Seismocardiography for Emotion Recognition: A Study on EmoWear With Insights From DEAP

Mohammad Hasan Rahmani, Rafael Berkvens, *Member, IEEE*, and Maarten Weyn, *Member, IEEE*

***Abstract*—Emotions have a profound impact on our daily lives, influencing our thoughts, behaviors, and interactions, but also our physiological reactions. Recent advances in wearable technology have facilitated studying emotions through cardio-respiratory signals. Accelerometers offer a non-invasive, convenient, and cost-effective method for capturing heart- and pulmonary-induced vibrations on the chest wall, specifically Seismocardiography (SCG) and Accelerometry-Derived Respiration (ADR). Their affordability, wide availability, and ability to provide rich contextual data make accelerometers ideal for everyday use. While accelerometers have been used as part of broader modality fusions for Emotion Recognition (ER), their stand-alone potential via SCG and ADR remains unexplored. Bridging this gap could significantly help the embedding of ER into real-world applications, minimizing the hardware, and increasing contextual integration potentials. To address this gap, we introduce SCG and ADR as novel modalities for ER and evaluate their performance using the EmoWear dataset. First, we replicate the single-trial emotion classification pipeline from the DEAP dataset study, achieving similar results. Then we use our validated pipeline to train models that predict affective valence-arousal states using SCG and compare them against established cardiac signals, Electrocardiography (ECG) and Blood Volume Pulse (BVP). Results show that SCG is a viable modality for ER, achieving similar performance to ECG and BVP. By combining ADR with SCG, we achieved a working ER framework that only requires a single chest-worn accelerometer. These findings pave the way for integrating ER into real-world, enabling seamless affective computing in everyday life.**

***Index Terms*—Seismocardiography, EmoWear, DEAP, emotion recognition, accelerometer.**

I. Introduction

EMOTION recognition enhances human-computer interaction by enabling systems to understand and respond to users' emotional states [1]. It has been studied and improved under the broader field of affective computing, with applications spanning healthcare [2], entertainment [3], education [4], and marketing [5]. *Emotions* are part of the broader category of *affective states*, although the terms have been used interchangeably in the literature [6], including in this paper. In a reciprocal relationship, affective states influence physiological responses of the Autonomic Nervous System (ANS), such as heart rate, breathing, and skin conductance [7], [8]. Therefore, measuring physiological signals, including ECG, BVP, respiration (RSP), Electrodermal Activity (EDA), and Skin Temperature (SKT), provides indirect insights into the affective states, and can be used for ER [6]. Recent advancements in wearable sensor technology have also facilitated unobtrusive and convenient ER research [9].

Wearable-based ER research relies on the assumed link between the subjective emotional experiences and their objective representation in physiological signals [6]. Emotional models try to explain these subjective experiences in structured and quantifiable terms. They are often categorized in two broad types: *discrete* and *dimensional* models [10]. Discrete models represent emotions using distinct categories, such as happiness, sadness, anger, and fear, with Ekman's basic emotions model [11] being a well-known example. Dimensional models, on the other hand, represent emotions in a continuous space, typically using two or three dimensions. Russell's circumplex model of affect [12] is a widely used dimensional model that represents emotions in a two-dimensional space: valence (pleasant versus unpleasant) and arousal (activation versus deactivation). The limbic system, plays a key role in processing emotions and influencing the ANS through its connections with the hypothalamus [13]. External stimuli can elicit emotional responses in the limbic system, which in turn manifest as physiological changes in the ANS, such as variations in breathing and heart rate [13], [14]. Researchers aim to develop models that map the objective representation of emotions in physiological signals to the subjective experiences of emotions, often measured using self-report or expert-assessed quantitative scales, such as the Self-Assessment Manikins (SAM) [15]. Fig. 1 illustrates the top-level overview of the explained hypothesis, on which this study is based.

Currently, cardiac signals used for ER include ECG and BVP [16], [17]. Both of these signals measure the heart's rhythmic activity, which is reflected in Heart Rate Variability (HRV) information. HRV is modulated by the ANS in response to emotional stimuli [14]. SCG, which measures heart-originated vibrations on the chest wall, also carries HRV information; however, its potential for ER remains unexplored [18]. SCG is a non-invasive method for sensing cardiac mechanical activity,

Received 29 November 2024; revised 20 May 2025; accepted 28 May 2025. Date of publication 2 June 2025; date of current version 3 December 2025. Recommended for acceptance by Y. Liu. *(Corresponding author: Mohammad Hasan Rahmani.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Ethics Committee for the Social Sciences and Humanities of the UAntwerp under Application No. SHW_22_035, and performed in line with the GDPR, Belgian data protection law, Royal Decree (7 May 2004), and EU/institutional ethics codes.

The authors are with the IDLab - Faculty of Applied Engineering, University of Antwerp - imec, 2000 Antwerp, Belgium (e-mail: mohammad.rahmani@uantwerpen.be; rafael.berkvens@uantwerpen.be; maarten.weyn@uantwerpen.be).

Digital Object Identifier 10.1109/TAFFC.2025.3575281

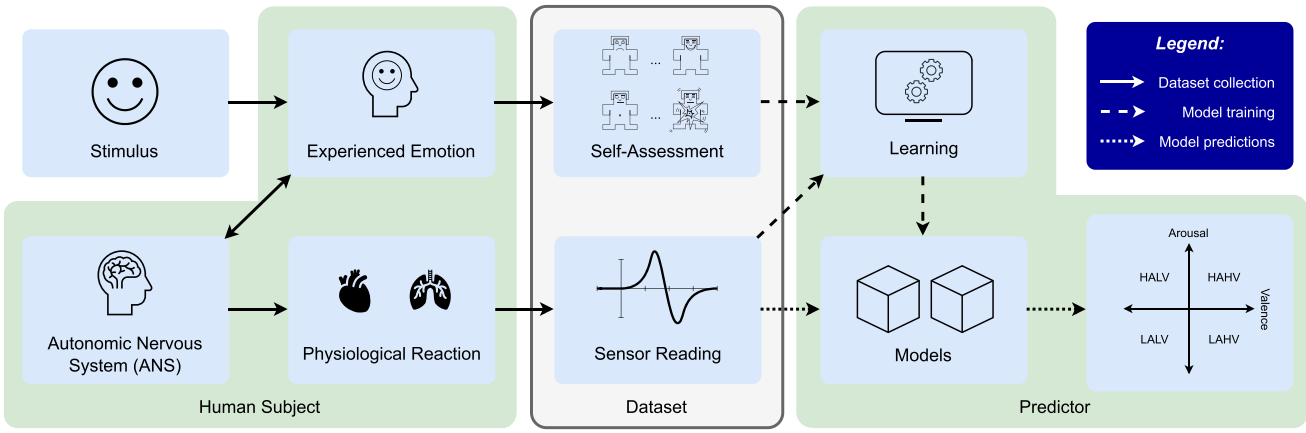


Fig. 1. Top-level overview of the hypothesis underlying this study: Emotion recognition research aims to map the objective representation of emotions in physiological signals to the subjective experiences of emotions, as captured by self-assessed emotional responses (HAHV: High Arousal-High Valence, HALV: High Arousal-Low Valence, LALV: Low Arousal-Low Valence, LAHV: Low Arousal-High Valence).

typically captured through chest-worn accelerometers [19]. It is often used for cardiac health monitoring through extracting the timing information [20].

SCG can be measured using a chest-worn accelerometer, offering a convenient and cost-effective method for capturing heart-induced vibrations on the chest wall. In the field of ER, SCG is expected to provide HRV information similar to ECG and BVP, with the added advantage of being measured using the same accelerometer used for various other applications. The way SCG is sensed, gives it a unique advantage over ECG and BVP in terms of versatility and cost-effectiveness. Accelerometers are widely available and can provide rich contextual data, making them ideal for everyday use. The same device can support multiple applications, such as activity recognition, fall detection, and gait analysis. In our previous work, we surveyed these applications and categorized their use cases across seven different domains [21]. Another signal that can be inferred from the same chest-worn accelerometer is the ADR, which reflects the chest wall's movement due to respiration [22]. By combining SCG with ADR, we can infer both cardiac and respiratory signals from a single chest-worn accelerometer, offering a more comprehensive view of the physiological responses to emotions.

Combining SCG with ADR for ER is a practically interesting approach, as both can be captured using the same chest-worn accelerometer. Theoretically, heart and respiration signals are tightly interconnected through the autonomic nervous system. Heart activity is influenced by both sympathetic and parasympathetic nervous systems, which respectively accelerate or decelerate heart rate. Respiratory rhythm is also controlled by autonomic inputs but can be modulated by conscious control. During an emotional event, sympathetic arousal (as in fear or excitement) will typically raise heart rate and blood pressure, and also increase respiration rate and tidal volume [23], [24]. Parasympathetic activity, on the other hand, promotes calm states, slowing the heart and stabilizing breathing. By monitoring both signals, an ER system can capture these concurrent changes.

Several datasets have been developed to facilitate ER research using physiological signals [25], [26], [27], [28], [29], [30], [31], [32], [33] (for a comprehensive comparison of the existing datasets, we refer the reader to Kang et al. [33]). The EmoWear dataset [18] is the only one to provide chest-worn accelerometer data validated for both SCG and ADR purposes. We collected and published this dataset to enable exploring SCG and ADR for ER. EmoWear builds upon DEAP [25] with the added elements of user mobility and context change. Therefore, in terms of experimental setup and data collection process, DEAP is the closest dataset to EmoWear. Both datasets are based on Russell's dimensional circumplex model of affect [12]. They both use video stimuli and collect self-assessed emotional responses using the SAM [15] scale. EmoWear uses stimuli from the same pool as DEAP (originally 40 stimuli, with 38 used in EmoWear). Both datasets provide cardiac and respiratory signal recordings, along with some additional peripheral modalities such as EDA and SKT. Essentially, EmoWear has all the peripheral modalities that DEAP has, except for Electromyography (EMG) and Electrooculography (EOG), but plus SCG and ADR. The similarities between the two datasets make DEAP an ideal benchmark for validating the EmoWear dataset in ER research, as it offers the highest resemblance when replicated.

This study evaluates SCG as a novel modality for ER, addressing a gap in the existing literature. Moreover, we explore the potential of using a single chest-worn accelerometer for ER by combining SCG with ADR. Motion was used as a modality for ER in the past [34], [35], but only in the context of measuring wrist or body movements. This is the first study to investigate SCG as HRV-providing alternative for ER. Our approach aims to pave the way for integrating wearable-based ER in real-world applications. Although our experiments are conducted on the EmoWear dataset, we also use the benchmark DEAP dataset to establish and strengthen our methodology and findings. Our main research questions are: 1) Can SCG be effectively used for ER? 2) How do results from SCG compare to the conventional cardiac signals, ECG and BVP?, and 3) How does a single chest-worn accelerometer perform for ER when combining SCG with ADR?

To address our research questions, we base our methodology on established best practices in the field. A review of the existing literature provides both the necessary background and reference

results for comparison. However, this review revealed several methodological limitations in the current literature that could affect the interpretability and generalizability of the reported findings. We aim to also address these limitations in our study and base our work on a robust and reliable methodology which is also comparable to existing results. Following contributions are made in this paper:

- A critical review of existing methodologies and reported results on the use of peripheral physiological signals (BVP and RSP, especially) from the DEAP dataset for ER.
- Replicating the DEAP dataset's single-trial emotion classification results, for reaffirming their findings and validating our analytical pipeline.
- Introduction of SCG as a novel gateway for ER and comparison of its performance against established cardiac signals.
- Examining the potential of a single chest-worn accelerometer for ER by combining SCG with ADR.
- Setting the first benchmarks on the EmoWear dataset for ER research.

The rest of the paper is organized as follows: Section II reviews existing ER research on peripheral physiological signals from the DEAP dataset. Section III presents the methodology used in this study, including the datasets, signal processing, feature extraction, and machine learning models. Section IV reports the experimental results, and Section V discusses the findings. Finally, Section VI concludes the paper with a summary of the contributions and implications of the study.

II. Related Work

Review [6] distinguishes between *classical feature-based* models and the *end-to-end* deep learning models for ER using physiological signals from wearable sensors. The former approach relies heavily on domain expertise for feature engineering, requiring the extraction and selection of meaningful physiological signal features before classification. In contrast, the latter processes raw signals directly, automatically capturing temporal and spatial patterns inherent in physiological data, and learning the features that are most relevant for the task. The end-to-end approach is a more recent development, used by only a minority of 12 % of their reviewed studies [6]. Although this method shows promise by bypassing the limitations of traditional feature engineering, its limited use in wearable-based field studies may be attributed to the limited dataset sizes and computational resources.

More recent studies have addressed the limitations associated with the dataset size by using self-supervised representation learning [36], [37] and transfer learning [38], [39]. In self-supervised learning the model is trained using automatically generated labels from the input data itself. It can be used to learn meaningful representations of specific signals (e.g., ECG) on a large dataset to be used for a specific downstream task on a smaller dataset. Transfer learning, on the other hand, is a technique where a model trained on one task is used for another related task. Both techniques help in undermining the limitations of the dataset, and have shown promising results in ER research.

While these techniques show promise, our current study uses the traditional feature-based approach as we focus on providing initial insights into the potential of SCG for ER and the EmoWear dataset [18]. EmoWear is a recently published dataset that offers the unique opportunity to explore the potential of SCG and ADR for ER. It shares methodological similarities (see Section III-A) with the benchmark DEAP dataset [25], which has been extensively used for ER research since its release in 2012. In both datasets, video stimuli are presented to elicit emotions, concurrent with the recording of physiological signals and self-assessed emotional responses.

With the aim of learning from past work and identifying best practices for the EmoWear dataset, this section reviews existing ER research on peripheral physiological signals from the DEAP dataset. This way we establish the rationale for our methodological approach as well as a ground to compare our results against. Although the DEAP dataset contains a range of physiological recordings, our review focuses on studies that predict emotions using the BVP and RSP modalities, as these align with the scope of our current study. Table I provides a list of the related work that report findings on these specific modalities. It shows a variety of combinations of methods used to measure the predictive power of different peripheral physiological signals for ER. Different combinations of modalities, classifiers, data split approaches, performance metrics, and statistical analysis of the results, have been used to classify emotions and report findings.

Before diving into the details of these studies, it is important to note that the reviewed aspects are not exhaustive. There are many other detail aspects differentiating the studies, that are not covered by the table, such as the preprocessing steps, feature extraction methods, and the hyperparameters of the classifiers. For a more comprehensive analysis of the associated methods for ER, we refer the readers to the original cited references and plenty of review papers on the topic [6], [9], [40], [41].

A deeper look at the studies in Table I reveal certain methodological limitations that potentially impact the interpretability and generalizability of the reported results. We investigate these limitations across two main pillars: "performance measurement and reporting", and "data splitting and validation". In what follows, we provide an overview of these limitations and highlight the key methodological gaps we found. At the end of this section, we conclude with a summary of the identified gaps and the need for a robust and reliable methodology to address them.

A. Performance Measurement and Reporting

We identified several shortcomings in the performance measurement and reporting of the reviewed studies, which we categorized in the following three areas:

1) Accuracy and the Class Imbalance: When the emotional valence and arousal in the DEAP dataset are binarized, a high class imbalance emerges [25]. Such imbalance can lead to biased classifiers that favor the majority class [6], an effect that can be hidden when accuracy is the sole performance metric for evaluation. Reporting robust performance metrics that are less sensitive to class imbalance, such as the $F1$ score, can help mitigate this issue [25]. Among the 18 reviewed studies, 7

TABLE I
OVERVIEW OF THE RELATED WORK THAT REPORT ON BVP AND RSP MODALITIES OF THE DEAP DATASET

Reference	Modalities	Classifier	Data split	Performance Metrics	Statistical Test
Koelstra et al. [25] (Original DEAP paper)	BVP, RSP, EDA, SKT, EMG, EOG, EEG	NB	LOVO CV, PS	A $F1_{ma}$	one-sample t-test
Godin et al. [42]	BVP, RSP, EDA, SKT, EMG, EOG	NB	LOSO CV	A	-
Chen et al. [43]	BVP, RSP, EDA, SKT, EMG, EOG, EEG	SVM	10-fold CV, MS	A $F1_{ma}$	t-test
Zhang et al. [44]	RSP	LR	80% training, 20% testing, MS	A	-
Yin et al. [45]	BVP, RSP, EDA, SKT, EMG, EOG, EEG	k-NN, LR, MESAE, NB, SVM	10-fold CV, PS	A $F1_{sc}$	paired t-test
Ayata et al. [46]	BVP, EDA	DT, RF, k-NN, SVM	10-fold CV, MS	A	-
Choi et al. [47]	BVP, EDA, EEG	LSTM	80% training, 20% testing, SS (except for 1)	$\frac{TP_{sc?}}{TN_{sc?}}$	-
Lee et al. [48] (2019)	BVP	CNN	80% training, 20% testing, MS	A	-
Lee et al. [49] (2020)	BVP	CNN	80% training, 20% testing, MS (selected 20 out of 32)	A	-
Wu et al. [50]	BVP, RSP, EDA, SKT, EMG, EOG, EEG	CNN, LSTM	-	A	-
Zhu et al. [51]	BVP, RSP, EDA, SKT, EMG, EOG, EEG, Face Video	CNN, MLP	80% training, 20% testing, MS? (selected 18 out of 32)	A	-
Elalamy et al. [52]	BVP, EDA	LR	LOSO CV	A $F1_{sc?}$	paired Wilcoxon test
Kang et al. [53]	BVP, EDA	CAE	64% training, 16% validation, 20% testing, MS	A $F1_{sc?}$	-
Pidgeon et al. [54]	BVP, RSP, EDA, SKT	CNN	Stratified 10-fold CV, MS	A $F1_{sc?}$	-
Siddharth et al. [55]	BVP, EDA, Face Video	-	LOSO CV	A $F1_{sc?}$	two-sample t-test
Shubha et al. [56]	BVP, EDA	DT, RF, SVM, LR, AdaBoost	-	A $F1_{sc?}$	-
Gohumpu et al. [57]	BVP, EDA, SKT	DT, k-NN, SVM	80% training, 20% testing, MS? + 2-fold CV	A $F1_{sc?}$	-
Bamonte et al. [58]	BVP, EDA	k-NN, DT, RF, SVM, GBM, CNN	Stratified shuffle split PS (80% training, 20% testing repeated 10 times PS)	A $F1_{sc?}$	-

Several combinations of the methods are used to develop and evaluate the models. PS: Per Subject, MS: Mixed Subjects, SS: Separate Subjects, CV: Cross Validation, LOSO: Leave-One-Subject-Out, LOVO: Leave-One-Video-out, A: Accuracy, *TP* : True Positive, TN: True egative. Ma: macro Averaged, sc: single class, -: not reported, ?: not explicitly reported.

(39 %) reported the accuracy as their sole performance metric, 1 (5 %) reported True Positive to True Negative ratio, and 10 (56 %) reported the $F1$ score along with the accuracy.

Some studies have used stratified data splitting techniques which ensures that the class distribution is preserved in the training and testing sets. Although stratification may improve evaluation, the model still sees imbalanced data during its training and may still learn the biased patterns. As a result, class stratification does not eliminate the need for robust performance metrics that are less sensitive to class imbalance. Among the reviewed studies, 2 (11 %) used stratified data splitting techniques, both commonly reporting the $F1$ score as a performance metric.

2) F1 Scoring Approach: For imbalanced datasets, the $F1$ score is considered a more reliable performance metric than accuracy [25]. The $F1$ score is a harmonic mean of the precision and recall, which is designed to consider the trade-off between the two. Both precision and recall are calculated based on the True Positive (TP), False Positive (FP), and False Negative (FN) counts in the classifier predictions. The first letter (T or F) in the abbreviations indicates the correctness of the prediction, and the second letter (P or N) indicates the actual class of the sample.

When calculating the $F1$ score, it is necessary to define the class of interest upfront. This is the class for which TP, FP, and FN will be counted. Defining the class of interest also implies that the $F1$ score is essentially a single-class performance metric, and the reported $F1$ score should be interpreted in the context of the pre-defined class. Therefore, the following approaches can be considered when reporting the $F1$ score: 1) reporting $F1$ for the class of interest only, or 2) reporting $F1$ for all classes separately, or 3) reporting an average $F1$ over all classes.

For the classification of "high" versus "low" emotions in the DEAP dataset, classes emerge from binarization of the continuous SAM [15] that measure the emotional dimensions (pleasant versus unpleasant for valence, and activation versus deactivation for arousal). In real-world applications, recognizing both high and low states is equally important as they represent the four quadrants of the valence-arousal space [12], [59]. In such scenario, reporting a single-class $F1$ score is hard to interpret [60], [61], as it 1) provides an incomplete picture of the classifier's performance, and 2) may lead to biased interpretations, especially if the chosen class is the majority class. As a result, the $F1$ score should either be reported separately per class, or as an average $F1$ score over both classes. The averaging can be weighted by the class distribution (weighted averaging), unweighted (macro averaging), or considered in the counting of the TP, FP, and FN (micro averaging). In the reviewed

studies and among the 10 studies that reported the $F1$ score, 7 (70 %) did not specify the class for which the $F1$ score was reported, 1 (10 %) reported the $F1$ score only for the "low" class, and 2 (20 %) reported the $F1$ score using the macro averaging approach. With 16 (89 %) of the reviewed studies that aim to classify emotions using BVP and RSP modalities of the DEAP dataset not reporting any form of imbalance-tolerant metrics or not reporting them sufficiently, we identify a significant gap in performance reporting. This gap may have led to severely inflated or biased conclusions in these papers.

3) Statistical Significance Analysis: The statistical significance analysis help determining whether the observed differences in performance are due to the model's learning ability or random chance. They strengthen the reliability of the reported performance and any conclusions drawn from it. When the performance are averaged over groups of samples (e.g., folds in Cross-Validation (CV) or individual results in subject dependent classification), statistical significance tests can give an indication of the reliability of the reported average. In case of comparing groups of results, these tests can help determine whether the observed differences are statistically significant [62]. Among the reviewed studies, 5 (28 %) conducted statistical significance tests, while 13 (72 %) did not report any statistical significance analysis. Different forms of *t*-tests, were the most commonly used statistical significance tests used by the reviewed studies.

B. Data Splitting and Validation

Data splitting strategies and validation techniques are crucial aspects of classification tasks. Below we discuss the limitations observed in the reviewed studies from two perspectives: reliable generalization and transparency in selection criteria.

1) Reliable Generalization: An essential aspect of model evaluation is ensuring that results generalize across different data subsets. The simplest way to achieve this is to split the dataset into two parts: a training set and a testing set. The model is trained on the training set and evaluated on the testing set. It is a simple and quick approach; however, it risks making the evaluation performance dependent on a very specific choice of data split. Furthermore, if the test set is repeatedly used during model selection or hyperparameter tuning, the model would essentially learn the test set in an implicit way. This may result in an overly optimistic estimate of performance that does not generalize to unseen data. A common workaround is to include a validation set in the initial split. The validation set is used for tuning and model selection, reserving the test set strictly for final evaluation. Without a validation set, it is hard to choose the best model or optimize hyperparameters without compromising the test set's role as an unbiased measure of generalization performance [63].

With a single, fixed train-(validation)-test split, the performance metrics may not be representative of the entire dataset, as a different split could produce significantly different results. A single, fixed split can lead to overfitting to the specific characteristics of that particular split. CV techniques help to mitigate bias that can arise from relying on such a single, fixed data split. CV evaluates the model on multiple subsets, providing a more comprehensive assessment of the model's generalization performance.

Among the studies reviewed, 6 (33 %) did not use any form of CV and instead reported performance based on a single data split. Of these, 5 studies (28 %) did not dedicate a separate validation set, increasing the risk of overfitting to the test data.

2) Transparency in Selection Criteria: Another limitation observed in certain studies is the selective use of part of the dataset without providing a transparent selection criteria. Selective sampling without explicit criteria risks introducing bias, especially if chosen samples share specific physiological traits or demographic characteristics. Transparent and consistent selection methodologies are important to make sure that findings are representative of the broader dataset [64]. Out of the reviewed studies, 2 (11 %) opted to use only a subset of the 32 subjects from the dataset but did not provide clear explanations for these selections. Another 2 (11 %) do not share any information about their data splitting methodology, making their results hard to interpret.

In summary, the reviewed studies reveal several methodological gaps that can impact the reliability, interpretability, and generalizability of their findings. Key issues in performance measurement and reporting include the lack of imbalance-tolerant metrics, imperfect $F1$ scoring approaches, and limited use of statistical significance testing. Moreover, the use of unreliable data splitting strategies and a lack of transparency in data selection further compromise the generalizability and trustworthiness of the findings in some of the reviewed studies. These limitations highlight the need for robust and transparent methodologies to ensure that conclusions drawn from the results are both reliable and meaningful. Future research must prioritize methodological rigor by systematically adopting best practices, including reliable data-splitting strategies, imbalance-tolerant metrics, proper statistical significance testing, and transparent reporting. When using $F1$ score, especial care should be taken to report a complete picture of the way it was achieved, taking into account that, unlike accuracy, it is essentially a single-class metric. Without these foundational elements, advancements in emotion recognition risk being built on fragile methodological grounds. We urge the community to integrate these practices as standard, rather than optional, to improve the reliability, interpretability, and generalizability of findings in this domain. In the next section, we present our method, designed to address these methodological gaps. We use a validated approach to deliver results that set reliable benchmarks on the EmoWear dataset. First, we validate the emotion classification pipeline used in the DEAP study on the DEAP dataset. Then, we apply the same pipeline to the EmoWear dataset.

III. Methodology

Our study can be divided into two main parts: the replication of the DEAP dataset's emotion classification pipeline, and the application of this pipeline to the EmoWear dataset. First, we replicated the single-trial emotion classification pipeline from the DEAP dataset and obtained similar results. Next, we applied the same approach to the EmoWear dataset, which offers

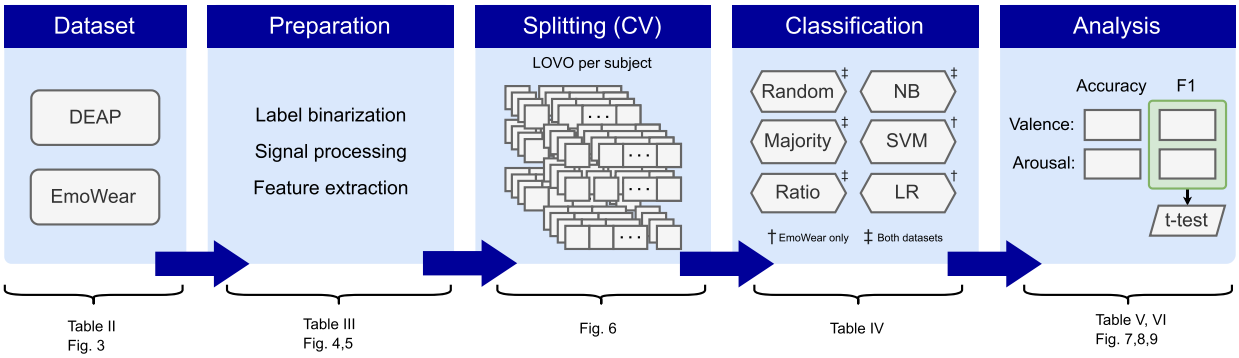


Fig. 2. Overview of the emotion classification pipeline used in this study. The pipeline is inspired by the single-trial emotion classification pipeline from the DEAP study, and is applied to both the DEAP and EmoWear datasets (CV: Cross Validation, LOVO: Leave-One-Video-Out, NB: Naïve Bayes, SVM: Support Vector Machine, LR: Logistic Regression).

TABLE II
CONTENT SUMMARY OF EMOWEAR AND DEAP DATASETS

	DEAP [25]	**EmoWear** [18], [65]
Participants	32 (15 F, 17 M)	48 (21 F, 27 M)
Recorded signals	EEG, EDA, BVP, RSP, SKT, EMG, EOG, Face Video	ACC (SCG & ADR), GYRO, ECG, BVP, RSP, EDA, SKT
Signal recording equipment	Non-portable sensors	Portable wearable sensors
Number of stimuli per participant	40 (10 HAHV, 10 LAHV, 10 LALV, 10 HALV)	38 (10 HAHV, 9 LAHV, 10 LALV, 9 HALV)
Rating method	Self-assessment manikins (SAM)	Self-assessment manikins (SAM)
Rating scales	Arousal, Valence, Dominance, Liking, Familiarity	Arousal, Valence, Dominance, Liking, Familiarity
Stimulus presenter / mediator GUI	Presentation ® [66]	ColEmo [67]
Publication year	2012	2024

F: Female, M: Male.

a new proxy signal, SCG, for ER. By first confirming our implementation using the DEAP benchmark, we ensured that we were applying a validated implementation to the EmoWear dataset, thereby better supporting our findings and providing more meaningful insights into the performance of the new proxy signal, SCG, for ER. Fig. 2 provides an overview of the emotion classification pipeline used in this study.

A. Dataset

We used the DEAP dataset [25] and the EmoWear dataset [18] for our study. These two datasets share the stimulus videos and the SAM rating scales, but differ in the recording equipment and the subjects. DEAP is a benchmark dataset for ER, recorded in a laboratory setting with non-portable sensors, while EmoWear is a recent dataset, recorded with portable wearable sensors. Looking at the sensor modalities of our interest, DEAP offers BVP for cardiac-, and RSP for respiratory-sensing. EmoWear offers BVP and ECG for cardiac-, and RSP for respiratory-sensing. EmoWear also offers the chest-recorded vibration data through Accelerometer (ACC) which we use to infer both the SCG and the ADR signals for cardiac- and respiratory-sensing, respectively. Table II provides a content summary of the two datasets.

Ethical Statement: In this study, we are using the publicly available datasets in accordance with their respective terms of use. The EmoWear dataset was collected by us in accordance with the ethical guidelines of the University of Antwerp, and following a positive ethical clearance decision. Our use of the DEAP dataset is bounded by their end-user license agreement which is limited for research purposes. We use the physiological signals present in the DEAP dataset from those subjects who have given their consent for the distribution of their physiological recordings.

Packages Used: We used the preprocessed DEAP dataset package and the second phase (elicit-assess-walk cycles) of the EmoWear dataset's CSV package [65]. The preprocessed DEAP dataset offers a ready-to-use downsampled, filtered, and segmented version of the physiological signals, offered in Matlab and Python pickle formats. We used the Python pickle format for our study. The EmoWear dataset is recorded in two phases: the first phase is the vocal vibration recording, and the second phase involves the elicit-assess-walk cycles. We used the second phase data which is the relevant phase for ER research. The dataset is offered in three packages: the raw data package, the Matlab data package, and the CSV package. The CSV package offers the synchronized and segmented data of the physiological signals and the self-assessed emotional responses, which we used for our study.

Subject Demographics: The DEAP dataset contains 32 subjects (15 females and 17 males) aged between 19 and 37 years (mean=26.9), while the EmoWear dataset contains 48 subjects (21 females and 27 males) aged between 21 and 45 years (mean=29.3). One subject in the DEAP dataset reported a history of migraine, while three subjects in the EmoWear dataset reported histories of ADHD - autism, panic attacks, and epilepsy.

Exclusion Criteria: We excluded subjects whose frontal chest accelerometer data was missing or corrupted (subjects 1, 3, 35 from the EmoWear dataset). We also excluded subjects whose emotion ratings were severely imbalanced. To this end, we considered an empirical threshold of 10 % for availability of high and low ratings for both binarized valence and arousal, below which the subject was excluded. This criterion resulted in the

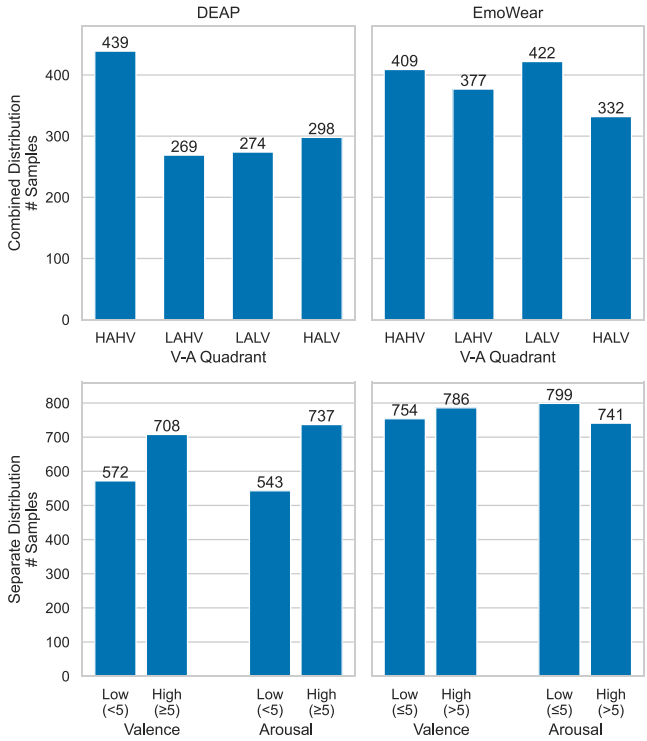


Fig. 3. Class distribution of the DEAP and EmoWear datasets, after applying the exclusion criteria.

exclusion of subjects 19, 21, and 32 from the EmoWear dataset. Also, we excluded subject 34 from the EmoWear dataset due to the use of a different sensor unit for the chest accelerometer data. The final number of subjects used in our study was 42 from the EmoWear dataset. In addition to the mentioned subject exclusion criteria, we excluded single trials identified as faulty in the original EmoWear data description [65]. These faulty trials were identified based on severe disruptions during the experiments where: 1) participant was distracted by external factors during stimulus presentation (accidental objects falling), and 2) video stimulus was not presented correctly (sound issues). For the DEAP dataset, no subjects or trials were excluded because none met any of the specified exclusion criteria. Fig. 3 shows the class distribution of the DEAP and EmoWear datasets after applying the exclusion criteria.

B. Signal Processing

Signal processing involved preprocessing the raw physiological signals, and extracting features from them. While all other signals have been directly recorded, the SCG and ADR signals are derived from the ACC signal. Therefore, we first describe the signal processing steps for the SCG and ADR signals, separately in detail, followed by the preprocessing steps for the other signals. All signal processing including preprocessing and feature extraction was done in Python 3.10.14 using the *NumPy* [68], *pandas* [69], [70], and the *NeuroKit2* [71] libraries.

Seismocardiography: The so-called R-peaks of the ECG signal represent ventricular depolarization, an electrical event that initiates the mechanical contraction of the left ventricle. Once the pressure in the left ventricle exceeds the pressure in the aorta, the aortic valve opens [72], leading to Aortic valve Opening (AO) peak in the SCG signal. The aortic valve opens shortly after the R-peak due to the time required for ventricular contraction to generate sufficient pressure to overcome aortic pressure [73], [74]. This corresponding relationship makes the AO peak our best alternative to the R-peak for cardiac activity sensing in the SCG signal. Fig. 4(a) shows an example of the ECG and SCG signals from the EmoWear dataset with marked R-peaks and AO-peaks.

We use the AO peaks for our analysis of the Heart Rate (HR) and HRV indexes from the SCG signal. We took the data of dorsoventral (z) axis from the frontal ACC_3 (LSM6DSOX) sensor of the EmoWear dataset. Since the accelerometer data was sampled irregularly [18], we interpolated the data to a regular 200 Hz sampling rate, which we used as our SCG signal.

Next, we detected the AO peaks in the SCG signal using the algorithm proposed by Massaroni et al. [75]. Our implementation involves an initial band-pass filtering with a pass-band of 10–20 Hz, envelope detection using the Hilbert transform, a second band-pass filtering with a pass-band of 0.5–2 Hz, and a final peak detection algorithm. Fig. 4(a) shows detected AO-peaks in an initially-filtered SCG signal.

Accelerometry-Derived Respiration: Chest-sensed acceleration signals also carry respiratory information [21]. Many studies have validated ADR against the gold standard respiratory signals, such as spirometry [76] and Respiratory Inductance Plethysmography (RIP) [22]. We derive the ADR signal from the chest-sensed ACC signal by applying a band-pass filter of 0.15–0.35 Hz to extract the respiratory frequency band, and then a baseline detrending (same methodology of BioSPPy toolbox [77]). Fig. 4(b) shows an example of the RSP (using capacitive breathing sensor of Zephyr BioHarness 3 [18]) and ADR signals from the EmoWear dataset, before and after the mentioned filtering/detrending steps.

Peripheral Physiological Signals: Fig. 5 shows a summary of the data sources used in our study and the signal processing pipeline applied to them. We used the BVP, RSP, EDA, EMG, EOG, and SKT signals from the DEAP dataset, and the ECG, BVP, SCG, RSP, ADR, EDA, and SKT signals from the EmoWear dataset.

Different signals in both datasets were preprocessed to remove noise and artifacts, and prepare them for feature extraction. The ECG signal was band-pass filtered between 8–20 Hz [78]. The BVP and EDA signals were detrended by subtracting a 256-point moving average from the signal [25]. Similar to the ADR, the RSP signal was band-pass filtered between 0.15–0.35 Hz [77]. Finally, the EMG and EOG signals were both band-pass filtered between 4–40 Hz [25]. The SKT signal was not preprocessed.

Feature Extraction: Different hand-crafted features were extracted from the preprocessed signals based on the literature. We extracted the set of features that were used by the DEAP dataset to enable replication of their results. For the cardiac (ECG, BVP, and SCG), and the respiratory (RSP and ADR) signals, we also extracted additional features than those reported in the DEAP dataset based on more recent common practices in the literature.

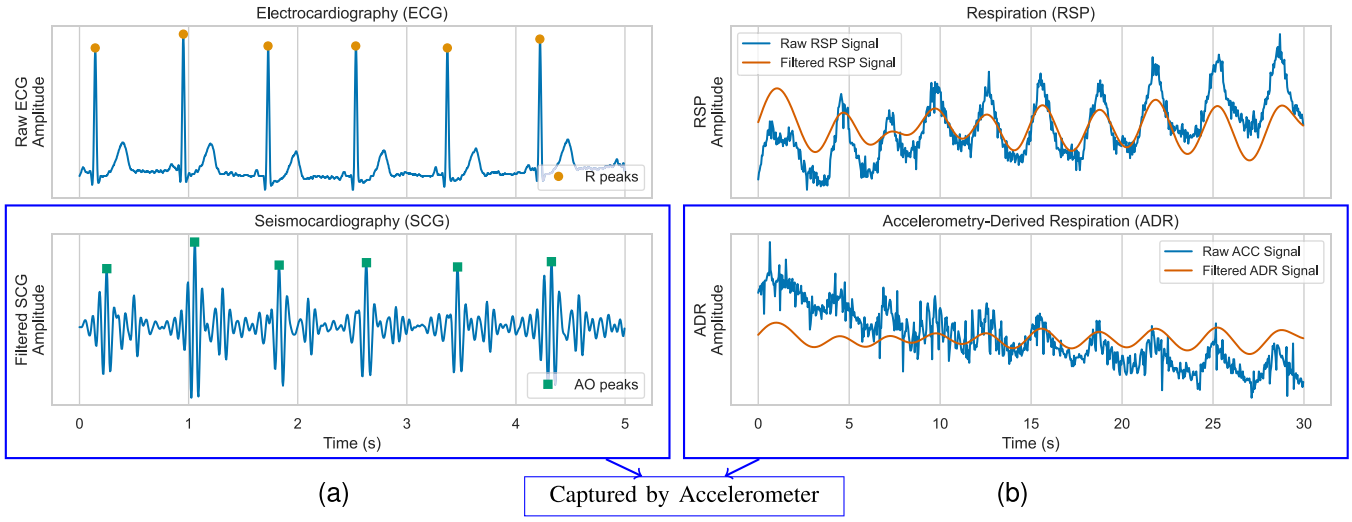


Fig. 4. Example signals from the EmoWear dataset representing how chest-worn accelerometer data hold cardio-respiratory information. (a) Electrocardiography (ECG) and Seismocardiography (SCG) signals with detected R-peaks and AO-peaks. (b) Respiration (RSP) and Accelerometry-Derived Respiration (ADR) before and after filtering.

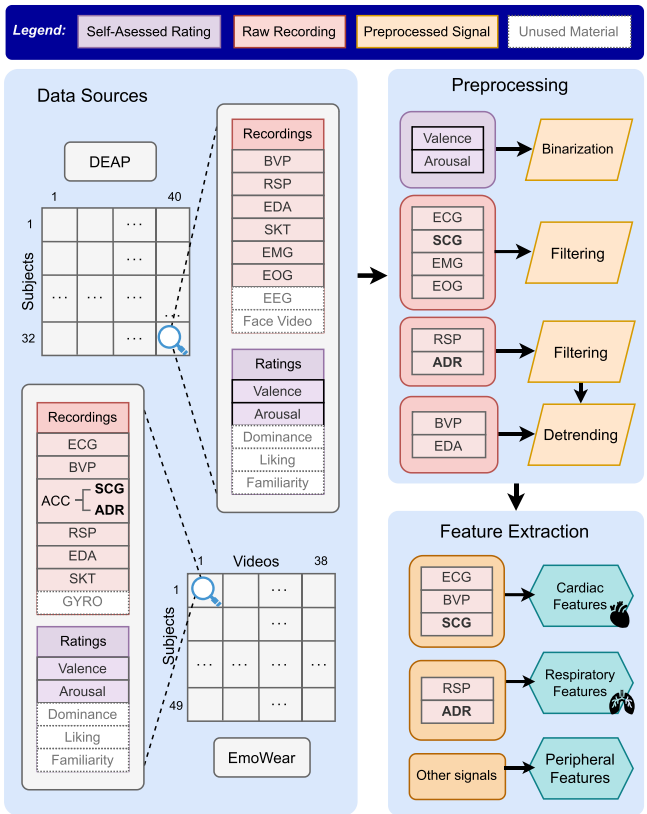


Fig. 5. Signal processing and feature extraction pipeline for different types of physiological signals in both DEAP and EmoWear datasets.

TABLE III
SUMMARY OF EXTRACTED FEATURES FROM PHYSIOLOGICAL SIGNALS

Signal	Extracted Feature
ECG, BVP, SCG	Mean and standard deviation of inter-beat intervals (IBI), heart rate (HR), and IBI derivatives. Frequency band power ratio (0.04–0.15 Hz vs. 0.15–0.5 Hz) for IBI. Spectral power in IBI bands: 0.1–0.2 Hz, 0.2–0.3 Hz, 0.3–0.4 Hz. Spectral power in IBI derivatives: low (0.01–0.08 Hz), medium (0.08–0.15 Hz), and high (0.15–0.5 Hz) bands.
	Additionals:* CVNN, CVSD, HFn, HTI, IQRNN, LFn, LnHF, MCVNN, MadNN, MaxNN, MedianNN, MinNN, Prc20NN, Prc80NN, RMSSD, SDRMSSD, TINN, TP, pNN20, pNN50
RSP, ADR	Mean and standard deviation of signal. Mean of signal derivatives. Breathing rate. Mean and median of peak-to-peak intervals. Logarithmic band energy difference (0.05–0.25 Hz and 0.25–5 Hz). Spectral centroid of breathing rhythm. Band power up to 2.4 Hz.
	Additionals:* CVSD, RMSSD, ApEn, CVBB, HF, LF, LFHF, MCVBB, MadBB, SD1, SDBB, SDSD, RVT
EDA	Mean resistance and mean derivative values. Mean decrease rate during decay periods. Proportion of negative derivative samples. Local minima count. Mean rise time. Band power up to 2.4 Hz. Zero-crossing rates for skin conductance slow, and very slow responses (SCSR and SCVSR). Mean peak magnitudes of SCSR and SCVSR.
SKT	Mean temperature, mean of derivatives, and spectral power in low-frequency bands (0–0.1 Hz and 0.1–0.2 Hz).
EMG	Signal energy, and basic statistical measures (mean, variance).
EOG	Same as EMG + eye blinking rate.

*For a full description of the terms of the additional features, visit the NeuroKit2 documentation [79].

The additional features were only extracted from the EmoWear dataset. Table III lists the features extracted from the different physiological signals.

C. Machine Learning

Classification Tasks: The circumplex model of affect [12] defines emotions in a two-dimensional space of valence and arousal. This model assumes independence between the two dimensions [6]. Based on this assumption, we considered two independent classification problems: the binary classification of either of the emotional dimensions - valence and arousal. The

TABLE IV
SUMMARY OF CLASSIFIER INPUT SCENARIOS EXPERIMENTED IN THE STUDY

Input Category	Dataset	Cardiac			Respiratory		Other Peripherals				Input Scenario	Presented Results
		BVP	ECG	SCG	RSP	ADR	EDA	SKT	EMG	EOG		
Replication	DEAP	✓			✓		✓	✓	✓	✓	BVP + all	Table V
All-Modality Sets	EmoWear	✓			✓		✓	✓			BVP + all	Table VI, Fig. 7, Fig. 9
			✓		✓		✓	✓			ECG + all	
				✓	✓	✓	✓	✓			**SCG** + all	
Cardio-Respiratory Sets		✓			✓						BVP + RSP	Table VI, Fig. 7, Fig. 8, Fig. 9
			✓		✓						ECG + RSP	
				✓	✓						**SCG** + RSP	
				✓		✓					**SCG + ADR**	

binarization of the continuous SAM ratings into high and low values defines the classes.

Input Sets: We fed the machine learning classifiers with the hand-crafted features explained in Section III-B. For the DEAP dataset, we used all the available peripheral physiological signals, to replicate their results (similar to the original DEAP study). For the EmoWear dataset, we experimented with two types of input set combinations: 1) all-modality sets: one of the cardiac recordings (ECG / BVP / SCG) along with all other available peripheral signals (RSP, ADR, EDA, and SKT), and 2) cardio-respiratory sets: one of the cardiac recordings (ECG / BVP / SCG) along with one of the respiratory recordings (RSP / ADR). In either case, we never combined ADR with ECG or BVP, as ADR is derived from the same sensor as SCG and is therefore only available when SCG is. The all-modality sets provide a ground for comparison with the DEAP dataset using various cardiac sources, while the cardio-respiratory sets allow us to further investigate the potential of the SCG and ADR signals (essentially, a single chest-worn accelerometer) for ER. Table IV summarizes the modality-combination input scenarios experimented in the study.

Target Labels: We binarized the self-assessed valence and arousal ratings into high and low values based on the mid-threshold of 5 in their 1–9 scales. Such binarization divides the valence-arousal space into four quadrants: High Arousal-High Valence (HAHV), Low Arousal-High Valence (LAHV), Low Arousal-Low Valence (LALV), and High Arousal-Low Valence (HALV). The dominance, liking, and familiarity dimensions were not used in our study.

Feature Selection: We use Fisher's linear discriminant score to rank the features based on their discriminative power. The Fisher score J of a feature f is calculated as:

$$J(f) = \frac{|\mu_1 - \mu_2|}{\sigma_1^2 + \sigma_2^2} \tag{1}$$

where μ_1 and μ_2 are the means, and σ_1 and σ_2 are the standard deviations of the feature f for the two classes. The feature was selected if its Fisher score was above the empirical threshold of 0.3. For the EmoWear dataset, we combined this threshold with a pre-determined minimum ranked feature count of 15 to ensure that the feature set did not become too small.

Classifiers: We used three classic classifiers: Gaussian Naïve Bayes (NB), Support Vector Machine (SVM), and Logistic Regression (LR). The NB classifier is a simple probabilistic classifier based on the Bayes theorem with strong independence assumptions between the features. Despite the fact that the assumption of independence does not apply to our type of extracted features (Table III), the NB classifier has been successfully employed by similar studies [6], [25]. However, our choice of this classifier was mainly based on: 1) replicating the DEAP dataset's analytical pipeline, 2) enabling a direct comparison of the results.

The choice of SVM and LR classifiers was based on our initial investigation of different classifiers and hyperparameters. We took k-Nearest Neighbors, Decision Trees, Random Forest, boosting classifiers (including Adaboost, Gradient Boosting, and XGBoost), and Neural Networks (including Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM)). We experimented with different hyperparameters and optimized their performance using a grid search. The goal was to maximize the macro-averaged F1 score over the high and low classes of the emotional dimensions in the entire EmoWear dataset. The best performing classifiers were the SVM and LR classifiers with L2 regularization, balanced class weights, and linear kernels. Based on this initial investigation, we limited our scope to these classifiers, only varying the regularization strength parameter (usually known as C parameter) for both classifiers, and the solver parameter for the LR classifier.

Baseline Scores: To establish a baseline for our study, we implemented three simple voting strategies: 1) random voting, which assigns a class label randomly to each given sample; 2) majority voting, which assigns the most frequently occurring class label from the subject's labels to each sample; and 3) ratio voting, which assigns a class label randomly based on a probability distribution proportional to the ratio of class labels for the subject. These strategies were used to compare the results with the expected values of analytically determined baseline scores.

Model Validation: Fig. 6 shows the CV scheme used in our study. We develop subject-dependent classification models for both datasets, where one video from a subject at a time is used for testing while the rest are used for training. This splitting is repeated for each video of the subject using a Leave-One-Video-Out (LOVO) CV scheme. Each subject receives an accuracy and a macro-averaged $F1$ score based on the confusion matrix of the subject's predictions. The final performance metrics are the averages of these subject-wise scores, calculated separately for each emotional dimension. This approach complies with the DEAP performance evaluation methodology, and allows for a direct comparison of the results. To assess statistical significance,

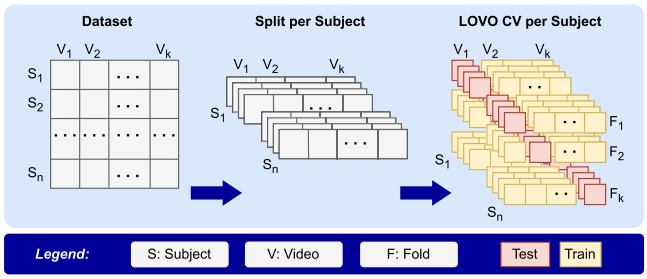


Fig. 6. Data splitting and Cross Validation (CV) scheme of the study. Leave-One-Video-Out (LOVO) CV was applied to every subject.

TABLE V
RESULTS OF REPLICATING SINGLE-TRIAL EMOTION CLASSIFICATION FROM THE ORIGINAL DEAP PAPER [25]

	Setup	Valence		Arousal	
		A	F1	A	F1
	Ref. Val. [25]	0.627	0.608**	0.570	0.533*
	Replication	0.631	0.606**	0.616	0.542*
Baseline	Random	0.500	0.494	0.500	0.483
Baseline	Majority	0.586	0.368	0.644	0.389
Baseline	Ratio	0.525	0.500	0.562	0.500

Numbers are averaged over subjects. (A: Accuracy).

a one-sample t-test is performed, comparing the average $F1$ score to the best value set by the baseline voting strategies.

IV. RESULTS

A. Replication of DEAP Results

As first step of our study, we replicated the single-trial emotion classification pipeline from the DEAP dataset. We used the pre-processed DEAP dataset package and focused on classifying the valence and arousal dimensions using peripheral physiological signals. The list of peripheral signals was identical to those used in the original DEAP paper: BVP, RSP, EDA, SKT, EMG, and EOG. We extracted the same set of features, applied the same classifier (Gaussian NB), used the same feature selection, and the same evaluation methodology.

Table V presents the results of our replication, averaged over subjects. The first row of the table shows the results reported in the original DEAP paper [25], while the second row shows our replication results. Our replication results are consistent with the original DEAP paper, with no significant differences in performance metrics. The slight discrepancies observed in the results can be attributed to the details of the signal processing and feature extraction steps, which are difficult to replicate exactly without access to the original codebase and parameters. We also replicated the baseline voting strategies, with the corresponding values shown in the last three rows of the table. These baseline values are identical to those reported in the original work (see Table VII of Ref. [25]).

B. EmoWear Dataset Results

The EmoWear dataset shares methodological similarities with the DEAP dataset (see Section III-A), enabling us to apply the same emotion classification pipeline. We kept the processing steps as consistent as possible between the two datasets, making only the following modifications: 1) we extracted additional features from the cardio-respiratory signals (see Table III); 2) we added a minimum feature count threshold to the feature selection step (see Section III-C); 3) we tested additional classifiers (SVM and LR); and 4) we examined different combinations of input sets. All above changes were made to enable a more comprehensive analysis on the EmoWear dataset, especially considering the new gateway modality, SCG.

TABLE VI
RESULTS OF SINGLE-TRIAL EMOTION CLASSIFICATION ON THE EMOWEAR DATASET

	Setup			Valence		Arousal	
	Clf.	C-sig.	Peri.	A	F1	A	F1
1	NB	ECG	all	0.576	0.554**	0.603	0.542*
2	NB	BVP	all	0.581	0.558**	0.609	0.547*
3	NB	SCG	all	0.573	0.552**	0.603	0.542*
4	SVM	ECG	all	0.600	0.584***	0.609	0.579***
5	SVM	BVP	all	0.595	0.585***	0.606	0.573**
6	SVM	SCG	all	0.596	0.587***	0.609	0.579***
7	LR	ECG	all	0.597	0.589***	0.608	0.577***
8	LR	BVP	all	0.600	0.591***	0.609	0.575***
9	LR	SCG	all	0.597	0.588***	0.608	0.577***
10	NB	ECG	RSP	0.576	0.554**	0.617	0.560**
11	NB	BVP	RSP	0.581	0.558**	0.611	0.555*
12	NB	SCG	RSP	0.581	0.558**	0.625	0.570***
13	SVM	ECG	RSP	0.600	0.584***	0.605	0.574**
14	SVM	BVP	RSP	0.595	0.585***	0.605	0.567**
15	SVM	SCG	RSP	0.595	0.585***	0.612	0.583***
16	LR	ECG	RSP	0.597	0.589***	0.605	0.571**
17	LR	BVP	RSP	0.600	0.591***	0.605	0.569**
18	LR	SCG	RSP	0.600	0.591***	0.611	0.579***
19	NB	SCG	ADR	0.543	0.522	0.587	0.520
20	SVM	SCG	ADR	0.572	0.561***	0.578	0.550***
21	LR	SCG	ADR	0.575	0.564***	0.586	0.555***
22	Baseline	Random		0.500	0.494	0.500	0.483
23	Baseline	Majority		0.580	0.366	0.636	0.386
24	Baseline	Ratio		0.521	0.500	0.560	0.500

$F1$-scores are macro-averaged over the high and low classes of the emotional dimensions. Significance levels are determined using a one-sample t-test, comparing the $F1$-score distribution to the best value among the baseline voting strategies (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.)
Numbers are averaged over subjects. (Clf.: Classifier, C-sig.: Cardiac signal (ECG, BVP, SCG), Peri.: Peripheral signals (RSP, ADR, EDA, SKT), A: Accuracy).

Table VI presents the results of single-trial emotion classification on the EmoWear dataset, averaged over subjects. The investigated setups can be described as different selections of cardiac signals (ECG, BVP, or SCG), combined with different selections of peripheral signals (either all available signals, or only one of the respiratory signals, RSP or ADR), classified using different models (NB, SVM, or LR). We analyze and discuss these results in the following section.

V. DISCUSSION

A. Methodological Approach

We based the methodology used in this study, first and most on avoiding common pitfalls, and second on the best practices learned from the literature. We considered both the limitations and strengths of the EmoWear dataset, such as its limited size, and the availability of the new modality, SCG. Based on the lessons learned from the reviewed studies, we ensured: 1) The use of a cross-validation scheme in our data splitting, which is crucial for small datasets to avoid overfitting and to address

the generalization of the models over the dataset, 2) The use of $F1$ score as the primary evaluation metrics, which is more robust to class imbalance than accuracy, 3) Averaging the $F1$ scores over both classes, which otherwise risks overestimating the performances by focusing on the $F1$ score of the best performing class, 4) The use of a one-sample *t*-test to assess the statistical significance of the results against some baseline voting strategies, making sure that the results are not due to random chance, 5) Transparency in reporting the methodology and results, which is crucial for reproducibility of the results, and 6) Validating our implementation of the full processing, learning, and evaluation pipeline by replicating the DEAP study results and ensuring consistency.

The consistency of the above methodological approach with the single-trial emotion classification pipeline of the DEAP dataset, allowed us to replicate their work. Our replication results revealed no significant differences in performance (see Section IV-A). The consistency of our replication results with the results reported in DEAP study highlights two key points: 1) It serves as a stand-alone contribution by reaffirming their findings, and 2) It ensures that our implementation of the pipeline is reliable, setting the way for the EmoWear dataset analysis, which we did next.

When comparing with existing accelerometer emotion recognition research, we note that this study is the first to analyze accelerometer data as a mechanical surrogate of cardiac activity (SCG) for ER. The use of accelerometer data for ER was previously limited to other motion characterization methods, such as blind motion measurement [80], [81], [82] or gate-analysis [35]. Using SCG provides a more direct measurement of the heart's mechanical activity, which is less affected by motion artifacts than other accelerometer-based methods.

B. Interpretation of Findings

Comparison to DEAP: In Table VI, the second row presents the classification results using the NB classifier when BVP is used as the cardiac signal, and all other peripheral signals present in the EmoWear dataset are used. As such, this setup is most comparable to the DEAP dataset, which also used BVP as the cardiac signal, along with all other peripheral signals, and the NB classifier. Results in this row show applicability of the DEAP setup on the EmoWear dataset, with comparable performance levels. The level of significance observed in these results are consistent with the reported level of significance in the DEAP study. It is important to note that we do not expect the results to be identical, as the datasets differ, most importantly the subjects are totally different. However, the applicability of the DEAP setup to the EmoWear dataset, achieving similar results is a strong indicator, on how well such a setup can generalize across the two datasets.

Applicability of SCG for Emotion Recognition: The results of primary interest are those related to the SCG signal, which has never been investigated for ER before. We discuss the outcomes of the SCG-based setups in three contexts: 1) when combined with all other peripheral signals (RSP, ADR, EDA, and SKT); 2) when combined with only the respiratory signal from a dedicated sensor (RSP); and 3) when combined with only the respiratory signal from the same sensor that provides the SCG (ADR). Fig. 7 shows distinct boxplots per the three mentioned contexts on the SCG column. It also allows for comparing the SCG results with the other cardiac signals, ECG and BVP columns, in the same contexts.

Considering the SCG in combination with 'all' peripheral signals, we observe no systematic trend in the differences of the average results between the SCG and the other cardiac signals. The SCG signal does not outperform the other cardiac signals, but it also does not underperform. The distribution is obviously different, but so is the distribution of the other two cardiac signals compared to each other. More or less, similar observations can be made when the SCG is combined with the RSP signal.

As outlined in the introduction, ER research relies on mapping objective physiological signals to subjective affective states. Our findings demonstrate that SCG, aligns well with this framework, with similar performance to the other cardiac signals. This is an important finding, as the SCG signal can be obtained from a simple chest-worn accelerometer, which has the potential to provide a lot more information than just the cardiac signals [21].

The most interesting insights are observed when the SCG signal is combined with the ADR signal, as this combination uses only a single chest-worn accelerometer to extract both signals. The boxplots of the 'SCG + ADR' combination in Fig. 7 show an evident drop of performance comparing to their adjacent boxplots, meaning that the ADR signal has not been able to reach the same level of performance as the RSP / all signals. Fig. 8 provides a more detailed view on the cardio-respiratory-based combination results. Fig. 8(a) confirms that the average $F1$ scores of the 'SCG + ADR' combination is comparable to the 'ECG / BVP + RSP' combinations, but slightly lower. We do not see this performance drop when the SCG is combined with the RSP signal. Therefore, the slight drop in performance of the 'SCG + ADR' combination can be attributed to the fact that the ADR signal tends to be noisier and less specific than a dedicated respiratory belt (RSP), limiting the signal fidelity. Fig. 8(b) shows the Pearson correlation heatmap of the $F1$ scores obtained across the four cardio-respiratory signal combinations. The heatmap suggests that the 'SCG + ADR' combination performs differently compared to the other signal combinations, within subjects. Noteworthily, the different distribution of the 'SCG + ADR' combination, does not mean that the combination is not effective. The *t*-test results in Table VI reveal otherwise: while the NB classifier fails to achieve significant results, the more robust SVM and LR classifiers demonstrate $F1$ score distributions that are significantly higher than best of the baselines ($p < 0.001$; see Table VI, rows 20–21). These findings confirm the applicability of a single chest-worn accelerometer for ER, which is a significant step towards embedding ER in everyday life scenarios.

Classifiers and Features: The three classifiers used in our study show similar performance trends across the different setups. However, the NB classifier shows lower performance compared to the SVM and LR classifiers. In Table VI, the NB classifier presents the least significant results most frequently. In the worst case, the NB classifier failed to achieve significant

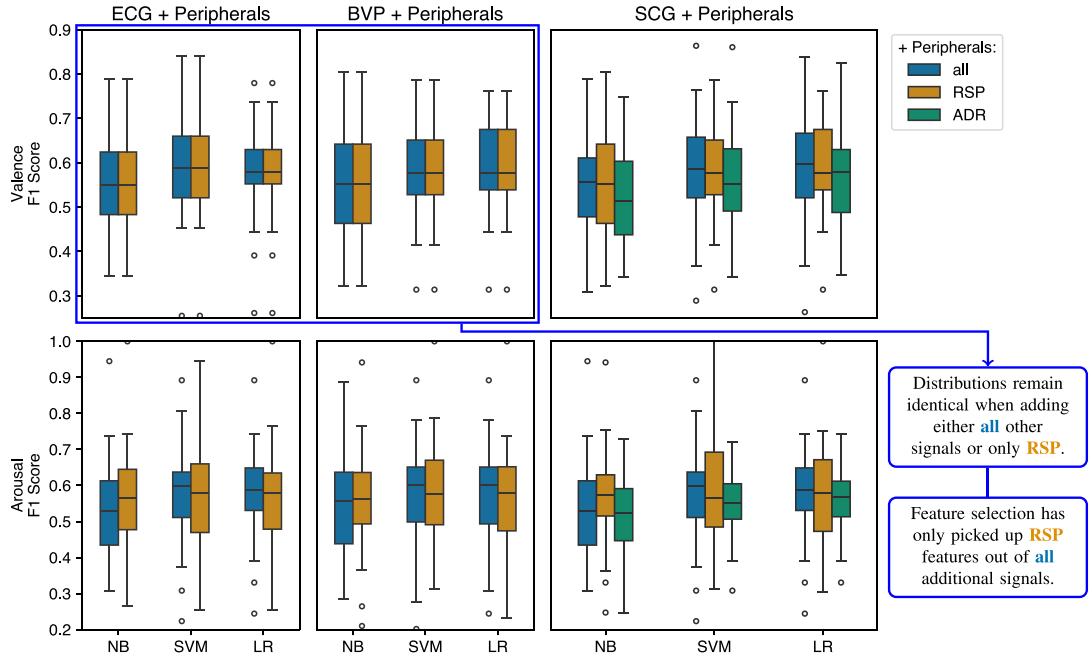


Fig. 7. Boxplots of the $F1$ scores over subjects per classifier type, categorized by the emotional dimensions, and the cardiac sources. Each boxplot corresponds to a cardiac signal that is combined with 1) all other peripheral signals, 2) only the RSP signal, or 3) only the ADR signal.

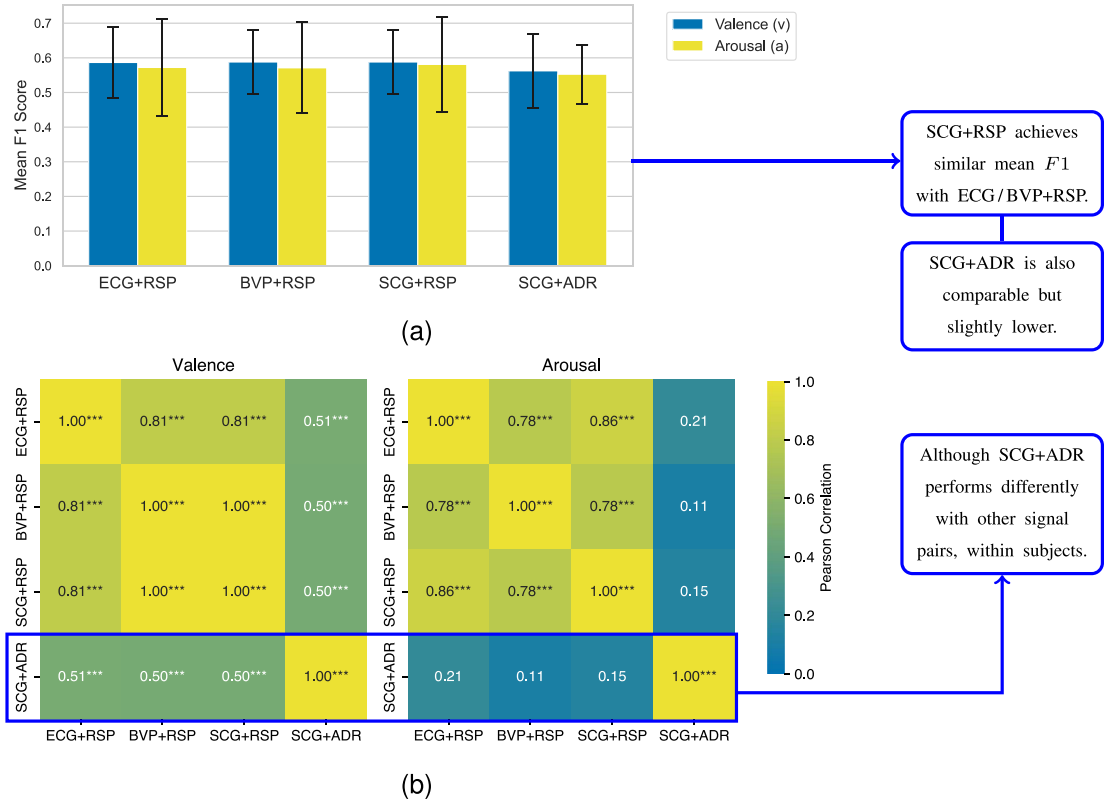


Fig. 8. Comparing cardio-respiratory combinations on F1-score results. (a) Averaged over subjects and classifiers (SVM and LR only) with error bars reflecting standard deviations. (b) Pearson correlation heatmap of F1 results obtained (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.).

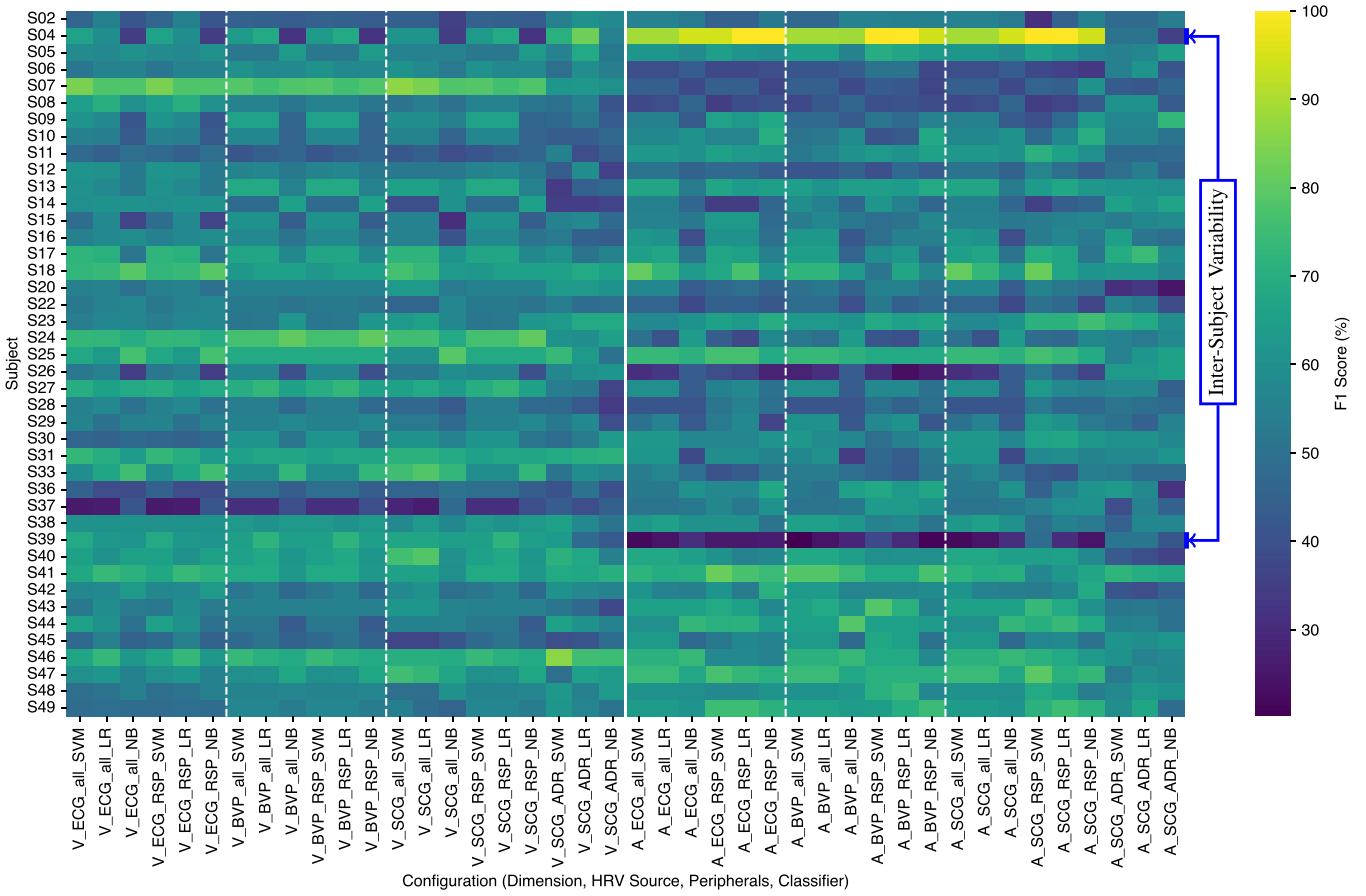


Fig. 9. Heatmap of all F1 scores per subject, for all configurations of emotional dimension, HRV source, peripherals, and classifiers. (V: Valence, A: Arousal.).

results for the 'SCG + ADR' combination (row 19), while the SVM and LR classifiers achieved significant results ($p < 0.001$). Similarly, the boxplots of Fig. 7 suggest consistent results across the SVM and LR classifiers, but slightly lower performance for the NB classifier. Weaker performance of the NB classifier was expected, as it comes with strong independence assumptions between features, which do not hold for our extracted features.

Another evident observation in Fig. 7 is the identical distribution of $F1$ scores achieved for valence classification of the 'ECG/BVP + all' combinations versus the 'ECG/BVP + RSP' combinations. We checked and discovered that our feature selection procedure had only picked up the RSP features even when all other peripherals where present. Such selection was not unexpected, as SKT contributed minimally to the feature vector (Table III), and the literature identifies EDA, the other available alternative, as rather being a primary indicator of emotional arousal [83].

Inter-Subject Variability: The variability of the obtained results between subjects is evident in the wide boxplots of Fig. 7. Fig. 9 provides a heatmap of all $F1$ scores obtained in the study. The clear horizontal patterns in the heatmap indicate that performance remains fairly consistent across different setups but varies significantly across subjects. Some subjects consistently achieve better or worse classification results than others, regardless of the setup (although, this pattern is not always observed for the 'SCG + ADR' setups.) The variability is more pronounced in the classification of the valence dimension for the subjects 7, and 37, and in the arousal dimension for the subjects 2, 26, and 39. This inter-subject variability reaffirms that emotional experience is inherently subjective and can be expressed differently by different individuals [84]. We also hope that the detailed heatmap of $F1$ scores in Fig. 9 will serve as a baseline for future studies on the EmoWear dataset, providing further insights into ER performance across subjects.

C. Limitations

Comparison to Related Work: In Section II, we reviewed related work that provide context for our study; however, several limitations prevent direct comparisons between our results and those of the reviewed studies. First, this study is the first to investigate ER using the EmoWear dataset, and the first to explore the potential of the SCG signal for this purpose. Second, the use of incompatible performance metrics and validation procedures in our studies complicate any direct comparison [6]. Finally, methodological gaps that were identified in Section II may have led to overly optimistic results in some of the reviewed studies, which we avoided in this study. However, following the single-trial emotion classification pipeline of the DEAP

study [25] allowed us to replicate their results and compare our findings. Moreover, our use of baseline voting strategies provided a reference point for our results, and statistical significance tests confirmed the validity of our findings.

Dataset Limitations: EmoWear and DEAP datasets have inherent limitations that should be considered when interpreting the findings. First, the demographic composition of the datasets may not fully represent the diversity of real-world users. Factors such as age, gender, and physiological differences may influence the physiological signals used for ER. While the datasets include both male and female participants, a more balanced and heterogeneous sample would be beneficial for ensuring broader applicability. Second, the experimental setting of the datasets may not entirely reflect real-world usage scenarios. Although the EmoWear dataset was designed to mimic practical conditions, controlled laboratory settings inherently differ from everyday environments where wearable ER systems would be deployed. The impact of movement artifacts, sensor positioning variations, and spontaneous emotional expressions remains an area for further investigation. Third, both datasets are relatively small, which limits the complexity of the models that can be trained. Deep learning models require large datasets to avoid overfitting which we address in the following part.

Model Complexity and Overfitting: As previously mentioned in Section III-C, our selection of the classifiers and their hyperparameters was based on an initial investigation of different classifiers and hyperparameters, aiming to maximize the macro-averaged $F1$ score over the high and low classes of the emotional dimensions in the entire EmoWear dataset. We experimented with CNN and LSTM models, but they failed to classify emotions effectively, consistent with our expectations based on the dataset size. Performing Leave-One-Video-Out cross-validation on a subject-dependent ER task, would leave us with only 37 training samples in each fold which is a very small dataset for deep learning models. On average over the folds, these models predicted the train data at about 98% $F1$, while the test data was predicted at a range of about 40-45%. Although these models could potentially outperform classic classifiers if trained on a larger dataset, we observed significant overfitting on the training data, even after applying various regularization techniques such as dropout, early stopping, L2 norm regularization, and batch normalization. We identify a potential direction for future research to address the dataset size limitation, which we discuss later in this section.

Subjectivity of Emotions: Emotions are complex and multifaceted, and can be influenced by a wide range of factors, including cultural background, personal experiences, and individual differences [84]. The inherent complexity and subjectivity of emotions is a major challenge in ER, especially in providing ground truth labels for training and testing the models [6]. A manifestation of these issues are reflected in the variability over subjects observed in our results, and discussed earlier.

D. Future Directions

Outcomes and limitations seen in this study suggest several potential directions for future research. These directions include but are not limited to: 1) Exploring the potential of the SCG signal for ER in more detail, including the development of dedicated feature extraction methods and models; 2) Investigating and improving signal processing techniques for more robust information inference from the SCG and ADR signals; 3) Examining the potential of lightweight deep learning models (e.g., MobileNet, TinyLSTM) for ER on the EmoWear dataset, in addition to the explored traditional classifiers; 4) Exploring deep learning approaches like CNNs, LSTMs, or transformers to process raw SCG signals directly, circumventing limitations of manual feature engineering; 5) Exploring transfer learning or self-supervised pre-training on large-scale physiological datasets to overcome the challenge of limited dataset size; 6) Investigating subject-independent ER problem using stronger models, capable of learning more complex patterns; 7) Uncovering the potential of chest-worn accelerometers for ER in everyday life scenarios, including the development of real-time ER systems; 8) Exploring the potential of the EmoWear dataset for other applications, such as voice activity detection, drinking detection, and activity recognition, which are also included in the dataset; 9) Expanding the existing dataset by collecting more data from a larger and more diverse group of participants, to improve the generalizability of the findings; and 10) Developing a chest-worn smartphone application that can be used by participants in their daily lives, to collect more data in an opportunistic manner.

VI. Conclusion

In this study, we explored the potential of SCG for emotion recognition. We replicated the single-trial emotion classification pipeline of the DEAP study, applying it on both the DEAP and EmoWear datasets. Replication results showed consistency with the original DEAP study, both confirming their findings and validating our implementation of the pipeline. Applying the same validated pipeline to the EmoWear dataset revealed that SCG can be used for emotion recognition, achieving similar performance to the ECG and BVP cardiac modalities. Moreover, we achieved significant results using only the combination of SCG with the ADR, both extracted from the ACC signal. 'SCG + ADR' combination results confirmed that a single chest-worn accelerometer can be effectively used as a physiological gateway for emotion recognition. Our results serve as the first benchmarks for future research on the EmoWear dataset. Employing deep learning models was limited by the small size of dataset; however, future work could explore transfer learning or self-supervised pre-training on larger datasets. Expanding research on SCG for emotion recognition can bring affective computing technology into everyday life.

References

[1] T. Wang et al., "Emotion recognition from full-body motion using multiscale spatio-temporal network," *IEEE Trans. Affect. Comput.*, vol. 15, no. 3, pp. 898–912, Third Quarter 2024.

[2] M. A. Hasnul et al., "Electrocardiogram-based emotion recognition systems and their applications in healthcare–A review," *Sensors*, vol. 21, Jul. 2021, Art. no. 5015.

[3] P. D. Paraschos and D. E. Koulouriotis, "Game difficulty adaptation and experience personalization: A literature review," *Int. J. Hum.–Comput. Interaction*, vol. 39, no. 1, pp. 1–22, 2023.

[4] H. Farman et al., "Facial emotion recognition in smart education systems: A review," in *Proc. 2023 IEEE Int. Smart Cities Conf.*, 2023, pp. 1–9.

[5] S. Renjith, A. Sreekumar, and M. Jathavedan, "An extensive study on the evolution of context-aware personalized travel recommender systems," *Inf. Process. Manage.*, vol. 57, Jan. 2020, Art. no. 102078.

[6] S. Saganowski, B. Perz, A. G. Polak, and P. Kazienko, "Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 1876–1897, Third Quarter 2023.

[7] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biol. Psychol.*, vol. 84, pp. 394–421, 2010.

[8] K. Wac and C. Tsiourti, "Ambulatory assessment of affect: Survey of sensor systems for monitoring of autonomic nervous systems activation in emotion," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 251–272, Third Quarter 2014.

[9] P. Schmidt et al., "Wearable-based affect recognition–A review," *Sensors*, vol. 19, Sep. 2019, Art. no. 4079.

[10] Y. Wang et al., "A systematic review on affective computing: Emotion models, databases, and recent advances," *Inf. Fusion*, vol. 83/84, pp. 19–52, Jul. 2022.

[11] P. Ekman and H. Oster, "Facial expressions of emotion," *Annu. Rev. Psychol.*, vol. 30, pp. 527–554, 1979.

[12] J. A. Russell, "A circumplex model of affect," *J. Pers. Social Psychol.*, vol. 39, pp. 1161–1178, 1980.

[13] J. Martin, "The limbic system and cerebral circuits for reward, emotions, and memory," in *Neuroanatomy Text and Atlas*, 4th ed. New York, NY, USA: McGraw-Hill Publishing, 2012, pp. 385–413.

[14] B. M. Appelhans and L. J. Luecken, "Heart rate variability as an index of regulated emotional responding," *Rev. Gen. Psychol.*, vol. 10, pp. 229–240, Sep. 2006.

[15] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry*, vol. 25, pp. 49–59, 1994.

[16] Y.-L. Hsu, J.-S. Wang, W.-C. Chiang, and C.-H. Hung, "Automatic ECG-Based emotion recognition in music listening," *IEEE Trans. Affect. Comput.*, vol. 11, no. 1, pp. 85–99, First Quarter 2020.

[17] L. A. Bugnon, R. A. Calvo, and D. H. Milone, "Dimensional affect recognition from HRV: An approach based on supervised SOM and ELM," *IEEE Trans. Affect. Comput.*, vol. 11, no. 1, pp. 32–44, First Quarter 2020.

[18] M. H. Rahmani et al., "EmoWear: Wearable physiological and motion dataset for emotion recognition and context awareness," *Sci. Data*, vol. 11, pp. 1–18, Jun. 2024.

[19] O. T. Inan et al., "Ballistocardiography and seismocardiography: A review of recent advances," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 4, pp. 1414–1427, Jul. 2015.

[20] D. Rai et al., "A comprehensive review on seismocardiogram: Current advancements on acquisition, annotation, and applications," *Mathematics*, vol. 9, no. 18, 2021, Art. no. 2243.

[21] M. H. Rahmani, R. Berkvens, and M. Weyn, "Chest-worn inertial sensors: A survey of applications and methods," *Sensors*, vol. 21, Apr. 2021, Art. no. 2875.

[22] A. Bricout, J. Fontecave-Jallon, D. Colas, G. Gerard, J.-L. Pépin, and P.-Y. Guméry, "Adaptive accelerometry derived respiration: Comparison with respiratory inductance plethysmography during sleep," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2019, pp. 6714–6717.

[23] H. D. Critchley and Y. Nagai, "How emotions are shaped by bodily states," *Emotion Rev.*, vol. 4, no. 2, pp. 163–168, 2012.

[24] R. W. Levenson, "Autonomic nervous system differences among emotions," *Psychol. Sci.*, vol. 3, no. 1, pp. 23–27, 1992.

[25] S. Koelstra et al., "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, First Quarter 2012.

[26] S. Saganowski et al., "Emognition dataset: Emotion recognition with self-reports, facial expressions, and physiology using wearables," *Sci. Data*, vol. 9, pp. 1–11, Apr. 2022.

[27] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, First Quarter 2012.

[28] K. Sharma et al., "A dataset of continuous affect annotations and physiological signals for emotion analysis," *Sci. Data*, vol. 6, pp. 1–13, Oct. 2019.

[29] M. Behnke et al., "Psychophysiology of positive and negative emotions, dataset of 1157 cases and 8 biosignals," *Sci. Data*, vol. 9, pp. 1–15, Jan. 2022.

[30] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-Based multimodal database for decoding affective physiological responses," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 209–222, Third Quarter 2015.

[31] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 147–160, Second Quarter 2018.

[32] K. Kutt et al., "BIRAFFE: Bio-reactions and faces for emotion-based personalization," in *Proc. Workshop Affect. Comput. Context Awareness Ambient Intell.*, 2019, pp. 1–17.

[33] S. Kang et al., "K-EmoPhone: A mobile and wearable dataset with in-situ emotion, stress, and attention labels," *Sci. Data*, vol. 10, no. 1, 2023, Art. no. 351.

[34] P. Schmidt et al., "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proc. 20th ACM Int. Conf. Multimodal Interaction*, 2018, pp. 400–408.

[35] M. A. Hashmi, Q. Riaz, M. Zeeshan, M. Shahzad, and M. M. Fraz, "Motion reveal emotions: Identifying emotions from human walk using chest mounted smartphone," *IEEE Sensors J.*, vol. 20, no. 22, pp. 13511–13522, Nov. 2020.

[36] 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.

[37] 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.

[38] D. Dresvyanskiy et al., "End-to-end modeling and transfer learning for audiovisual emotion recognition in-the-wild," *Multimodal Technol. Interaction*, vol. 6, Jan. 2022, Art. no. 11.

[39] C. T. Yen and K. H. Li, "Discussions of different deep transfer learning models for emotion recognitions," *IEEE Access*, vol. 10, pp. 102860–102875, 2022.

[40] J. Zhang et al., "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Inf. Fusion*, vol. 59, pp. 103–126, Jul. 2020.

[41] H. Z. Wijasena, R. Ferdiana, and S. Wibirama, "A survey of emotion recognition using physiological signal in wearable devices," in *Proc. Int. Conf. Artif. Intell. Mechatron. Syst.*, 2021, pp. 1–6.

[42] C. Godin et al., "Selection of the most relevant physiological features for classifying emotion," in *Proc. 2nd Int. Conf. Physiol. Comput. Syst.*, 2015, pp. 17–25.

[43] S. Chen, Z. Gao, and S. Wang, "Emotion recognition from peripheral physiological signals enhanced by EEG," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 2827–2831.

[44] Q. Zhang et al., "Respiration-based emotion recognition with deep learning," *Comput. Ind.*, vol. 92/93, pp. 84–90, Nov. 2017.

[45] Z. Yin et al., "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model," *Comput. Methods Programs Biomed.*, vol. 140, pp. 93–110, Mar. 2017.

[46] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion based music recommendation system using wearable physiological sensors," *IEEE Trans. Consum. Electron.*, vol. 64, no. 2, pp. 196–203, May 2018.

[47] E. J. Choi and D. K. Kim, "Arousal and valence classification model based on long short-term memory and DEAP data for mental healthcare management," *Healthcare Informat. Res.*, vol. 24, pp. 309–316, Oct. 2018.

[48] M. S. Lee et al., "Fast emotion recognition based on single pulse PPG signal with convolutional neural network," *Appl. Sci.*, vol. 9, Aug. 2019, Art. no. 3355.

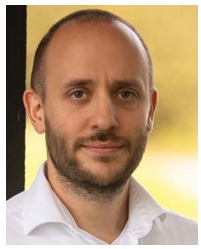
[49] M. S. Lee et al., "Emotion recognition using convolutional neural network with selected statistical photoplethysmogram features," *Appl. Sci.*, vol. 10, May 2020, Art. no. 3501.

[50] X. Wu et al., "Multimodal physiological signal emotion recognition based on convolutional recurrent neural network," in *Proc. IOP Conf. Series: Mater. Sci. Eng.*, 2020, Art. no. 032005.

[51] Q. Zhu, G. Lu, and J. Yan, "Valence-arousal model based emotion recognition using EEG, peripheral physiological signals and facial expression," in *Proc. ACM Int. Conf. Mach. Learn. Soft Comput.*, 2020, pp. 81–85.

[52] R. Elalamy, M. Fanourakis, and G. Chanel, "Multi-modal emotion recognition using recurrence plots and transfer learning on physiological signals," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interaction*, 2021, pp. 1–7.
[53] D. H. Kang and D. H. Kim, "1D convolutional autoencoder-based PPG and GSR signals for real-time emotion classification," *IEEE Access*, vol. 10, pp. 91332–91345, 2022.
[54] M. Pidgeon et al., "End-to-end emotion recognition using peripheral physiological signals," in *Proc. 35th Brit. HCI Conf. Towards Hum.-Centred Digit. Soc.*, 2022, pp. 1–10.
[55] S. T. P. Jung and T. J. Sejnowski, "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 96–107, First Quarter 2022.
[56] B Shubha, N Poornima, V. M. Gowda, U. Sushma, Y. R. Meghana, and T. S. Bhoomika, "Emotion recognition based on PPG and GSR signals using DEAP dataset," in *Proc. 2023 Int. Conf. Netw. Multimedia Inf. Technol.*, 2023, pp. 1–7.
[57] J. Gohumpu, M. Xue, and Y. Bao, "Emotion recognition with multi-modal peripheral physiological signals," *Front. Comput. Sci.*, vol. 5, Dec. 2023, Art. no. 1264713.
[58] M. F. Bamonte, M. Risk, and V. Herrero, "Emotion recognition based on galvanic skin response and photoplethysmography signals using artificial intelligence algorithms," in *Proc. Adv. Bioeng. Clin. Eng.*, Springer Nature, Switzerland, 2024, pp. 23–35.
[59] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology*. Cambridge, MA, USA: MIT Press, 1974.
[60] G. Canbek, T. T. Temizel, and S. Sagiroglu, "PToPI: A comprehensive review, analysis, and knowledge representation of binary classification performance measures/metrics," *SN Comput. Sci.*, vol. 4, pp. 1–30, Jan. 2023.
[61] L. A. Jeni, J. F. Cohn, and F. D. L. Torre, "Facing imbalanced data - Recommendations for the use of performance metrics," in *Proc. 2013 Humaine Assoc. Conf. Affect. Comput. Intell. Interaction*, 2013, pp. 245–251.
[62] A. Field and G. Hole, *How to Design and Report Experiments*. Newcastle upon Tyne, U.K.: SAGE, 2002.
[63] Y. Xu and R. Goodacre, "On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning," *J. Anal. Testing*, vol. 2, pp. 249–262, Jul. 2018.
[64] R. A. Poldrack et al., "Scanning the horizon: Towards transparent and reproducible neuroimaging research," *Nature Rev. Neurosci.*, vol. 18, no. 2, pp. 115–126, Jan. 2017.
[65] M. H. Rahmani, M. Symons, R. Berkvens, and M. Weyn, "EmoWear data," Dec. 2023.
[66] Presentation software. 2003. [Online]. Available: https://www.neurobs.com
[67] M. H. Rahmani, R. Berkvens, and M. Weyn, "ColEmo: A flexible open source software interface for collecting emotion data," in *Proc. 11th Int. Conf. Affect. Comput. Intell. Interaction Workshops Demos*, 2023, pp. 1–8.
[68] C. R. Harris et al., "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, Sep. 2020.
[69] T. pandas development team, "pandas-dev/pandas: Pandas," Apr. 2024, doi: 10.5281/zenodo.10957263.
[70] W. McKinney, "Data structures for statistical computing in python," in *Proc. 9th Python Sci. Conf.*, 2010, pp. 56–61.
[71] D. Makowski et al., "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behav. Res. Methods*, vol. 53, pp. 1689–1696, Feb. 2021.
[72] J. D. Pollock and A. N. Makaryus, *Physiology, Cardiac Cycle*. Treasure Island, FL, USA: StatPearls Publishing, 2025.
[73] P. Dehkordi et al., "Comparison of different methods for estimating cardiac timings: A comprehensive multimodal echocardiography investigation," *Front. Physiol.*, vol. 10, Aug. 2019, Art. no. 452771.
[74] E. Vaini, P. Lombardi, and M. D. Rienzo, "Aortic-finger pulse transit time vs. R-derived pulse arrival time: A beat-to-beat assessment," *Comput. Cardiol.*, vol. 42, pp. 253–256, Feb. 2015.
[75] C. Massaroni et al., "Heart rate and heart rate variability indexes estimated by mechanical signals from a skin-interfaced IMU," in *Proc. 2022 IEEE Int. Workshop Metrol. Ind. 4.0 IoT*, 2022, pp. 322–327.
[76] S. Lapi et al., "Respiratory rate assessments using a dual-accelerometer device," *Respir. Physiol. Neurobiol.*, vol. 191, pp. 60–66, Jan. 2014.
[77] C. Carreiras et al., "BioSPPy: Biosignal processing in python," 2015. [Online]. Available: https://github.com/PIA-Group/BioSPPy/
[78] M. Elgendi, M. Jonkman, and F. D. Boer, "Frequency bands effects on QRS detection," in *Proc. 3rd Int. Conf. Bio-inpsired Syst. Signal Process.*, V. Mahadevan and J. Zhou, Eds., 2010, vol. 1, pp. 428–431.
[79] D. Makowski, "Neurophysiological data analysis with NeuroKit2," 2021. [Online]. Available: https://neuropsychology.github.io/NeuroKit/
[80] J. C. Quiroz, E. Geangu, and M. H. Yong, "Emotion recognition using smart watch sensor data: Mixed-design study," *JMIR Ment. Health*, vol. 5, no. 3, 2018, Art. no. e10153.
[81] P. Schmidt, R. Dürichen, A. Reiss, K. Van Laerhoven, and T. Plötz, "Multi-target affect detection in the wild: An exploratory study," in *Proc. 2019 ACM Int. Symp. Wearable Comput.*, 2019, pp. 211–219.
[82] O. Piskioulis, K. Tzafilkou, and A. Economides, "Emotion detection through smartphone's accelerometer and gyroscope sensors," in *Proc. 29th ACM Conf. User Model. Adapt. Personalization*, New York, NY, USA, 2021, pp. 130–137.
[83] D. Caruelle et al., "The use of electrodermal activity (EDA) measurement to understand consumer emotions–A literature review and a call for action," *J. Bus. Res.*, vol. 104, pp. 146–160, Nov. 2019.
[84] J. E. LeDoux and S. G. Hofmann, "The subjective experience of emotion: A fearful view," *Curr. Opin. Behav. Sci.*, vol. 19, pp. 67–72, Feb. 2018.

Mohammad Hasan Rahmani received the BSc degree in electrical engineering from the K.N. Toosi University of Technology, and the MSc degree in biomedical engineering from the Amirkabir University of Technology (Tehran Polytechnic), both in Tehran, Iran. He is currently working toward the PhD degree in applied engineering with the University of Antwerp, Belgium, focusing on biomedical signal processing and affective computing. As the Principal architect of the ColEmo interface, he developed an open-source tool for collecting emotion data, which was instrumental in creating the EmoWear dataset–a comprehensive resource for emotion recognition and context awareness through wearable sensors. He contributed to the NudJIT project, focusing on adaptive interventions to promote healthy behavior through just-in-time nudges. His research interests include wearable technology, sensing, machine learning, robotics and the integration of physiological signals for enhanced human-computer interaction.



Rafael Berkvens (Member, IEEE) received the master's and PhD degrees from IDLab-imec, Department of Electronics, Information and Communication Technology, Faculty of Applied Engineering, University of Antwerp. He is an assistant professor with IDLab-imec, Department of Electronics, Information and Communication Technology, Faculty of Applied Engineering, University of Antwerp. His research focuses on understanding the information available about individuals and their environment through the study of wireless communication signals. During his PhD, he utilized RatSLAM—a computational model of a rodent's brain—replacing camera input with Wi-Fi signals to explore spatial perception. He has developed the concept of opportunistic sensing, aiming to perceive the world using radio frequency transmitters as if they were light sources. He is currently advancing fundamental aspects of this research, including reception, identification, localization, and prediction, with a major focus on Integrated Sensing and Communication (ISAC) and 6G technologies. He has published more than 90 journal articles and conference proceedings in his field and is a member of the IEEE Communications Society.

Maarten Weyn (Member, IEEE) is a full professor and vice-rector of Research and Impact with the University of Antwerp. He teaches wireless communication system. His research at imec-IDLab focuses on ultra-low power sensor communication, embedded systems, sub-1 GHz communication, sensor processing, and localization. He co-founded spin-offs Aloxy, CrowdScan, IoSa, and AtSharp, and contributed to 1OK and Viloc.