrics provide practical guidance for deployability and are commonly used in wearable and embedded machine learning studies.

1) UPSTREAM TRAINING COST ACROSS LABORATORY DATASETS

Table 10 summarizes the computational characteristics of upstream SSL runs on the laboratory datasets (AffectHRI and VERBIO). Clear trade-offs between model capacity and training cost can be observed. Lightweight convolutional encoders (CNN) exhibit the smallest parameter footprint and the lowest total training time, whereas deeper residual

TABLE 10. Computational characteristics of upstream self-supervised pretraining across laboratory datasets and encoder architectures.

Pretraining Dataset	Encoder	#Params	Epochs	Total Time (s)	Avg / Epoch (s)
AffectHRI (SSL)	CNN	57,676	25	42,947.56	1,717.90
AffectHRI (SSL)	TCN	62,048	2	28,603.61	14,301.81
AffectHRI (SSL)	CNN-Transformer	672,226	19	120,261.47	6,329.55
AffectHRI (SSL)	ResNet34-1D	4,177,536	19	154,154.62	8,113.40
VERBIO (SSL)	CNN	57,676	25	39,470.88	1,578.84
VERBIO (SSL)	TCN	62,048	14	25,764.89	1,840.35
VERBIO (SSL)	CNN-Transformer	672,226	21	116,057.38	5,526.54
VERBIO (SSL)	ResNet34-1D	4,177,536	23	150,029.64	6,523.03

TABLE 11. Measured computational metrics for supervised VERBIO training (CNN–Transformer was used).

Setting	#Params	Total Time (s)	Latency Mean (s)	Latency (Best F1)
VERBIO (Supervised)	672,226	70,126.73	0.133	0.185

TABLE 12. Relative inference latency ranking based on architectural complexity.

Encoder	Relative Inference Latency
CNN	Very Low
TCN	Low
CNN–Transformer	Medium
ResNet34-1D	High

and attention-based architectures require substantially higher computational resources.

2) SUPERVISED VS. SELF-SUPERVISED TRAINING COST

To contextualize the SSL computational cost, Table 11 reports the measured training time and inference latency for the supervised VERBIO experiment using the CNN–Transformer encoder. Compared to its SSL counterpart, supervised training exhibits lower total training time, but it relies on labeled source data and is inherently tied to dataset-specific label definitions and stress induction conditions, as discussed in earlier sections.

The reported inference latency values quantify the wall-clock time required for a forward pass of the trained model under the measurement setup used in this study. Specifically, *Latency Mean (s)* corresponds to the average inference time per prediction during evaluation, whereas *Latency (Best F1)* reports the inference time measured at the checkpoint that achieved the best F1-score. For example, a mean latency of approximately 0.133 s indicates that, under the current evaluation configuration, the model requires on the order of one-tenth of a second to produce a prediction. These values should be interpreted as indicative rather than absolute, as latency depends on hardware, batch size, and implementation details.

3) INFERENCE LATENCY CONSIDERATIONS

For SSL experiments, we provide a relative inference latency ranking based on encoder complexity and parameter scale, which remains independent of hardware-specific implementations. As shown in Table 12, convolution-only models offer the lowest inference cost, while deep residual and attention-based architectures introduce higher computational overhead.

Overall, this analysis highlights a clear trade-off between representational capacity and computational efficiency. While deeper and more complex architectures such as ResNet34-1D and CNN–Transformer can yield strong cross-dataset performance, lightweight CNN-based models remain attractive for real-time, resource-constrained wearable applications. Importantly, the results demonstrate that self-supervised learning enables practitioners to navigate this trade-off by selecting architectures that balance performance and efficiency according to deployment requirements.

I. DISCUSSION

This study demonstrates that multimodal self-supervised learning (SSL) provides a robust and scalable foundation for physiological stress recognition under realistic cross-dataset and cross-subject evaluation settings. Across all transfer directions—laboratory-to-laboratory, laboratory-to-daily life, and daily-to-laboratory—SSL consistently enables strong generalization despite substantial differences in stress induction protocols, recording conditions, and subject populations.

A central insight emerging from our experiments is the importance of data scale and representational diversity during pretraining. Large and heterogeneous datasets, particularly unconstrained daily life recordings, yield more transferable representations than smaller or highly controlled laboratory datasets. This finding challenges the common assumption that clean, protocol-aligned data necessarily provide the best pretraining signal. Instead, the variability inherent in real-world data appears to act as an effective regularizer for contrastive objectives, promoting subject-invariant and context-robust representations.

Our results further clarify that SSL performance gains are not driven by individual physiological modalities. While supervised learning can outperform SSL in carefully aligned unimodal cases—most notably for EDA—such advantages remain modality-specific and do not translate into a substantial multimodal margin. In contrast, SSL consistently benefits from multimodal integration, achieving strong performance without reliance on a single dominant signal. This behavior is particularly relevant for wearable applications, where sensor quality and availability may vary across users and contexts.

Architectural comparisons indicate that inductive bias plays a critical role in cross-subject generalization. Residual

and convolutional architectures exhibit the most stable behavior under LOSO evaluation, whereas more complex attention-based models show increased sensitivity to data scale and distribution. These findings suggest that, in physiological signal modeling, architectural complexity alone does not guarantee better transferability and must be carefully matched to dataset characteristics.

Taken together, these observations position multimodal SSL as a principled alternative to supervised transfer learning for cross-dataset stress recognition. By decoupling representation learning from dataset-specific labels and stressor definitions, SSL mitigates key sources of bias and variability that have historically limited generalization in affective computing. From a practical perspective, this makes SSL particularly well suited for long-term, real-world stress monitoring systems, where large volumes of unlabeled physiological data are readily available but reliable annotation is costly or infeasible.

Beyond predictive performance, the reported parameter counts, training time, and inference latency highlight important deployability trade-offs for wearable and edge-device scenarios. The results show that lightweight CNN and TCN-based encoders achieve favorable efficiency–performance balance, while deeper architectures such as ResNet34-1D and CNN–Transformer offer higher representational capacity at increased computational cost. Notably, several self-supervised configurations achieve competitive cross-dataset performance with substantially lower model complexity and inference latency, supporting their practical suitability for real-world wearable stress monitoring where energy efficiency and responsiveness are critical constraints.

VI. CONCLUSION

In this work, we presented a comprehensive multimodal self-supervised learning (SSL) framework for physiological stress recognition and evaluated its generalization capability under a strict leave-one-subject-out (LOSO) protocol. By removing the reliance on labeled data during pretraining, the proposed approach addresses fundamental limitations of supervised transfer learning, including label subjectivity, dataset-specific bias, and reduced robustness under cross-subject and cross-dataset evaluation.

Across extensive experiments involving multiple laboratory and daily life datasets, we showed that self-supervised pretraining enables highly transferable representations across diverse recording conditions, stress elicitation protocols, and subject populations. In particular, pretraining on large-scale and heterogeneous daily life data consistently resulted in strong transfer performance to laboratory settings, in several cases approaching or surpassing intra-dataset baselines. These findings demonstrate that representational diversity and variability are more critical for generalization than strict experimental control within an SSL framework.

Our modality ablation analyses further revealed that SSL performance gains arise from synergistic multimodal representation learning rather than reliance on a single dominant physiological signal. While supervised transfer learning can achieve higher performance in carefully aligned unimodal cases—most notably for EDA—this advantage remains modality-specific and does not translate into a substantial multimodal margin. Under LOSO evaluation, supervised and self-supervised multimodal performance differs only marginally, highlighting SSL as a robust and label-efficient alternative for cross-dataset scenarios.

While modality ablation provides direct evidence for the contribution of multimodal fusion, more fine-grained attribution analyses remain an important direction for future work. Techniques such as Shapley-style contribution modeling or class activation and attention-based alignment could further elucidate how specific signal segments contribute to stress representations across domains. However, applying such methods to multimodal physiological time series under strong domain shift presents non-trivial challenges related to temporal alignment and interpretability. Addressing these challenges is left for future investigation.

Architecture-wise evaluations indicate that inductive bias plays a key role in cross-subject generalization. Residual and convolutional architectures provide the most stable performance under LOSO evaluation, while more complex attention-based hybrids show increased sensitivity to data scale and distribution. These results suggest that architectural complexity alone does not guarantee improved generalization in physiological signal modeling and must be carefully matched to dataset characteristics.

Overall, this study establishes multimodal self-supervised learning as a principled and practical approach for building robust affective computing systems. By leveraging large volumes of unlabeled physiological data and contrastive objectives, SSL offers a scalable pathway toward real-world stress recognition on wearable platforms. Future work will explore context-aware and temporally aligned self-supervised objectives, as well as adaptive personalization mechanisms, to further enhance robustness and applicability in long-term, real-life monitoring scenarios.

REFERENCES

[1] *Chronic Stress Puts Your Health at Risk*. Accessed: Jan. 11, 2024. [Online]. Available: https://www.mayoclinic.org/healthy-lifestyle/stress-management/in-depth/stress/art-20046037

[2] S. Gedam and S. Paul, "A review on mental stress detection using wearable sensors and machine learning techniques," *IEEE Access*, vol. 9, pp. 84045–84066, 2021.

[3] O. T. Ba aran, Y. S. Can, E. André, and C. Ersoy, "Relieving the burden of intensive labeling for stress monitoring in the wild by using semi-supervised learning," *Frontiers Psychol.*, vol. 14, Jan. 2024, Art. no. 1293513.

[4] M. Benchekroun, P. E. Velmovitsky, D. Istrate, V. Zalc, P. P. Morita, and D. Lenne, "Cross dataset analysis for generalizability of HRV-based stress detection models," *Sensors*, vol. 23, no. 4, p. 1807, Feb. 2023.

[5] K. Matton, R. Lewis, J. Guttag, and R. Picard, "Contrastive learning of electrodermal activity representations for stress detection," in *Proc. Conf. Health, Inference, Learn.*, 2023, pp. 410–426.

[6] S. Eom, S. Eom, and P. Washington, "SIM-CNN: Self-supervised individualized multimodal learning for stress prediction on nurses using biosignals," in *Proc. Workshop Mach. Learn. Multimodal Healthcare Data*. Cham, Switzerland: Springer, 2023, pp. 155–171.

[7] F. P. Akbulut, "Hybrid deep convolutional model-based emotion recognition using multiple physiological signals," *Comput. Methods Biomechanics Biomed. Eng.*, vol. 25, no. 15, pp. 1–13, Nov. 2022.

[8] P. Prajod, B. Mahesh, and E. André, "Stressor type matters! – exploring factors influencing cross-dataset generalizability of physiological stress detection," in *Proc. Int. Conf. Multimodel Interact.*, Nov. 2024, pp. 508–517.
[9] P. Prajod and E. André, "On the generalizability of ECG-based stress detection models," in *Proc. 21st IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2022, pp. 549–554.
[10] L. J. Cronbach and P. E. Meehl, "Construct validity in psychological tests," *Psychol. Bull.*, vol. 52, no. 4, pp. 281–302, Jul. 1955.
[11] S. Messick, "Validity of psychological assessment," *Amer. Psychol.*, vol. 50, no. 9, pp. 741–749, 1995.
[12] *Standards for Educational and Psychological Testing*. Washington, DC, USA: American Educational Research Association, 2014.
[13] R. S. Lazarus and S. Folkman, *Stress, Appraisal, and Coping*. Cham, Switzerland: Springer, 1984.
[14] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," *J. Health Social Behav.*, vol. 24, no. 4, pp. 385–396, Dec. 1983.
[15] S. P. Reise, "The rediscovery of bifactor measurement models," *Multivariate Behav. Res.*, vol. 47, no. 5, pp. 667–696, Sep. 2012.
[16] W. Meredith, "Measurement invariance, factor analysis and factorial invariance," *Psychometrika*, vol. 58, no. 4, pp. 525–543, Dec. 1993.
[17] Richard J. Vandenberg and Charles E. Lance, "A review and synthesis of the measurement invariance literature," *Organizational Res. Methods*, vol. 3, no. 1, pp. 4–70, 2000.
[18] Y. S. Can, D. Gokay, D. R. Kılıç, D. Ekiz, N. Chalabianloo, and C. Ersoy, "How laboratory experiments can be exploited for monitoring stress in the wild: A bridge between laboratory and daily life," *Sensors*, vol. 20, no. 3, p. 838, Feb. 2020.
[19] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 400–408.
[20] M. Yadav, M. N. Sakib, E. H. Nirjhar, K. Feng, A. H. Behzadan, and T. Chaspari, "Exploring individual differences of public speaking anxiety in real-life and virtual presentations," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1168–1182, Jul. 2022.
[21] J. S. Heinisch, J. Kirchhoff, P. Busch, J. Wendt, O. V. Stryk, and K. David, "Physiological data for affective computing in HRI with anthropomorphic service robots: The AFFECT-HRI data set," *Sci. Data*, vol. 11, no. 1, p. 333, 2024.
[22] E. Smets, E. Rios Velazquez, G. Schiavone, I. Chakroun, E. D'Hondt, W. De Raedt, J. Cornelis, O. Janssens, S. Van Hoecke, S. Claes, I. Van Diest, and C. Van Hoof, "Large-scale wearable data reveal digital phenotypes for daily-life stress detection," *npj Digit. Med.*, vol. 1, no. 1, p. 67, Dec. 2018.
[23] Y. Wu, M. Daoudi, and A. Amad, "Transformer-based self-supervised multimodal representation learning for wearable emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 15, no. 1, pp. 157–172, Jan. 2023.
[24] Y. S. Can and E. André, "Performance exploration of RNN variants for recognizing daily life stress levels by using multimodal physiological signals," in *Proc. Int. Conf. MULTIMODAL Interact.*, New York, NY, USA, Oct. 2023, pp. 481–487.
[25] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. Keong Kwoh, X. Li, and C. Guan, "Time-series representation learning via temporal and contextual contrasting," 2021, *arXiv:2106.14112*.
[26] Y. S. Can, M. Benouis, B. Mahesh, and E. André, "Application of multimodal self-supervised architectures for daily life affect recognition," *IEEE Trans. Affect. Comput.*, vol. 16, no. 3, pp. 2454–2465, Jul. 2025.
[27] J. Y. Cheng, H. Goh, K. Dogrusoz, O. Tuzel, and E. Azemi, "Subject-aware contrastive learning for biosignals," 2020, *arXiv:2007.04871*.

YEKTA SAID CAN received the B.Sc., M.Sc., and Ph.D. degrees from Bogazici University, Türkiye, in 2012, 2014, and 2020, respectively. He was a Teaching Assistant with Bogazici University for six years during his Ph.D. After obtaining his Ph.D. degree, he was a Postdoctoral Researcher in an European Union's Horizon 2020 ERC Project (UrbanOccupations) for applying computer vision techniques to retrieve information from historical documents for two years. He is currently a Postdoctoral Researcher with Augsburg University, working on recognizing emotions and stress. His research interests include biometrics, document analysis, physiological signal processing, affective and wearable computing, and machine learning.

MOHAMED BENOUIS received the Ph.D. degree in computer science from Ahmed Benbella University (Oran 1), in 2017. From 2013 to 2022, he was an Assistant Professor with M'sila University. From 2023 to 2025, he was a Postdoctoral Researcher with the Chair for Human-Centered Artificial Intelligence, University of Augsburg, Germany. His research interests include privacy-preserving decentralized learning, affective computing, biometrics, and biomedical data-driven modeling.

ELISABETH ANDRÉ (Senior Member, IEEE) is currently a Full Professor of computer science and the Founding Chair of the Human-Centered Artificial Intelligence, Augsburg University, Germany. She has a long track record in multimodal human–machine interaction, embodied conversational agents, social robotics, affective computing, and social signal processing. Her work has won many awards, including the Gottfried Wilhelm Leibniz Prize, the most important research funding award in Germany. She is a member of the prestigious Academy of Europe, the German Academy of Sciences Leopoldina, and the CHI Academy. In 2013, she was awarded a EurAI Fellowship (European Association for Artificial Intelligence). In 2019, she was named one of the ten most influential figures in the history of AI in Germany by the National Society for Informatics (GI). From 2019 to 2022, she was the Editor-in-Chief of IEEE TRANSACTIONS ON AFFECTIVE COMPUTING.