Application of Multimodal Self-Supervised Architectures for Daily Life Affect Recognition

Yekta Said Can, Mohamed Benouis, Bhargavi Mahesh, and Elisabeth André

***Abstract*—The recognition of affects (an umbrella term including but not limited to emotions, mood, and stress) in daily life is crucial for maintaining mental well-being and preventing long-term health issues. Wearable devices, such as smart bands, can collect physiological data including heart rate variability, electrodermal activity, skin temperature, and acceleration facilitating daily life affect monitoring via machine learning models. However, accurately labeling this data for model evaluation is challenging in affective computing research, as individuals often provide subjective, inaccurate, or incomplete labels in their daily lives. This study introduces the adaptation of self-supervised learning architectures for multimodal daily life stress and emotion recognition tasks, focusing on self-representation and contrastive learning methods. By leveraging unlabeled multimodal physiological signals, we aim to alleviate the need for extensive labeled data and enhance model generalizability. Our research demonstrates that self-supervised learning can effectively learn meaningful representations from physiological data without explicit labels, offering a promising approach for developing robust affect recognition systems that can operate in dynamic and uncontrolled environments. This work represents a significant improvement in recognizing affects in the wild, with potential implications for personalized mental health support and timely interventions.**

***Index Terms*—Wearable computing, self supervised learning, daily emotion recognition, deep learning, Transformer, CNN, physiological signals.**

I. Introduction

RECOGNIZING stress and negative emotions early is vital for maintaining mental well-being and preventing potential health complications. In recent years, there has been a surge in interest in leveraging wearable devices to detect affective states such as stress and emotion in real-time. These devices can gather heart activity, electrodermal activity (EDA), skin temperature (ST), and acceleration (ACC) signals, which are frequently used physiological measurements in research. [1]. Wearable-based affect recognition research commences in the laboratory environment [1], [2], [3], [1]. Recognizing affect levels in this setting is relatively easier due to the ability to monitor subjects, get time stamps and contextual information. The use of gold-standard devices and the restricted movement of subjects contribute to improved data quality. The research direction then shifts to controlled real-life environments like offices, cars, and classrooms, where cameras and sensors can still monitor and control conditions and movements are still limited. Researchers observe that emotions and stress in artificial settings differ from the ones encountered in daily life, which holds greater significance to individuals [4]. Additionally, it is found that subjects are reluctant to wear obtrusive golden standard equipment for stress measurement due to discomfort. Consequently, affect recognition research moves beyond the laboratory and controlled environments, aiming to create an unobtrusive stress recognition system for everyday use, offering the potential for timely interventions and personalized support.

However, recognizing affect levels in daily life remains a challenge. The effectiveness of such systems is often hindered by problems such as relying on subjective self-reports as the golden standard, unrestricted movements, low data quality of unobtrusive wrist-worn devices which are suitable for daily life usage, and limited battery life [5]. The most prominent issues are subjective, noisy, and missing labels. Traditionally, affect recognition models heavily rely on labeled data, which poses significant challenges, especially in real-world scenarios. Subjective self-reports, the primary method of labeling affective states, can be inherently biased and unreliable. In laboratory environments, the context is known at all times which increases the reliability of labels. Furthermore, there is the possibility of getting experts to observe the behavior of participants and get more reliable labels in laboratory environments [6]. Moreover, daily life affect data often lack crucial context information and are limited by factors such as unrestricted movement, leading to inaccuracies in recognizing affective states. In the wild, the movements are unlimited and subjects can be involved in different levels of physical activities (such as walking, running to catch a bus, and going to the gym) which completely alters the characteristics of physiological signals and creates a significant amount of noise. Therefore, in order to deal with the altered signals and increased amount of noise, more advanced architectures are needed for daily life environments.

Due to the mentioned reasons, daily life affect recognition has a number of significant issues and it is almost a completely different problem than recognizing affects in laboratory or controlled environments. Addressing these challenges requires innovative approaches that can alleviate the burden of labeling, particularly in dynamic and uncontrolled environments. Self-supervised learning architectures, such as contrastive learning and self-representation learning, offer promising solutions by leveraging

Received 18 November 2024; revised 14 March 2025; accepted 16 April 2025. Date of publication 21 April 2025; date of current version 15 September 2025. This work was supported by the German Research Foundation under the project Health-Relevant Effects of Different Urban Forest Structures (LEAF). Recommended for acceptance by A. Liu. *(Corresponding author: Yekta Said Can.)*

The authors are with the Chair for Human-Centered Artificial Intelligence, Universität Augsburg, 86159 Augsburg, Germany (e-mail: yekta.can@uni-a.de).

Digital Object Identifier 10.1109/TAFFC.2025.3562552

TABLE I
EMOTION RECOGNITION STUDIES THAT USE UNOBTRUSIVE SMARTBANDS AND SMARTWATCHES

Study,Year	Signals	ML Type	Classifier	V/A/S	Environment
Akbulut et al. [7](2022)	ECG, GSR, ST, SpO2	Supervised	LSTM, CNN, CNN-RF	5 Emotions	Lab
Sharma et al. [8](2021)	EEG, ECG	Supervised	SVM, kNN	V/A	Lab
Malviya et al. [9] (2023)	PPG, EDA, ST, Acceleration	Supervised	SVM, RF, kNN, LSTM	S	Lab
Soni et al. [10] (2022)	PPG, EDA, ST, Acceleration	Supervised	LSTM, Transformer	S	Lab
Khan et al. [3] (2022)	PPG, EDA, ST, Acceleration	Semi-Supervised	Bi-LSTM	S	Lab
Sarkar et al. [11] (2020)	ECG	Self Supervised	Transformer	S	Lab
Matton et al. [12] (2023)	EDA	Self Supervised	CNN	S	Lab
Wu et al. [1] (2023)	PPG, EDA, ST, Acceleration	Self Supervised	Transformer	S	Lab
Can et al. [13] (2020)	PPG, EDA, ST, Acceleration	Supervised	MLP, RF, SVM, kNN	S	Daily
Shui et al. [14](2023)	PPG, EDA, Acceleration	Supervised	CNN, Attention, SVM	A,V	Daily
Ahmed et al. [15] (2022)	PPG, EDA, Acceleration	Supervised	Logistic Regression (LR)	A,V	Daily
Abdalalim et al. [16] (2024)	ECG, EDA, Acceleration, ST	Supervised	SVM, LR, kNN, RF	S	Daily
Basaran et al. [17](2024)	PPG, EDA, ST, Acceleration	Semi Supervised	Deep Autoencoder	S	Daily
Yu et al. [18] (2022)	ECG, EDA, Acceleration, ST	Semi Supervised	LSTM Autoencoder	S	Daily
Yu et al. [19] (2023)	ECG, EDA, Acceleration, ST	Semi Supervised	LSTM Autoencoder	S	Daily
Proposed System (2024)	ECG, PPG, EDA, Acceleration, ST	Self Supervised	CNN, Transformer	A,S	Daily

Signal abbreviations are PPG: Photoplethysmography, EDA: Electrodermal Activity, ST: Skin Temperature, ECG: Electrocardiogram. Machine learning (ML) classifier abbreviations are as follows: SVM: Support Vector Machine, RF: Random Forest, CNN: Convolutional Neural Networks, LSTM: Long Short Term Memory, NB: Naive Bayes, LR: Logistic Regression. V stands for Valence, A stands for Arousal and S stands for Stress.

unlabeled data to learn meaningful representations directly from the data itself. In contrastive learning, the model learns by comparing similar and dissimilar data points, refining its ability to distinguish important features through these comparisons, like recognizing a friend by seeing their photo next to others. Self-representation learning, however, allows the model to learn directly from the data without explicit comparisons, focusing on finding patterns and structures within the data itself. Both methods are pivotal for creating rich, useful representations, especially in fields where labeled data is limited. By capturing inherent relationships within the data, these approaches have the potential to enhance the robustness against noise and generalizability of affect recognition models, particularly in wild conditions where labeled data is scarce or unreliable.

In this study, we first implement a self-supervised deep learning model by using a public dataset recorded in the laboratory. After that, we optimize and modify selected prominent deep learning architectures for the multimodal daily life data which makes it a self-supervised learning framework that is adjusted for daily life multimodal physiological data. To the best of our knowledge, this study is the first to develop multimodal self-supervised (contrastive and representation learning) stress and emotion recognition architectures specifically suited for daily life multimodal physiological data.

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, selected self-supervised affect recognition architectures are introduced. In Section V, the experimental results of the proposed systems are discussed. In Section VI, we summarize the study, and future work of the current research is presented.

II. RELATED WORK

The initial approach to reducing the labeling burden in affective computing involved semi-supervised learning techniques (see Table I), which leverage a small subset of labeled data to generate pseudo-labels for unlabeled data. In this context, the labeled subset acts as an upstream task, aiding in representation learning that benefits downstream applications such as affect recognition. This approach has shown promising results, particularly in controlled laboratory settings. For instance, Khan et al. [3] applied semi-supervised learning to laboratory datasets, achieving satisfactory performance on downstream tasks. Extending this to "in-the-wild" datasets, researchers developed a sequence-to-sequence LSTM auto-encoder (LSTM-AE) that combines semi-supervised learning with data augmentation and consistency regularization techniques [19]. Applied to the SWEET [20] and TILES [21] datasets, demonstrating moderate success in real-world scenarios. Basaran et al. [17] subsequently improved the downstream task performance by employing graph-based label propagation on a local daily-life dataset.

Following this, self-supervised learning has gained traction across various domains, including medical imaging, natural language processing, and computer vision. In self-supervised learning, the model typically begins with a pretext task to learn useful representations from unlabeled data. This learned knowledge is then transferred to downstream tasks, where it can improve the performance of target applications. Self-supervised methods in affective computing are still limited but include contrastive learning approaches, such as SimCLR [22], which involve pretext tasks aimed at learning separable representations by contrasting positive and negative pairs of data samples. In other works, pseudo-labeling is used to create surrogate labels for auxiliary tasks [1].

The potential of self-supervised pretext tasks was first explored in affective computing by Sarkar and Etemad [11], who used a multi-task CNN with ECG signals to improve representation learning. In this study, they leveraged a pretext task that involved categorizing augmented ECG signals, and the resulting representations were later fine-tuned for the downstream task of stress recognition, achieving performance gains. Cheng et al. [23] developed a self-supervised learning model for biosignal classification that focuses on EEG and ECG data, addressing the upstream challenges of noisy labels and intersubject variability. Their subject-aware approach incorporates subject-specific contrastive loss and adversarial training, promoting subject-invariance in learned representations during the pretext phase.

These embeddings, when applied to the downstream classification task, demonstrated competitive performance comparable to fully supervised methods, highlighting the effectiveness of subject-invariance in enhancing representation quality.

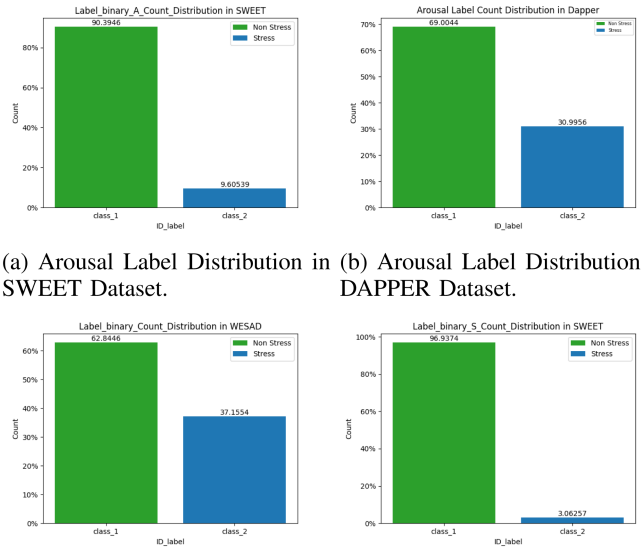
Most studies utilizing self-supervised learning in affective computing concentrate on single modalities, varying the backbone architecture or modality to optimize pretext task performance [24], [25], [26], [27]. Matton et al. [12] introduced a unique approach by designing EDA-specific augmentations for contrastive learning, enabling improved representation learning for stress classification in EDA data. Despite these advances, uni-modal methods often struggle with accuracy, prompting researchers to explore multimodal transformers. Wu et al. [1] employed transformers for multimodal self-supervised learning, defining a pretext task that captures inter-modality relationships, with the aim of generating discriminative representations for downstream classification tasks. Most existing studies apply their approaches to datasets like WESAD, VERBIO, SWELL, and DREAMER, typically collected in laboratory or controlled environments. However, in-the-wild settings pose additional challenges due to increased data variability. To our knowledge, this study is the first to tailor self-supervised learning and contrastive methods to in-the-wild multimodal affect recognition tasks.

III. Description of Datasets and Preprocessing

For the fine-tuning of stress detection models, the selection of appropriate datasets is essential. We incorporate a laboratory dataset WESAD for this purpose and all available public "in-the-wild" physiological signal datasets namely DAPPER, SWEET, and LabToDaily to develop daily life stress and emotion recognition models. The laboratory dataset provides a structured environment where stress levels and emotional states are precisely labeled, enabling model fine-tuning with well-defined physiological patterns. In contrast, the "in-the-wild" datasets, collected in natural, everyday contexts, offer broader variability, capturing stress and emotional responses influenced by diverse environmental and situational factors.

A. WESAD

The Wearable Stress and Affect Detection (WESAD) dataset, retrieved from 15 participants in a laboratory setting, encompassed conditions such as amusement, stress, meditation, and recovery. Self-reports includes 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 [28], 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 versus non-stress (see Fig. 1 for label distribution).



(a) Arousal Label Distribution in SWEET Dataset.
(b) Arousal Label Distribution in DAPPER Dataset.
(c) Stress Label Distribution in WESAD Dataset.
(d) Stress Label Distribution in SWEET Dataset.

Fig. 1. Label distributions in the used datasets.

B. DAPPER

The DAPPER dataset, recorded in an ambulatory environment, differed from the aforementioned datasets by being collected outside laboratory settings. 142 participants contributed psychological recordings, with only 88 providing physiological recordings over five days [29]. Emotions were annotated using the experience sampling method (ESM), capturing detailed descriptions of everyday emotional experiences through the day reconstruction method. ESM included arousal and valance ratings, and PANAS questions for ten selected emotions. Physiological data included blood volume pulse (PPG), EDA, and acceleration data. To ensure data quality, incomplete or discontinuous segments were excluded, resulting in 1801 30-minute data segments from 86 participants. Labels were not utilized, as the dataset served solely for self-supervised upstream tasks.

C. SWEET

Imec's SWEET (Stress in the Work Environment) dataset [20] 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.

D. LabToDaily

This dataset is recorded from 14 university students aged between 20 and 25 for one week, twelve hours a day during their daily routine [13]. 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.

E. Preprocessing

Following established preprocessing methods [28], 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 [1], 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 [1] and using raw data directly with them provides better performance when compared to handcrafted features [30].

IV. Methodology

We selected the self-supervised architectures that prove successful performance in various tasks by using physiological signals. We selected two contrastive learning-based CNN architectures [12] and [23], created and optimized one contrastive learning LSTM architecture, and one transformer-based [1] self-representation learning architecture. We adapted these architectures to multimodal physiological signal data, used fine-tuned hyperparameters for multi-modal affect recognition tasks, and improved them. We described their original versions and adaptation techniques in this section. We also mentioned our validation and evaluation strategies.

A. Self Supervised Contrastive Learning Architectures

1) Applied Transformation Types: In contrastive learning, transformations are applied to data to create diverse views or augmentations of the same input. These transformations alter the input data in various ways, such as introducing noise, negating values, scaling, permuting the order of data points, or time shifting (TS) the signal. Tuples from these transformations are used in contrastive learning to pair different augmentations or transformations of the same input data. By creating tuples of augmented views, the model is trained to minimize the difference between positive pairs (augmented views of the same data) and maximize the difference between negative pairs (augmented views of different data). This encourages the model to learn representations that capture the underlying structure or semantics of the data, as it must recognize similarities and differences between various transformations of the same input. The applied transformations (augmentations) can be briefly explained as:

a) Noising: Adding random noise to the input signal. This helps the model learn to be robust against noise in the data and generalize better to unseen variations.

b) Negating: Inverting or changing the sign of values in the input signal. This transformation helps the model learn invariant representations that are not affected by changes in polarity.

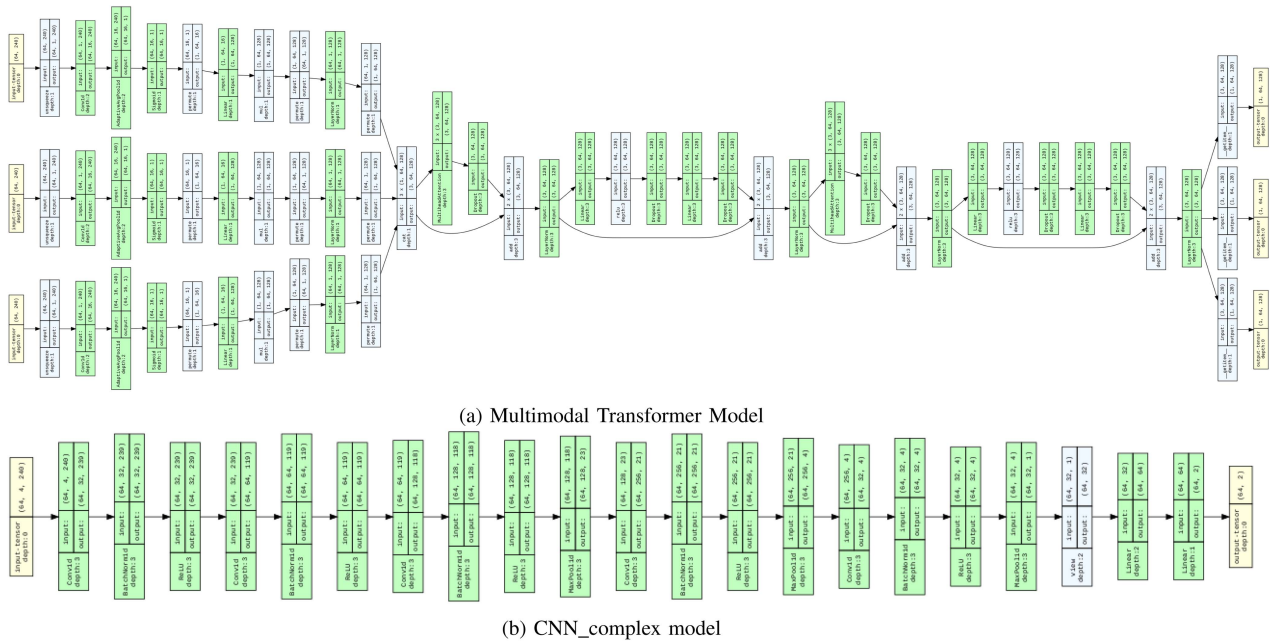
c) Scaling: Adjusting the magnitude or scale of values in the input signal. This helps the model learn representations that are invariant to changes in magnitude, making it more generalizable.

d) Permuting: Shuffling the order of data points within the input signal. This transformation helps the model learn to be invariant to the temporal order of data points, enhancing its robustness to variations in sequence length.

e) Time Shifting (TS): Temporally shifting the signal along the time axis. This transformation helps the model learn to be invariant to temporal shifts in the data, making it more robust to changes in timing or alignment.

f) TS-TCC (Temporal and Contextual Contrasting): This is a specific contrastive learning method mentioned in the study [31]. It involves creating contrasting views of time-series data by considering both temporal dynamics and contextual information. This method aims to learn robust representations by contrasting different aspects of the data, such as temporal features and contextual contexts.

2) Basic CNN Architecture. a) Overview of the Original Architecture: We first applied a basic CNN architecture for the upstream task [12]. A convolutional Neural Network (CNN) encoder is configured to process input data with a dimensionality of 240 (see Fig. 3). It employs a 7×7 convolutional kernel with a stride of 1, producing 64-dimensional feature vectors as output. Notably, dropout, a regularization technique, is disabled (dropout_prob: 0), implying that during training, no units are randomly dropped. Moreover, the weights of the network are not frozen (freeze=False), allowing them to be updated during the training process. Overall, this code segment sets up a CNN encoder network with specific architectural parameters suitable



(a) Multimodal Transformer Model

(b) CNN_complex model

Fig. 2. Architecture diagrams for multimodal transformer and CNN_complex architectures.

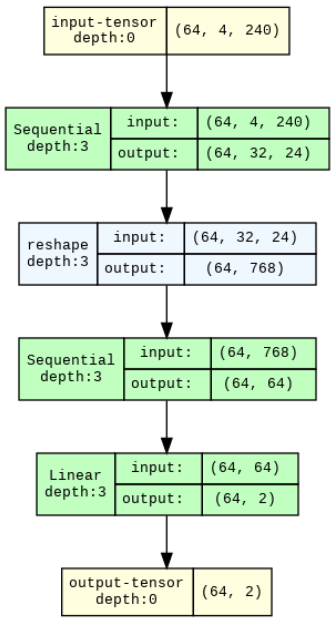


Fig. 3. Basic CNN architecture.

for processing input data of a given dimensionality and generating feature representations of a defined size.

b) Adaptation to the Multimodal Affect Recognition Data: First of all the architecture uses EDA-specific transformation based on phasic and tonic signal components. However, since we are using multimodal data which has three other modalities in addition to the EDA signal, we tested time series related well known generic signal processing transformations and chose the best-performing tuple. Second, the architecture is developed for a single modality and we changed the input layers by concatenating different modalities into one input and providing this concatenated multimodal data as input. Lastly, since an important parameter of contrastive learning, namely temperature, is not reported in the original study, we need to fine-tune this hyperparameter and use the best-performing temperature value.

3) Complex CNN Architecture. a) Overview of the Original Architecture: We also applied a more complex CNN architecture [23] which is applied to a laboratory-recorded MIT-BIH ECG dataset [32] for recognizing heart arrhythmia. The CNN encoder architecture begins with a convolutional layer that takes input from 2 channels and generates 32 feature maps using a 13 $\times$ 13 kernel (see Fig. 2(b)). This is followed by a residual block, maintaining 32 feature maps and employing a convolutional layer with an 11 $\times$ 11 kernel. Subsequently, a max-pooling layer with a 4 $\times$ 4 pool size downsamples the feature maps. Another residual block ensues, expanding the feature maps to 64 and utilizing a 9 $\times$ 9 kernel convolutional layer, followed by another max-pooling operation. The architecture then incorporates a residual block that increases the feature maps to 128 via a convolutional layer with a 7 $\times$ 7 kernel, succeeded by a max-pooling layer with a 2 $\times$ 2 pool size. Finally, a residual block reduces the feature maps back to 64 with a 7 $\times$ 7 kernel convolutional layer. An Exponential Linear Unit (ELU) activation function is applied before flattening the output, preparing it for further processing. This sequential arrangement forms a robust CNN encoder adept at extracting hierarchical representations from input data.

b) Adaptation to the Multimodal Affect Recognition Data: First, this architecture is also developed for a single modality, and we increased the number of input layers to four. After this change, we also optimized the other layers so that we had the same size output representation layer 64. The updated layers can be described as follows: Furthermore, we test time series related

TABLE II
HYPERPARAMETER OPTIMIZATION RESULTS ON THE WESAD DATASET, SHOWING THE IMPACT OF DIFFERENT TRANSFORMATION PAIRS AND TEMPERATURE VALUES ON F1 SCORE

Temperature	Transformation Tuples												
Value	NoisePermute	NoiseTS	NoiseNegate	NoiseScale	NegateTS	ScaleTS	PermuteTS	NegateScale	NegatePermute	ScalePermute	TSTCC	MaskNoise	MaskPermute
0.1	91.1	89.8	89.9	89.9	89.7	89.7	89.7	90.9	90.9	89.8	90.9	90.3	90.4
0.2	91.8	90.4	91.0	91.0	90.2	90.2	90.2	90.6	90.6	90.3	90.9	91.4	90.6
0.4	94.5	91.2	92.5	93.6	91.9	91.9	91.9	91.9	91.9	91.1	91.5	91.5	91.5
0.8	91.8	90.7	91.9	91.9	90.2	90.2	90.2	91.7	91.7	91.8	90.9	91.7	91.6

TABLE III
LSTM HYPERPARAMETER OPTIMIZATION ON SWEET DATASET

# of Layers	128 Neurons	256 Neurons	512 Neurons	1024 Neurons
1	94.67	94.57	94.87	95.35
2	94.65	95.06	95.13	**95.76**
3	94.47	94.72	94.97	95.51
4	94.55	94.71	95.01	95.64

Layers from 1 to 4 and neurons from 128 to 1024 are tested. F1 scores are presented.

to well-known generic signal processing transformations and choose the best-performing transformation tuple for contrastive learning, and we also fine-tune the temperature value and use the best-performing temperature value.

4) LSTM Architecture. a) Overview of the Original Architecture: An LSTM model was also implemented. The model is built around a long short-term memory (LSTM) network, which processes sequential input features and extracts meaningful representations. The final hidden state from the LSTM is passed through a fully connected layer to generate the output, which is then normalized using L2 normalization. This ensures that embeddings are of unit length, an essential property for contrastive learning. The training framework relies on contrastive loss, which is key for learning robust self-supervised feature representations. Instead of relying on manually labeled emotional categories, the model learns by comparing pairs of augmented physiological signals and optimizing their similarity. A temperature parameter of 0.4 was set to control the sensitivity of the similarity function, ensuring effective distance scaling between samples.

b) Hyperparameter Optimization: To enhance model performance, hyperparameter optimization was conducted, varying the number of LSTM layers from 1 to 4 to explore the impact of model depth on feature extraction (see Table III). The hidden dimension was adjusted between 128 and 1024 neurons to balance expressiveness and computational efficiency. The optimization process also tested different learning rates and batch sizes, ultimately setting batch size to 64 for stable training. The model was trained using AdamW optimization, which provides adaptive learning rates and improved weight decay, and a learning rate scheduler (ExponentialLR) was used to dynamically reduce the learning rate over time, ensuring better convergence.

B. Self-Representation Based Self Supervised Multimodal Transformer Architecture

1) Applied Self Representations: The pretext task of signal transformation recognition is employed in self-representation-based SSL. A number of signal transformations are defined, and the system tries to recognize these transformations in a classification task. Signal transformation recognition has been proven effective in learning generalized representations for downstream tasks like action and emotion recognition. The transformations used in SSL can be categorized into magnitude domain transformations and time domain transformations. These transformations are applied to all modalities of the signal data, and the resulting transformed data is input into the SSL model alongside the original data. By recognizing the types of signal transformations, the model learns to extract robust and generalized representations against disturbances in the magnitude or time domains.

Magnitude domain transformations include Gaussian noise addition and Magnitude-warping. Gaussian noise addition disturbs the original signal with white Gaussian noise, simulating real-world noise scenarios. Magnitude-warping alters the magnitudes of the signal by applying a random smooth curve, which can mimic measurement errors or signal artifacts.

Time domain transformations consist of Permutation, Time-warping, and Cropping. Permutation disrupts the temporal order of segments in the signal, prompting the model to capture time-domain dependencies. Time-warping stretches or squeezes segments of the signal to simulate duration variations in emotional responses. Cropping randomly selects and resamples segments of the signal to enhance robustness to temporal changes in emotional events.

2) Multimodal Transformer Architecture: The multimodal transformer architecture described in the paper is a novel framework designed for wearable emotion recognition, leveraging peripheral physiological signals. The architecture consists of two main components: modality-specific encoders and a shared transformer-based encoder. The modality-specific encoders use temporal convolutional networks to process individual modalities such as blood volume pulse (BVP), electrodermal activity (EDA), and temperature (TEMP), generating low-level features for each. These features are then fed into the shared transformer-based encoder, which integrates multimodal information through cross-modal attention and self-attention mechanisms.

The shared encoder employs a transformer block with four-head attention, a feedforward layer dimension of 128, ReLU activation, and a Dropout rate of 0.2. It processes the stacked multimodal embeddings without positional encoding, as the features from each modality are generated by different encoders. The output from the shared encoder is then passed through modality-specific classification heads, which include 1D global average pooling, fully connected layers, batch normalization, and ReLU activation, followed by a final fully connected layer for emotion classification. The architecture is trained in a self-supervised manner using a pretext task of signal transformation recognition, which automatically labels a large amount

of unlabeled data. This pre-training allows the model to learn generalized multimodal representations that are later fine-tuned on supervised emotion recognition tasks.

3) Adaptation to the Multimodal Affect Recognition Data: For the feature extraction component, this study used temporal convolutional networks. We changed this part and adopted the inception time architecture for the feature extractor encoder because it provides a better representation along a multi-series dimension and yields better results with time series [33] (see Fig. 2(a)). Typically, this encoder consists of five inception modules, Global Average Pooling (GAP), and two dense layers with ReLU activation to map into a sequence of segment embeddings with dimensions n x d. For more details, refer to the paper by Ismail et al. [33].

C. Validation Strategy

For the evaluation of the machine learning algorithms, we selected realistic and challenging techniques. We separated training, validation, and test sets by considering the participants. We used a 60%-20%-20% split by dividing the data by the number of subjects. We guaranteed that all different sets consist of data obtained from different subjects, and there is no overlap between training, validation, and test sets. If there are 80 subjects, randomly selected 48 subjects (%60) were assigned to the training set, randomly selected another 16 subjects were used for validation purposes, and the remaining 16 were used for testing. We rounded the numbers for the float number of subjects, and the percentages can vary slightly in these cases since we can not use the float number of (e.g. 48,31) subjects.

The second issue is the class imbalance in all of the datasets. To get more meaningful results, we opted for the F1 score metric instead of the accuracy metric. In this way, we achieved a fairer comparison between algorithms and state-of-the-art.

V. Experimental Results and Discussion

In this section, we first fine-tune the critical hyperparameters for contrastive learning in a stress recognition task from a public multimodal physiological dataset. We use these hyperparameters in the contrastive learning architectures. We apply the selected and adapted architectures to available daily life affect recognition datasets and report the results.

A. Hyperparameter Optimization

Temperature plays a significant role in controlling the behavior of contrastive learning [34]. It determines the level of penalty on hard negative samples. It has a significant impact on the classification performance [34]. Furthermore, the used transformation type has a prominent effect on affect recognition performance [12]. We fixed the different CNN and Transformer architectures since they are applied to similar affect recognition tasks and achieved robust results in controlled environments. However, since the best temperature values and best transformation pairs changed from paper to paper, and especially since temperature values are not reported in most of the papers, we tested all different values in hyperparameter optimization, found the best setting, and used it throughout the paper. As can be seen in Table II, the best transformation tuple is noise-permute, and the best temperature value is 0.4. These values are fixed for the remaining experiments.

TABLE IV
The Performance of Single Modalities for Recognizing Affects

Dataset name	EDA	TEMP	PPG	Label
DAPPER	78.00	N/A	75.00	Arousal
LD	75.70	71.00	76.20	Stress
SWEET	90.20	90.00	89.80	Arousal

The F1 score is used to demonstrate the performance.

TABLE V
The Performance of the Selected Multimodal Architectures for Recognizing Affects

Dataset name	CNN_basic	CNN_complex	MM_Transformer	Label
DAPPER	71.60	75.30	72.36	Arousal
LD	75.50	75.00	75.69	Stress
SWEET	97.80	96.00	96.56	Stress
SWEET	89.50	90.00	91.05	Arousal

The F1 score is used to demonstrate the performance.

B. Contribution of Modalities to Affect Recognition Performance

Our main focus is to apply a multimodal self-supervised system for daily life datasets. However, it is also important to demonstrate the effect of single modalities and their contribution when used alone. We skipped the acceleration modality because it mostly reflects physical activity levels, and using it alone for affect recognition can be misleading. The CNN_complex architecture [23] is used for the comparison. The best modality changes with the dataset (see Table IV). EDA achieves the best results with DAPPER and SWEET. On the other hand, PPG achieves slightly higher performance when compared to EDA.

C. Multimodal SSL Architectures for Daily Life Affect Recognition

We tested the selected three self-supervised architectures for the three in-the-wild datasets (see Table V). For emotion recognition, we used only the Arousal label since it is more clearly reflected in the physiological signals. For the SWEET dataset, we used both the Arousal and Stress labels. Both CNN architectures achieve similar accuracies for all the tasks. Although a multimodal transformer is a more complex architecture, it could not outperform the best CNN-based architecture in most cases. The nature of the datasets used can significantly impact model performance. CNNs, designed for local feature extraction, may excel when data exhibits strong local dependencies. In contrast, Transformers are adept at capturing global relationships, which may not always be necessary for tasks that do not require modeling long-range dependencies. Furthermore, Transformers typically require larger datasets to effectively learn patterns due to their complexity and lack of inherent inductive biases. In situations with limited data, CNNs might perform better as they can generalize more effectively from smaller datasets. Therefore, in our case, Transformers might require larger datasets to further

improve their performance. We also tested LSTM architecture for the SWEET dataset, and it achieved similar results (95.76 F1 score) with transformer and CNN_complex architectures.

The SWEET dataset has higher accuracies than the other two datasets. One key factor contributing to the superior performance observed in the SWEET dataset is the difference in sampling rates and decision intervals. In our study, we standardized the sampling rates across datasets by downsampling to the lowest available frequency. For the SWEET dataset, this was derived from Electrocardiography (ECG) data at 1 sample per minute, which led to all other signals being downsampled accordingly. Given our window size of 240 samples, this resulted in a decision interval of 240 minutes. In contrast, for other datasets, the lowest frequency was 4 samples per second, leading to a decision interval of only 1 minute. This longer interval in SWEET allows for more stable and averaged physiological measurements, potentially enhancing model accuracy by reducing the impact of transient fluctuations. This aligns with findings in signal processing literature, where appropriate sampling rates and window sizes are crucial for improving signal representation while minimizing noise. Additionally, dataset-specific differences in data distribution and sensor modalities likely contributed to performance variations. Each dataset was collected under different conditions, with varying participant demographics and recording environments, which may have influenced generalization. Notably, the SWEET dataset employs ECG for heart activity monitoring, whereas others use Photoplethysmography (PPG). ECG provides a direct measure of the heart's electrical activity, offering higher precision, while PPG is more susceptible to motion artifacts, potentially affecting classification performance. These factors collectively suggest that the SWEET dataset's superior results may stem from its longer decision intervals, improved signal stability, and the use of ECG for heart rate measurement.

D. Visualization of Emotion Separability Across Datasets

To further investigate how physiological signals encode emotional states, we analyzed the t-SNE projections of feature representations extracted from our model across different datasets. Fig. 4 illustrates the differences in emotion separability between the controlled WESAD dataset and three daily-life datasets: DAPPER, SWEET, and LabToDaily. In WESAD, where participants experience controlled emotional stimuli, the clusters of different emotional states are clearly distinguishable. This indicates that under structured experimental conditions, physiological signals provide strong and distinct affective representations, supporting the effectiveness of our model in detecting emotions. However, as we transition to daily-life datasets, the separability of emotions is notably reduced. In DAPPER, SWEET, and LabToDaily, we observe increased cluster overlap, reflecting the challenges introduced by real-world noise, motion artifacts, and self-reported labels. Despite these challenges, the partial clustering observed in the real-world datasets suggests that physiological signals still retain emotional information, even in dynamic, uncontrolled environments. This supports the viability of our model in wearable-based emotion recognition tasks.

TABLE VI
CROSS-DATASET EVALUATION RESULTS USING THE SELF-SUPERVISED CNN_COMPLEX MODEL TRAINED ON DAPPER

Training Dataset	Testing Dataset	F1 Score
DAPPER	LabToDaily (LD)	76.11
DAPPER	SWEET	90.06

E. Investigating Generalizability Through Cross-Dataset Evaluations

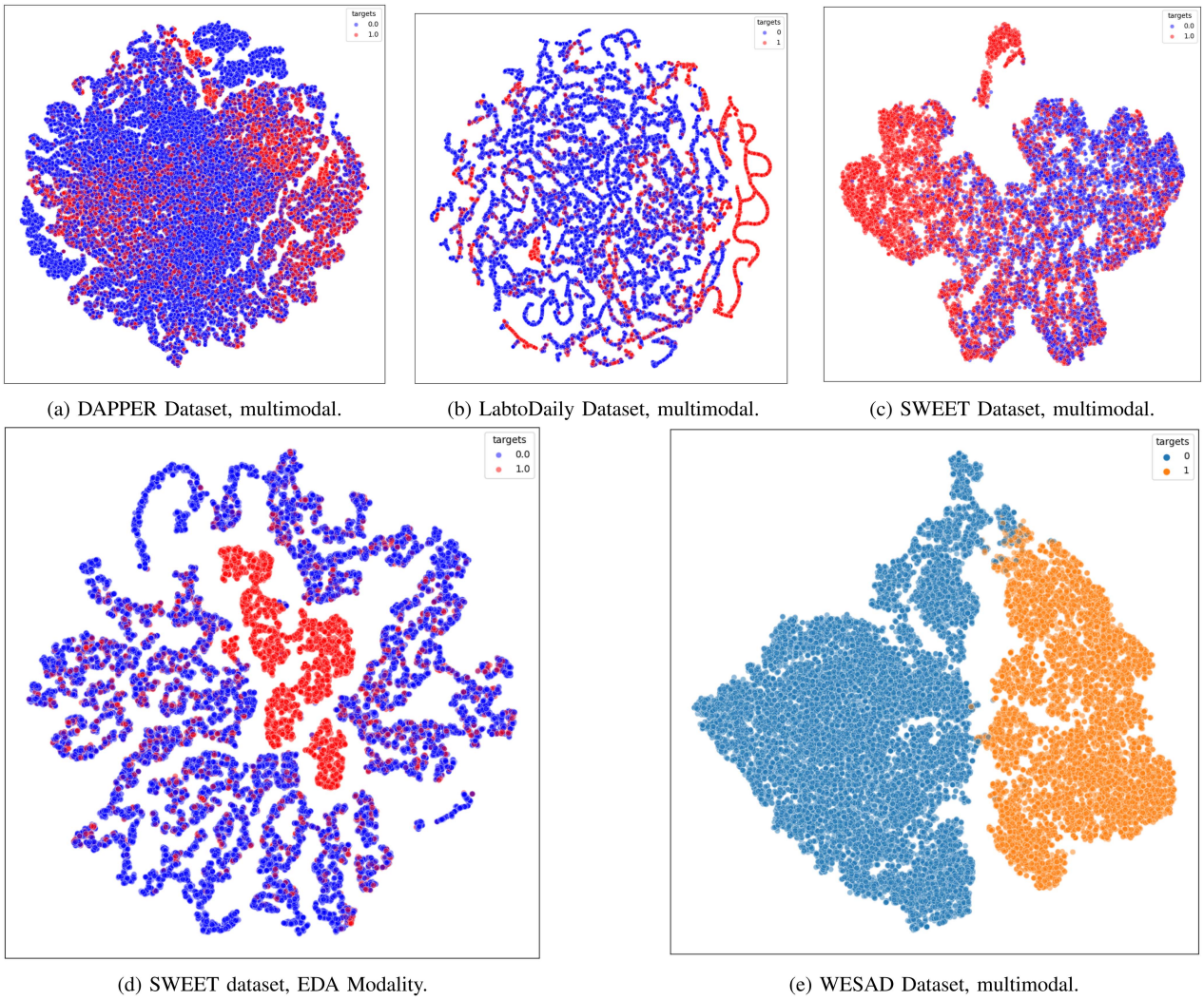
To assess the generalization potential of our self-supervised learning approach, we performed cross-dataset fine-tuning experiments. Rather than training models independently on each dataset, we first pre-trained our model using DAPPER, the most diverse and largest dataset in our study, and then fine-tuned it on SWEET and LabToDaily (LD). This approach allows us to evaluate whether features learned in a self-supervised manner on DAPPER can be effectively transferred to new datasets.

The results presented in Table VI show that the DAPPER-trained self-supervised model, when fine-tuned on SWEET and LD, achieves comparable accuracy to models trained directly on these datasets. This suggests that our learned representations capture fundamental physiological patterns related to emotion, enabling effective adaptation to unseen conditions. The ability to generalize well across datasets is particularly important for real-world applications, where labeled physiological datasets are often scarce, noisy, and collected in varying conditions.

While these findings confirm the transferability of our self-supervised learning approach, further research is required to systematically evaluate its behavior across more datasets, population groups, and recording conditions. In future work, we will extend this investigation, incorporating domain adaptation strategies and exploring methods to further improve cross-dataset consistency in affect recognition models.

F. Comparison With the State of the Art

As mentioned, the proposed system is the first self-supervised in-the-wild affect recognition study. In this section, we compared our performances with other studies that applied supervised architectures to these datasets. In this way, we will see the contribution of self-supervised techniques more clearly. It is important to note that even achieving similar and comparable performance with supervised architectures shows the superiority of self-supervised architectures since it relieves the burden of intensive labeling. We will compare and discuss the results for each dataset (see Table VII). For the DAPPER dataset, the studies applied supervised architectures and achieved a maximum of 71% accuracy for binary arousal detection. Our self-supervised approach outperforms these performances without using labeled data. Both supervised and semi-supervised techniques were tested with the SWEET dataset. Our system outperformed the semi-supervised systems, and our results are slightly below supervised results for this dataset. However, the decrease from 98.29% to 97.80% seems acceptable if we take into account the labeling burden. For the LabtoDaily dataset, our system achieves better results than traditional supervised



(a) DAPPER Dataset, multimodal. (b) LabtoDaily Dataset, multimodal. (c) SWEET Dataset, multimodal.

(d) SWEET dataset, EDA Modality. (e) WESAD Dataset, multimodal.

Fig. 4. t-SNE visualizations of physiological feature representations across different datasets. The first four plots show daily-life datasets: DAPPER, SWEET, and LabToDaily (LD), while the last plot represents the WESAD dataset (lab-controlled). Red and Orange points are stress labels, whereas blue ones are non-stress.

TABLE VII
AFFECT RECOGNITION STUDIES THAT USE DAILY LIFE DATASETS

Study,Year	Dataset	Type	Classifier	Accuracy (%)
[17](2024)	LabToDaily	Semi Supervised	Deep Autoencoder	77%
[13] (2020)	LabToDaily	Supervised	MLP, RF, SVM, kNN	71.40%
[14](2023)	DAPPER	Supervised	CNN, Attention, SVM	61.55% (Arousal)
[15] (2022)	DAPPER	Supervised	Logistic Regression (LR)	63.9% (Arousal)
[35] (2025)	DAPPER	Supervised	Transformer	71.5% (Negative, Positive Affect)
[16] (2024)	SWEET	Supervised	SVM, LR, kNN, RF	98.29 (Stress)
Proposed Technique (2024)	LabtoDaily	Self Supervised	CNN, Transformer	75.69%
Proposed Technique (2024)	DAPPER	Self Supervised	CNN, Transformer	75.30%
Proposed Technique (2024)	SWEET	Self Supervised	CNN, Transformer	97.80% (Stress)

The machine learning classifier abbreviations are as follows: SVM: Support Vector Machine, RF: Random Forest, CNN: Convolutional Neural Networks, LSTM: Long Short-Term Memory, AE: Autoencoder, NB: Naive Bayes, LR: Logistic Regression. The results are for 2-class affect recognition.

architectures and comparable results with semi-supervised architectures (75.69 - 77.00). These results show that our system achieves robust results with all three datasets.

VI. CONCLUSION

This study has explored the application of self-supervised learning architectures for the recognition of stress and emotions in daily life scenarios, leveraging multimodal physiological signals captured by wearable devices. By employing self-representation learning and contrastive learning methods, we aimed to mitigate the reliance on extensive labeled data and improve the generalizability of affect recognition models across different datasets and real-world conditions.

The study tested the effectiveness of the self-supervised learning models on several in-the-wild datasets, including DAPPER,

SWEET, and LabToDaily. The models demonstrated robust performance, with the best F1 scores achieved being 97.80% for stress recognition on the SWEET dataset, 75.30% for arousal detection on the DAPPER dataset, and 75.69% for stress recognition on the LabToDaily dataset. These results indicate that the self-supervised learning approaches can achieve comparable or even superior performance to traditional supervised methods that rely heavily on meticulous data labeling.

The findings suggest that self-supervised learning can effectively learn meaningful representations from physiological data, offering a promising avenue for the development of unobtrusive and robust affect recognition systems operating in dynamic and uncontrolled environments. The system's ability to integrate into daily life routines holds the potential for timely interventions and personalized support for mental well-being.

Limitations and Future Work: Despite the promising results, the study acknowledges its limitations. In the wild affect recognition has two important issues different from laboratory studies. The first one is the reliance on subjective and often missing self-reports as the sole ground truth, and the second one is the unlimited movements and artifacts and noise caused by them. While our method effectively mitigates labeling issues, addressing motion artifacts remains an ongoing challenge. Wearable emotion recognition systems rely on physiological signals (e.g., heart activity, electrodermal activity) that are often distorted by motion artifacts. In ambulatory settings, even moderate movements can introduce noise in sensor readings that are unrelated to genuine emotional states. For example, electrodermal activity (EDA) signals can be severely affected by motion; high-frequency spikes from movements may mimic or obscure true skin conductance responses to the point that entire data segments become unusable [36]. In one study, weeks of EDA recordings had to be discarded due to motion-induced signal degradation [36]. Similarly, low-cost photoplethysmography (PPG) sensors in smartwatches and smartbands are "easily affected by motion artifacts" [37]. Physical activity can be associated with heart-activity-based emotion metrics – a raised heart rate from running or hand motion may be falsely interpreted as emotional arousal. Indeed, researchers note that only data from static or mildly active states can be reliably used for emotion monitoring with wrist PPG, whereas data during intense movement (e.g., running) are often excluded due to artifact contamination [38]. Even ECG-based emotion recognition suffers in real-life use; motion artifacts can "lead to the decline of the distinguish[ing] ability of ECG features", making emotion classification far less accurate [39]. In summary, acceleration (motion) data interfere by introducing noise, degrading the quality of emotion-related signals, and potentially triggering false emotion inferences if not properly handled, underscoring the need for artifact mitigation in wearable emotion recognition.

Mitigation Strategies for Motion Artifacts: A common approach is to filter out frequency bands associated with motion noise. For instance, band-pass filtering PPG signals (e.g., 0.5–8 Hz) removes high-frequency jitters and baseline drift, helping suppress motion-induced disturbances [37]. Beyond static filters, adaptive algorithms are used to cancel motion noise in real time. Techniques like Kalman smoothing or adaptive filtering (e.g., LMS filters) leverage reference noise signals to subtract motion artifacts from physiological data [40]. Many wearable devices also embed noise-reduction algorithms and recommend tight sensor contact to minimize movement artifacts at the source [38]. These filtering methods significantly improve signal quality by attenuating the artifact components while preserving genuine emotion-related patterns. Combining motion sensors with physiological sensors enables smarter artifact handling. Accelerometers can detect when a user is in motion and either flag those data segments or help correct them. One strategy is activity gating – using accelerometer data to identify high-movement periods and exclude or down-weight those physiological readings in emotion analysis [38]. Another fusion approach is to use the accelerometer as a reference input for artifact removal algorithms. Adaptive noise cancelation filters often take accelerometer readings as the reference noise signal, dynamically filtering motion artifacts out of the primary biosignal [41].

Data-driven techniques have been developed to automatically detect and compensate for motion artifacts. Instead of simple threshold rules, machine learning models can learn the subtle differences between genuine emotional signal patterns and artifact noise. For example, Hossain et al. trained a classifier to distinguish clean versus motion-corrupted EDA segments, achieving about 95% accuracy in detecting artifact-contaminated windows [36]. Such classifiers use statistical and time-frequency features to identify artifacts, outperforming basic heuristics and preserving more valid data. On the modeling side, modern deep learning frameworks incorporate multi-modal sensor inputs to handle motion context as in this study. For instance, Transformer-based models and convolutional neural networks have been used to fuse heart activity, EDA, and accelerometer data, allowing the network to internally learn motion-artifact compensation and focus on emotion-relevant features [35]. By leveraging patterns across sensors and large datasets, these ML approaches can adapt to complex real-world noise.

In future research, we aim to integrate pretrained foundation models for physiological signals to enhance representation learning in emotion recognition tasks. Recent advancements in large-scale self-supervised learning have demonstrated that foundation models trained on extensive wearable biosignal datasets can generalize effectively across various applications. For instance, Abbaspourazad et al. (2024) [42] developed foundation models using large-scale photoplethysmography (PPG) and electrocardiogram (ECG) data collected via wearable consumer devices, showcasing the potential of such models in health monitoring contexts. Building upon these findings, we plan to investigate the application of pretrained physiological signal models in emotion recognition, aiming to improve performance while reducing reliance on large labeled datasets.

ACKNOWLEDGMENT

This work was carried out within the framework of the AI Production Network Augsburg.

REFERENCES

[1] 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, First Quarter 2024.

[2] S. Samyoun, M. M. Islam, T. Iqbal, and J. Stankovic, "M3sense: Affect-agnostic multitask representation learning using multimodal wearable sensors," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 6, no. 2, Jul. 2022, Art. no. 73.

[3] N. Khan and N. Sarkar, "Semi-supervised generative adversarial network for stress detection using partially labeled physiological data," 2022, *arXiv:2206.14976*.

[4] R. W. Picard, "Automating the recognition of stress and emotion: From lab to real-world impact," *IEEE MultiMedia*, vol. 23, no. 3, pp. 3–7, Third Quarter 2016.

[5] A. Liapis, C. Katsanos, D. Sotiropoulos, M. Xenos, and N. Karousos, "Stress recognition in human-computer interaction using physiological and self-reported data: A study of gender differences," in *Proc. 19th Panhellenic Conf. Inform.*, New York, NY, USA, 2015, pp. 323–328.

[6] M. Norden, O. T. Wolf, L. Lehmann, K. Langer, C. Lippert, and H. Drimalla, "Automatic detection of subjective, annotated and physiological stress responses from video data," in *Proc. 10th Int. Conf. Affect. Comput. Intell. Interaction*, 2022, pp. 1–8.

[7] F. P. Akbulut, "Hybrid deep convolutional model-based emotion recognition using multiple physiological signals," *Comput. Methods Biomech. Biomed. Eng.*, vol. 25, pp. 1678–1690, 2022.

[8] L. D. Sharma and A. Bhattacharyya, "A computerized approach for automatic human emotion recognition using sliding mode singular spectrum analysis," *IEEE Sensors J.*, vol. 21, no. 23, pp. 26931–26940, Dec. 2021.

[9] L. Malviya et al., "mental stress level detection using LSTM for wesad dataset," in *Proc. Data Analytics Manage.*, Springer, 2023, pp. 243–250.

[10] J. Soni, N. Prabakar, and H. Upadhyay, "A multi-layered deep learning approach for human stress detection," in *Proc. Int. Conf. Intell. Hum. Comput. Interaction*, Springer, 2022, pp. 7–17.

[11] P. Sarkar and A. Etemad, "Self-supervised ecg representation learning for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1541–1554, Third Quarter 2022.

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

[13] Y. S. Can et al., "How laboratory experiments can be exploited for monitoring stress in the wild: A bridge between laboratory and daily life," *Sensors*, vol. 20, no. 3, 2020, Art. no. 838.

[14] A. Ahmed, J. Ramesh, S. Ganguly, R. Aburukba, A. Sagahyroon, and F. Aloul, "Evaluating multimodal wearable sensors for quantifying affective states and depression with neural networks," *IEEE Sensors J.*, vol. 23, no. 19, pp. 22788–22802, Oct. 2023.

[15] A. Ahmed, J. Ramesh, S. Ganguly, R. Aburukba, A. Sagahyroon, and F. Aloul, "Investigating the feasibility of assessing depression severity and valence-arousal with wearable sensors using discrete wavelet transforms and machine learning," *Information*, vol. 13, no. 9, 2022, Art. no. 406.

[16] M. Abd Al-Alim, R. Mubarak, N. M. Salem, and I. Sadek, "A machine-learning approach for stress detection using wearable sensors in free-living environments," *Comput. Biol. Med.*, vol. 179, 2024, Art. no. 108918.

[17] O. T. Başaran, 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," *Front. Psychol.*, vol. 14, 2024, Art. no. 1293513.

[18] H. Yu and A. Sano, "Semi-supervised learning and data augmentation in wearable-based momentary stress detection in the wild," 2022, *arXiv:2202.12935*.

[19] H. Yu and A. Sano, "Semi-supervised learning for wearable-based momentary stress detection in the wild," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 7, no. 2, Jun. 2023, Art. no. 80.

[20] E. Smets et al., "Large-scale wearable data reveal digital phenotypes for daily-life stress detection," *NPJ Digit. Med.*, vol. 1, no. 1, 2018, Art. no. 67.

[21] K. Mundnich et al., "TILES-2018, a longitudinal physiologic and behavioral data set of hospital workers," *Sci. Data*, vol. 7, no. 1, 2020, Art. no. 354.

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2020, pp. 1597–1607.

[23] J. Y. Cheng, H. Goh, K. Dogrusoz, O. Tuzel, and E. Azemi, "Subject-aware contrastive learning for biosignals," 2020, *arXiv: 2007.04871*.

[24] X. Wang, Y. Ma, J. Cammon, F. Fang, Y. Gao, and Y. Zhang, "Self-supervised EEG emotion recognition models based on CNN," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1952–1962, 2023.

[25] J. Vazquez-Rodriguez, G. Lefebvre, J. Cumin, and J. L. Crowley, "Transformer-based self-supervised learning for emotion recognition," in *Proc. 26th Int. Conf. Pattern Recognit.*, 2022, pp. 2605–2612.

[26] R. Hu, J. Chen, and L. Zhou, "Spatiotemporal self-supervised representation learning from multi-lead ECG signals," *Biomed. Signal Process. Control*, vol. 84, 2023, Art. no. 104772.

[27] R. K. Nath, J. Tervonen, J. Närväinen, K. Pettersson, and J. Mäntyjärvi, "Towards self-supervised learning of ECG signal representation for the classification of acute stress types," in *Proc. Great Lakes Symp. VLSI*, 2023, pp. 85–90.

[28] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. V. Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proc. 20th ACM Int. Conf. Multimodal Interaction*, 2018, pp. 400–408.

[29] X. Shui, M. Zhang, Z. Li, X. Hu, F. Wang, and D. Zhang, "A dataset of daily ambulatory psychological and physiological recording for emotion research," *Sci. Data*, vol. 8, no. 1, 2021, Art. no. 161.

[30] Y. S. Can and E. André, "Performance exploration of rnn variants for recognizing daily life stress levels by using multimodal physiological signals," in *Proc. 25th Int. Conf. Multimodal Interact.*, New York, NY, USA, 2023, pp. 481–487.

[31] E. Eldele et al., "Time-series representation learning via temporal and contextual contrasting," 2021, *arXiv:2106.14112*.

[32] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, May/Jun. 2001.

[33] H. I. Fawaz et al., "Inceptiontime: Finding alexnet for time series classification," *Data Mining Knowl. Discov.*, vol. 34, no. 6, pp. 1936–1962, 2020.

[34] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2495–2504.

[35] F. Li and D. Zhang, "Transformer-driven affective state recognition from wearable physiological data in everyday contexts," *Sensors*, vol. 25, no. 3, 2025, Art. no. 761.

[36] M. -B. Hossain, H. F. Posada-Quintero, Y. Kong, R. McNaboe, and K. H. Chon, "A preliminary study on automatic motion artifacts detection in electrodermal activity data using machine learning," 2021, *arXiv:2107.07650*.

[37] G. Hwang, S. Yoo, and J. Yoo, "Emotion recognition using PPG signals of smartwatch on purpose of threat detection," *Sensors*, vol. 25, no. 1, 2024, Art. no. 18.

[38] L. Shu et al., "Wearable emotion recognition using heart rate data from a smart bracelet," *Sensors*, vol. 20, no. 3, 2020, Art. no. 718.

[39] W. He, Y. Ye, T. Pan, Q. Meng, and Y. Li, "Emotion recognition from ECG signals contaminated by motion artifacts," in *Proc. Int. Conf. Intell. Technol. Embedded Syst.*, 2021, pp. 125–130.

[40] B. Lee et al., "Improved elimination of motion artifacts from a photoplethysmographic signal using a kalman smoother with simultaneous accelerometry," *Physiol. Meas.*, vol. 31, no. 12, 2010, Art. no. 1585.

[41] A. R. Relente and L. G. Sison, "Characterization and adaptive filtering of motion artifacts in pulse oximetry using accelerometers," in *Proc. 2nd Joint 24th Annu. Conf. Annu. Fall Meeting Biomed. Eng. Soc.*, 2002, pp. 1769–1770.

[42] S. Abbaspourazad, O. Elachqar, A. Miller, S. Emrani, U. Nallasamy, and I. Shapiro, "Large-scale training of foundation models for wearable biosignals," in *Proc. 12th Int. Conf. Learn. Representations*, 2024, pp. 1–24.

Yekta Said Can received the BSc MSc and PhD degrees from Bogazici University, Turkey, in 2012, 2014 and 2020, respectively. He also worked as a teaching assistant with Bogazici University for six years during his PhD After obtaining his PhD degree, he worked as a postdoctoral researcher with 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 working on recognizing emotions and stress in Augsburg University as a postdoctoral researcher. His research interests include biometrics, document analysis, physiological signal processing, affective and wearable computing and machine learning.

Mohamed Benouis received the PhD degree in computer science from the University of Ahmed Benbella, Oran 1, in 2017. He also worked as a teaching assistant with M'sila University for eight years during 2013-2022. He is currently working on privacy in affective computing and computer vision in Augsburg University as a postdoctoral researcher. His research interests include biometrics, face analysis, physiological signal processing, and biomedical-based data-driven modeling.

Bhargavi Mahesh received the MSc degree in autonomous systems from the Bonn-Rhein-Sieg University of Applied Sciences, in 2019. She is a PhD researcher with Chair for Human-Centered Artificial Intelligence, University of Augsburg, Germany. Her doctoral work focuses on developing real-time interventions to support emotion regulation, with a broader research focus on affective computing and biosignal processing.

Elisabeth André is a full professor of computer science and founding chair of Human-Centered Artificial Intelligence with 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, and 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 10 most influential figures in the history of AI in Germany by National Society for Informatics (GI). From 2019–2022, she was serving as the editor-in-chief of *IEEE Transactions on Affective Computing*.