IM-TWIN: from Intrinsic Motivations to Transitional Wearable INtelligent companions for autism spectrum disorder

a European funded project

Personalised affect classification and feedback

Deliverable 3.2

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952095.

Project duration 36 months (November 2020, October 2023). Consortium: Consiglio Nazionale dele Ricerche (ITA), Universiteit Utrecht (NLD), Centre de Recherches Interdisciplinaires (FRA), Università degli Studi di Roma La Sapienza (ITA), Plux-Wireless Biosignals S.A. (PRT).

Deliverable data

Work Package:	2 Affective signal processing through the integration of multiple sources
Work Package leader:	CNR
Deliverable beneficiary:	UU
Dissemination level:	Public
Due date:	31th May 2023 (Month 31)
Type:	Report
Authors:	Lukas P.A. Arts, E.L. van den Broek

Acronyms of partners

CNR-ISTC	Consiglio Nazionale delle Ricerche, Istituto di Scienze e Tecnologie della Cognizione (Italy)
UU	Universiteit Utrecht (The Netherlands)
CRI	Centre de Recherches Interdisciplinaires (France)
LA SAPIENZA	Università degli Studi di Roma La Sapienza (Italy)
PLUX	Plux - Wireless Biosignals S.A. (Portugal)

Table of contents

1. Overview of the deliverable

This deliverable contains the final version of the personalised affect classification pipeline based on the biosignals recorded by the IM-TWIN T-Shirt. The core of the deliverable is sectioned as follows:

2. *Data,* containing the subsections
   - 2.1. *Recordings*, describing the characteristics of the data acquired by CNR, Sapienza, and CRI and the processing of the annotation data
   - 2.2. *ECG features*, describing the ECG feature extraction and aggregation using dynamic and static segmentation of the recordings
   - 2.3. *EDA features*, describing the EDA feature extraction and aggregation using dynamic and static segmentation of the recordings
3. *Statistical analysis,* containing the subsections
   - 3.1. *MANOVA*, describing the omnibus analysis and all assumption checks performed to assess the dataset's learnability
   - *3.2. Follow-up analysis,* describing the PCA and LDA analyses performed to inspect the interplay between the significant features found in 3.1
   - 3.3. *Classification*, describing the validation of a LDA-based classifier using a challenging real-world validation scheme.

2. Data

2.1 Recordings

Data was collected by two institutes: i) CNR & Sapienza and ii) CRI. CNR and Sapienza conducted experiments on 4 autistic children and CRI collected and annotated 12 records from 10 typically developed (TD) children. The experiments with the autistic children did not follow a strict protocol. Instead, the children were instructed to play with the PlusMe toy and the focus of these experiments were largely on signal quality and comfortability of the T-Shirt. UU did a initial analysis of signal quality using the specially designed Signal Quality Indicator (SQI) (see Report D2.2). The results are shown in Figure 1a.

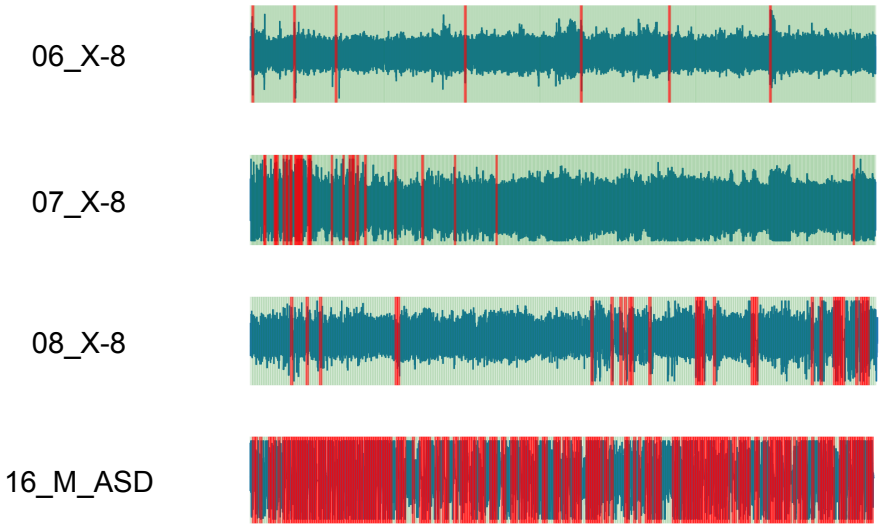

Figure 1a: Visual representation of signal quality of four T-Shirt recordings from autistic children playing with the PlusMe toy. The children were instructed to play with the PlusMe toy. As such, the records do not have annotations of emotional state. However, it does show the potential of the IM-TWIN system to capture high-quality biosignals in ambulatory situations.

On the other hand, CRI employed a study specifically designed to evoke positive, negative, and low-arousal emotional states in the typically developed children. In this report, we will focus on these recordings as these contain annotations of emotional state. Nevertheless, the four recordings with autistic children really shows IM-TWIN's potential of capturing high-quality biosignals from playing children.

Figure 1b visually presents the annotations, highlighting regions in each recording corresponding to the various emotional states. Table 1 provides a breakdown of the duration for each type of annotation per recording. Both Figure 1b and Table 1 make it apparent that the dataset is unbalanced. Specifically, the experimentally challenging negative arousal state is underrepresented, being annotated for a total of only 9 minutes—seven times less than the baseline state.

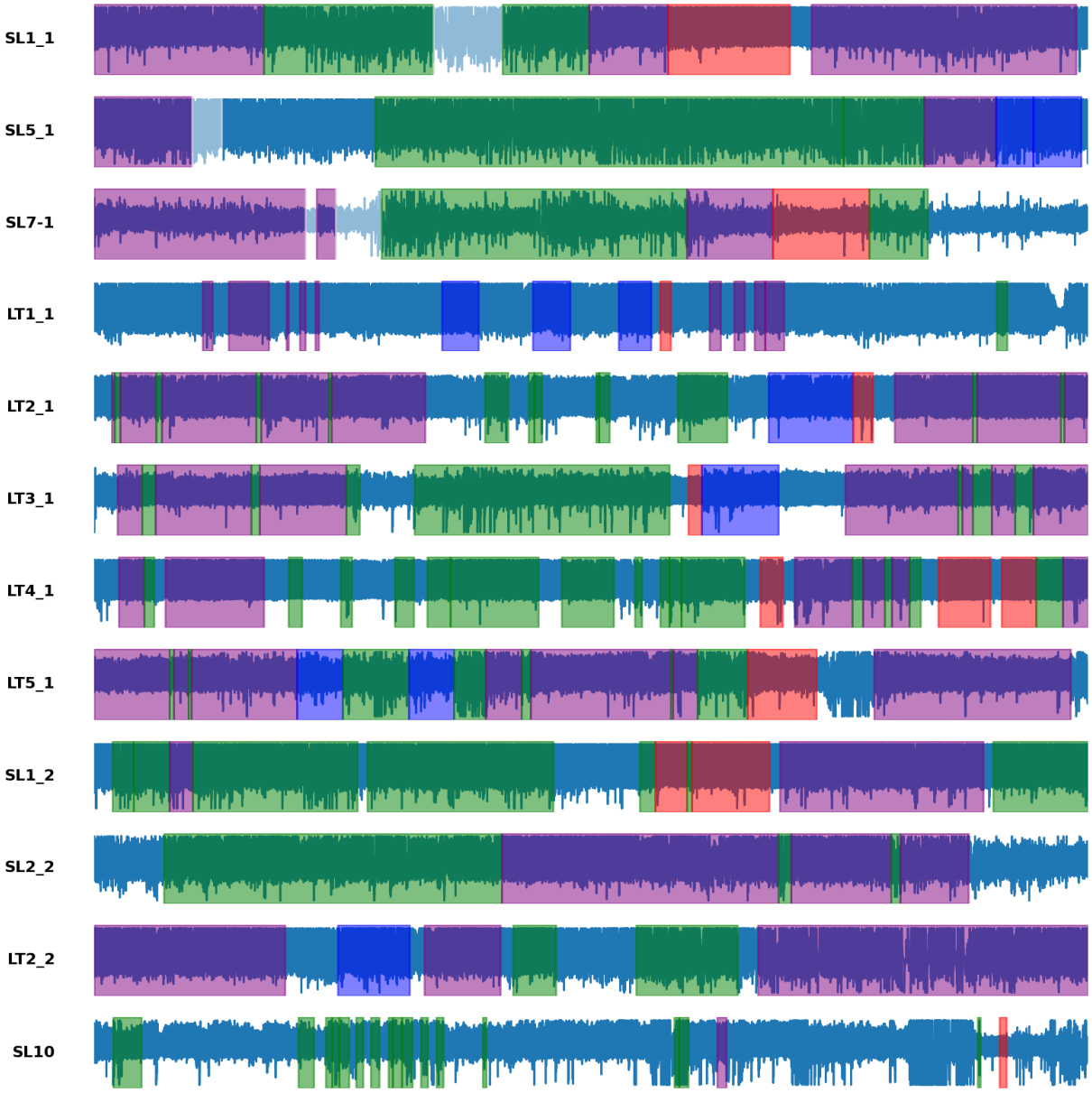

*Figure 1b: Overview of all annotated recordings. **Purple: Baseline**, **Green: Positive aroused**, **Red: Negative aroused**, **Blue: Low aroused**.*

Table 1: All recordings and the amount of data per affective state. Times are noted as minutes:seconds.

name	Baseline	Positive	Negative	Low engagement
SL1_1	*7:57*	*3:58*	*1:53*	
SL5_1	*2:34*	*8:11*		*1:16*
SL7-1	*5:14*	*6:13*	*1:36*	*0:45*
LT1_1	*1:52*	*0:11*	*0:11*	*1:50*
LT2_1	*7:42*	*1:49*	*0:17*	*1:22*
LT3_1	*6:33*	*5:18*	*0:13*	*1:07*
LT4_1	*4:04*	*4:23*	*1:39*	
LT5_1	*11:58*	*3:19*	*1:23*	*1:51*
SL1_2	*3:38*	*10:15*	*1:39*	
SL2_2	*4:41*	*3:48*		
LT2_2	*8:41*	*2:05*		*1:02*
SL10	*0:16*	*3:38*	*0:12*	
Total	***01:05:10***	***53:08***	***09:03***	***09:13***

In some recordings, the annotations were highly detailed, capturing affective states as brief as 10 seconds. However, biosignals typically do not respond quickly enough to reflect such short-lived emotional states except when highly intense. To address this, two morphological filters were applied to the annotations. The first filter merged annotations of the same type if they were less than 120 seconds apart and not separated by annotations of a different type. The second filter removed annotations shorter than 60 seconds if they were situated between annotations of another type. These two morphological operations, collectively referred to as merge filtering, reduced the number of short, fragmented annotations and increased the length of the remaining annotations. Figure 2 displays the annotations post-filtering.

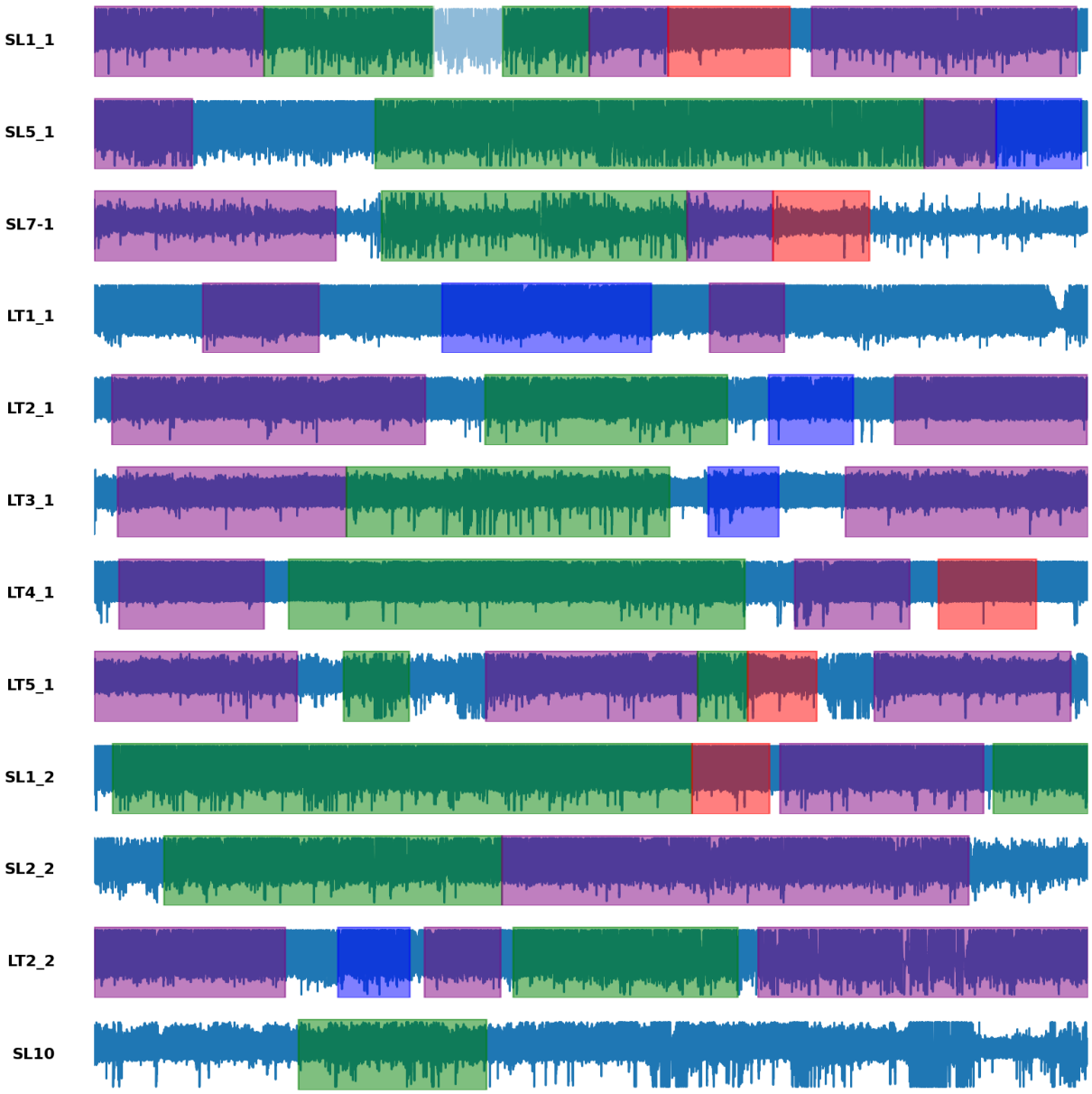

Figure 2: Overview of annotations after merge filtering. The merge filtering removed very short annotations (<30s) and merged scattered annotations when they were closer than 120s together without being separated by annotations of a different kind. ***Purple: Baseline****,* ***Green: Positive aroused****,* ***Red: Negative aroused****,* ***Blue: Low aroused****.*

Table 2: Amount of data per affective state per recording after merge filtering.

name	Baseline	Positive	Negative	Low engagement
SL1_1	*7:57*	*3:58*	*1:53*	
SL5_1	*2:34*	*8:19*		*1:17*
SL7-1	*5:25*	*5:04*	*1:36*	
LT1_1	*3:19*			*3:37*
LT2_1	*8:07*	*3:53*		*1:22*
LT3_1	*7:35*	*5:09*		*1:07*
LT4_1	*3:54*	*6:51*	*1:28*	
LT5_1	*12:23*	*2:19*	*1:23*	
SL1_2	*3:16*	*12:49*	*1:15*	
SL2_2	*4:55*	*3:34*		
LT2_2	*8:41*	*3:15*		*1:02*
SL10		*5:30*		
Total	**01:08:06**	**01:00:41**	**07:35**	**08:25**

Next, feature signals were calculated from the preprocessed ECG and EDA signals. These feature signals were then segmented using the filtered annotation regions.

2.2 ECG features

Report D2.2 outlines a robustly designed pipeline for extracting QRS peaks of high quality from signals contaminated with substantial noise bursts. The processing pipeline employed a two-stage approach. Initially, the signal's quality was assessed, after which QRS peaks were extracted using a robust deep learning network. Subsequently, peaks were locally readjusted to enhance timing accuracy and reduce jitter.

The extracted interbeat intervals were then subjected to analysis, focusing on a broad spectrum of HR (heart rate) and HRV (heart rate variability) features. Existing literature identifies numerous features, which can be grouped into three categories: time-domain, frequency-domain, and nonlinear features [1,3]. Time-domain features encompass the mean and standard deviation of all interval durations within a specified time window. Frequency-domain features entail interpolating the interbeat interval (IBI) signal by fitting a low-frequency sine wave to the interbeat intervals, followed by assessing its frequency spectrum in terms of low, medium, and high-frequency power. The ratios between these powers are also considered features. Nonlinear dynamics, on the other hand, employ Poincare plots, phase space representations, or metrics from information theory to quantify the degree of nonlinearity or chaos, and therefore, variability, among the intervals [2]. Table 3 provides a list of the most

commonly utilised features within each category, the minimum duration required and its physiological basis when known.

Table 3: Overview of interbeat interval based HR and HRV metrics. Based on combined information from [3] and [4].

Name	**Description**	**Minimum requirement**	**Physiological basis**	**Used**
Time domain				
Mean NN	Average NN interval	1-2 minutes		Yes
SDNN	Standard deviation of NN intervals	10-60s [4]	Cyclic components	Yes
RMSSD	Root Mean Square of Successive Differences	10-60s [4]	Vagal tone	Yes
SDRMSSD	Ratio between SDNN and RMSSD	10-60s [4]	Correlates with LFHF	Yes
Frequency domain				
VLF	Power in very low frequency range (<0.04Hz)	>5 minutes	Long-term regulation mechanics	No
LF	Power in low frequency range (0.04-0.15Hz)	2-5 minutes	Sympathetic and vagal activity	Yes
HF	Power in high frequency range (0.15-0.4Hz)	1-2 minutes	Vagal tone	Yes
LFHF	Ratio between low and high frequency power	2-5 minutes	Sympathethic and vagal actvitiy	Yes
LFn	Normalized low frequency power using total frequency power	2-5 minutes	See LF	Yes
HFn	Normalized high frequency power using total frequency power	1-2 minutes	See HF	Yes
LnHF	Log normalized high frequency power	1-2 minutes	See HF	Yes
Non-linear				
SD1	Poincare plot spread along the line of identity	>5 minutes		No

SD2	Poincare plot spread perpendicular to the line of identity	>5 minutes		No
SD1/SD2	Ratio between SD1 and SD2	>5 minutes		No
ApproxEn	Approximate entropy	>5 minutes		No
SampEn	Sample Entropy	>5 minutes		No
CD	Correlation dimension (min. num. of variables to describe the system	>5 minutes		No

IM-TWIN's requirement for real-time processing necessitated the selection of short-term HRV (heart rate variability) features. Consequently, non-linear features were excluded. Two segmentation windows were utilised: a shorter, non-overlapping 30-second window for time-domain features [4] and a longer 120-second window with a 90-second overlap for extracting high and low-frequency features [3]. The HRV features selected include MeanNN, SDNN, RMSSD, SDRMSSD, LF, HF, LFHF, LFn, HFn, and LnHF. Notably, RMSSD and HF-derived features are significant due to their physiological basis in reflecting vagal tone activity [5]. Vagal tone refers to the activity of the vagus nerve, part of the parasympathetic nervous system that activates during rest [1]. Consequently, a high vagal tone indicates a low-aroused state.

Features were extracted employing two methodologies: dynamic and static segmentation. In dynamic segmentation, time and frequency features were initially extracted using 30s and 120s sliding windows, respectively. The result is a feature signal that produces one feature value every 30 seconds, applicable to both 30s and 120s windows due to the 90s overlap. Feature values within an annotation region were then averaged to yield a single value per region, forming the dynamic feature set. In contrast, static segmentation directly uses annotation regions to segment ECG data. Following recommendations to use windows of uniform duration [3], annotation regions were either shortened or expanded from the centre to fit a 120s window, the shortest feasible window for extracting both time and frequency features. The outcome is the static feature set. This methodology is exclusively performed on merge-filtered annotations to minimise overlap between adjusted annotation regions. Both methodologies are schematically illustrated in Figure 3, while Figures 4 and 5 display examples of each segmentation approach.

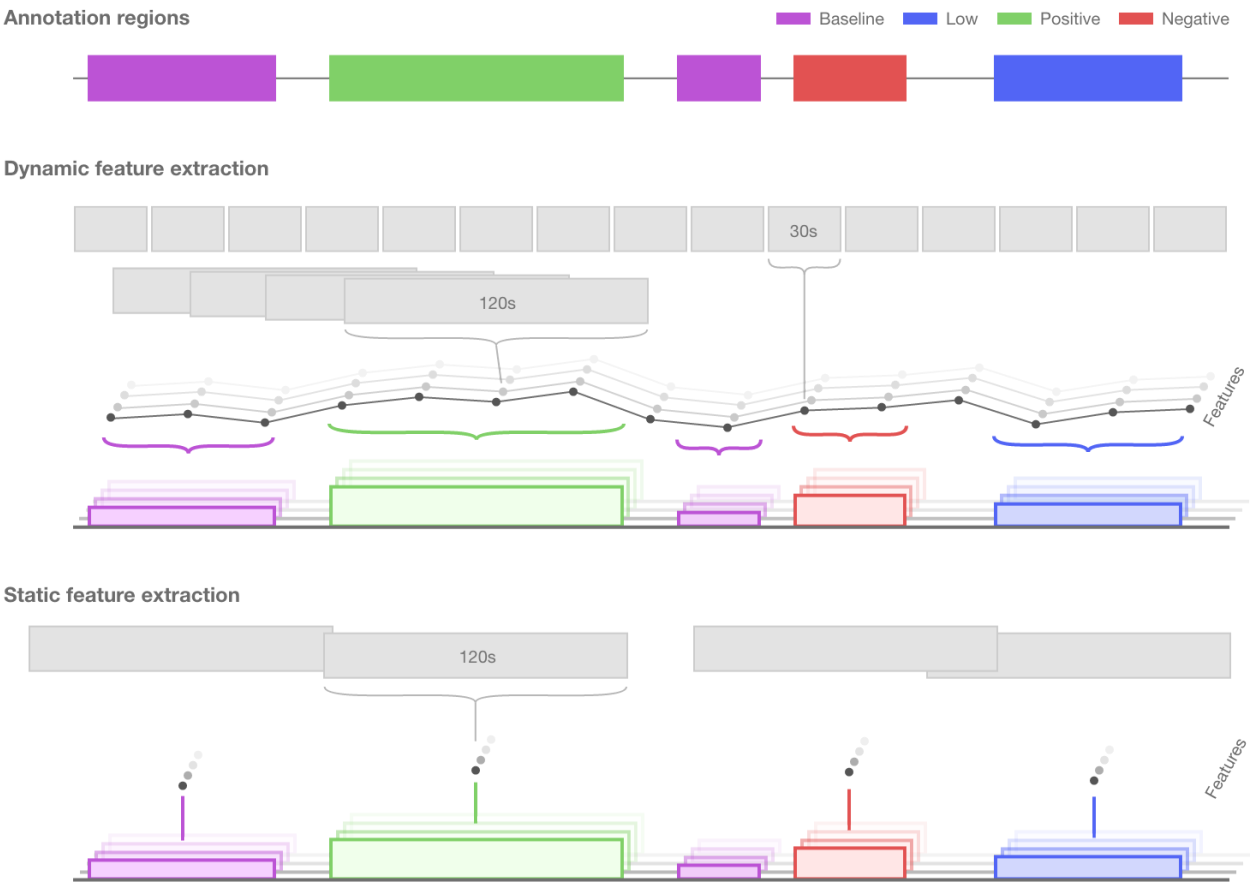

Figure 3: We use two segmentation methodologies to go from the biosignals to one feature value per annotation. Dynamic feature extraction uses two segmentation windows to generate an intermediate feature signal that has one value every 30s. Values within one region are then averaged. Static segmentation defines segmentation windows based on the annotations. Short annotations are skipped to minimise window overlap.

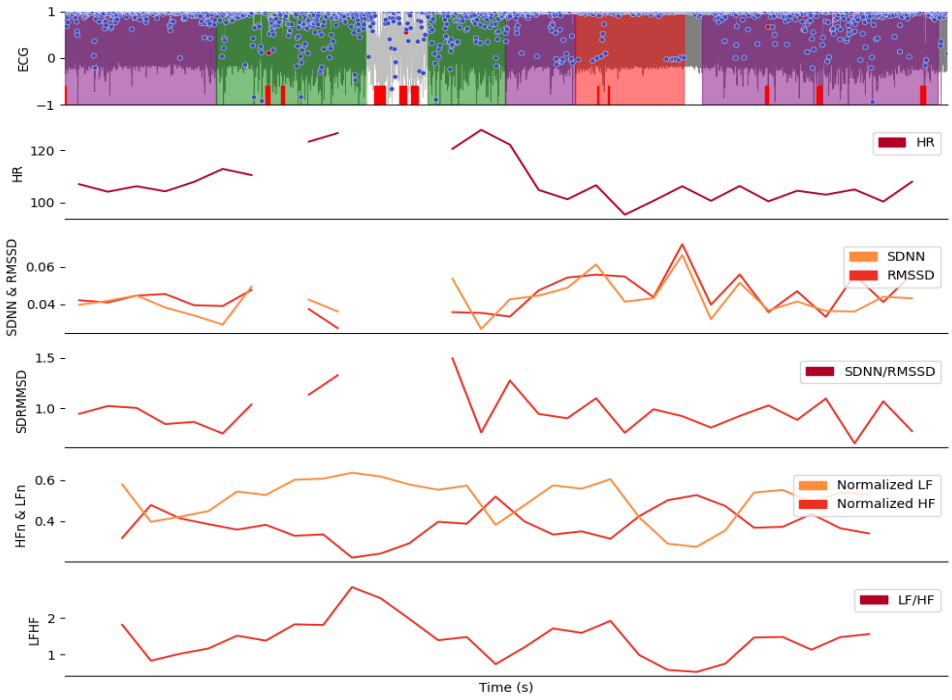

Figure 4: Example of dynamic HRV feature extraction that uses two segmentation windows for the time and frequency based HRV features, respectively.

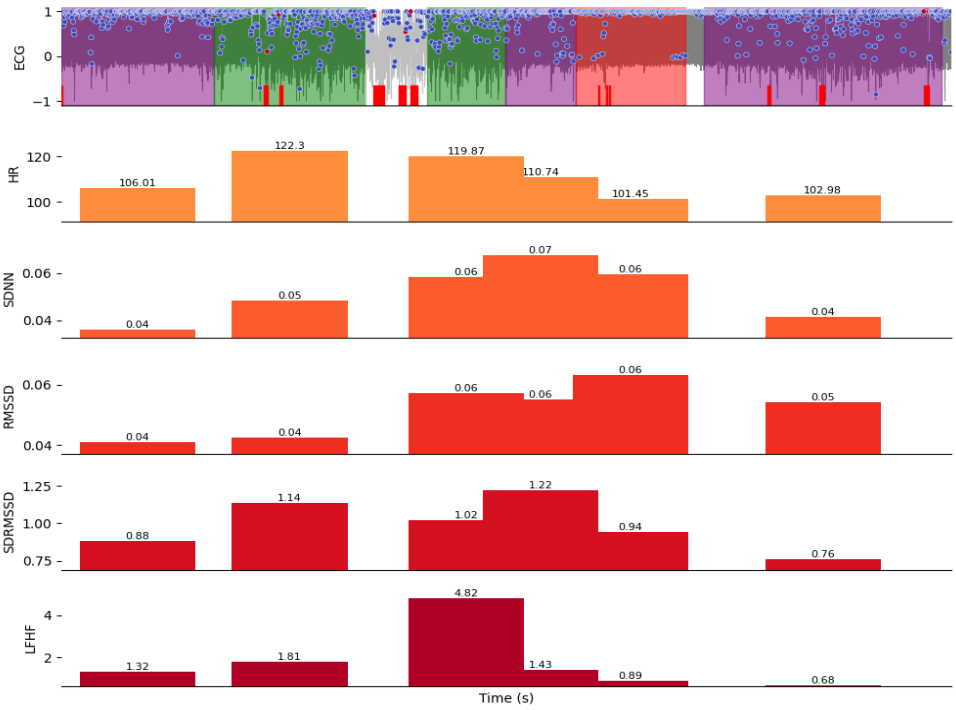

Figure 5: Example of static HRV feature extraction that uses one fixed width segmentation window located in the centre of each annotation region.

HR and HRV features were extracted for all 12 recordings using both segmentation methodologies. Figures 6 and 7 present the resulting distributions for each annotation type in the form of boxplots, showcasing the median, interquartile range, 95% confidence interval, and outliers. To mitigate interpersonal variances, features were individually normalised by subtracting the mean of the values derived from baseline segments. Consequently, a negative overall distribution indicates that the feature holds a value lower than the baseline, and vice versa. It's important to note that the baseline distribution does not necessarily have to be zero, as individuals may have multiple baseline-annotated segments.

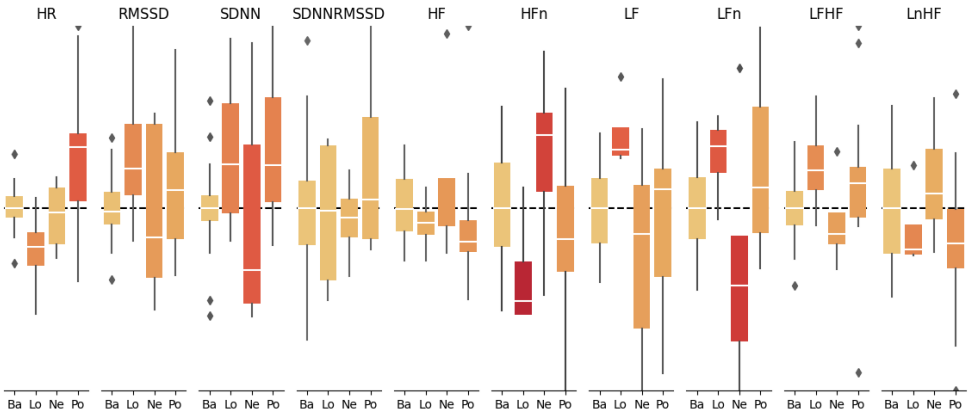

*Figure 6: Distribution of all HR and HRV features for all four annotation types using **dynamic segmentation**. The features are individually baseline corrected by subtracting the mean of all baseline segments. Ba: Baseline, Lo: Low aroused, Ne: Negative aroused, Po: Positive aroused*

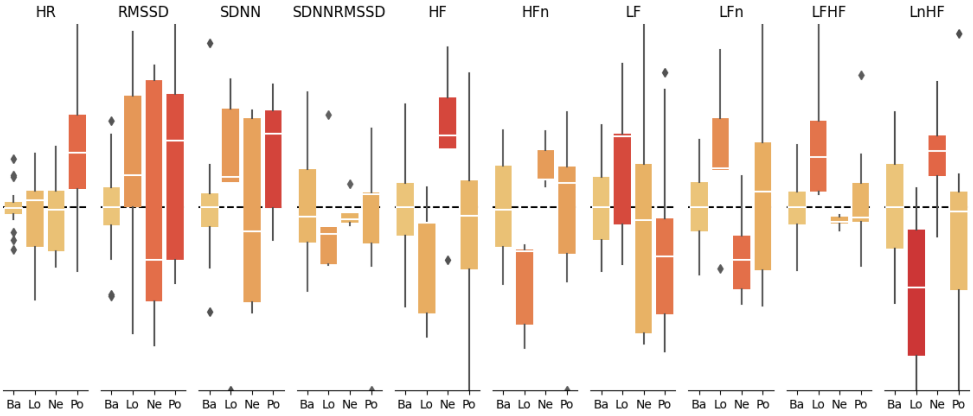

*Figure 7: Distribution of all HR and HRV features for all four annotation types using **static segmentation**. The features are individually baseline corrected by subtracting the mean of all baseline segments. Ba: Baseline, Lo: Low aroused, Ne: Negative aroused, Po: Positive aroused*

2.3 EDA features

As described in report D2.2, the Electrodermal Activity (EDA) signal is first low-pass filtered using an adaptive wavelet-based filter to remove muscle activity (EMG) interference. Afterwards, the signal is split into its low-frequency tonic and high-frequency phasic component. The tonic component, reflecting the Skin Conductance Level (SCL), can be used without further processing as its value is directly correlated to the activity of the sympathetic nervous system [6]. Table 4 shows commonly used tonic features based on moment-theory [7].

Table 4: Commonly used features to describe the tonic EDA component or SCL within a segmentation window.

Name	**Description**	**Physiological basis**
MeanTonic	Mean value	General SCL
SDTonic	Standard deviation	Reflects SCL activity
SkewTonic	Skewness	SCL trend
KurtTonic	Kurtosis	SCL trend

The phasic component of the EDA signal, which reflects the Skin Conductance Response (SCR), indicates the short-term activity of the signal. As this signal is event-based (meaning that specific or non-specific stimuli cause noticeable variations or "bumps" in the signal), it needs to be processed first, particularly by extracting the peaks in the signal [6]. Subsequent feature extraction focuses on the frequency, timing, and characteristics of these bumps. Unfortunately, as described in report D2.2, it is not feasible to extract and analyse peaks in our EDA data due to a strong noise component that exists within the same frequency band as the peaks.

To still extract phasic EDA features, we devised a new feature based on the power of the frequency band. As the noise component remains relatively constant, an increase in the power of the [0.04-0.24] Hz frequency band [10] typically corresponds with an increase in SCR peaks (see D2.2). Because the resulting signal describes power over time at a rate of 500Hz, we use descriptive features to aggregate the signal within each segmentation window. Table 5 presents a list of descriptive features in the time, frequency, and nonlinear domains that we employ.

Table 5: Descriptive features used to describe the behaviour of frequency power within a segmentation window.

Name	**Description**	**Physiological basis**
Time domain		

TFMean	Mean value	Average SCR activity
TFSD	Standard deviation	Reflects SCL activity
TFSkew	Skewness	SCL trend
TFKurt	Kurtosis	SCL trend
Frequency domain		
LF	Low frequent variations in SCR activity	
HF	High frequent variations in SCR activity	
LFn	LF divided by total power	
HFn	HF divided by total power	
LFHF	Ratio between LF and HF	
Non-linear		
SampEn	Sample entropy	
ApEn	Approximate entropy	
CD	Correlation dimension	

We selected all features listed in Table 5 as there are currently no guidelines about feature extraction for the phasic time-frequency signal. Figure 8 plots the individually baseline corrected distributions for all EDA features as boxplots showing the median, interquartile range, 95% confidence intervals, and outliers.

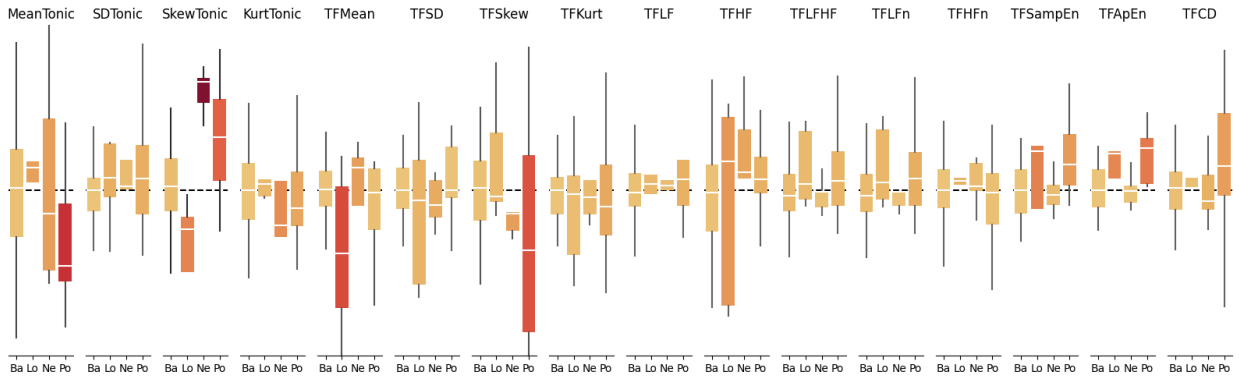

*Figure 8: Distribution of all EDA features for all four annotation types using **dynamic segmentation**. The features are individually baseline corrected by subtracting the mean of all baseline segments. Ba: Baseline, Lo: Low aroused, Ne: Negative aroused, Po: Positive aroused*

For statistical analysis, both the HR&HRV and EDA feature sets are combined into one set. This is done for both the dynamic and static segmentation strategy. The resulting feature sets include 26 features each.

3. Statistical Analysis

Before applying machine learning techniques to detect different states of affect, we first analyse the data using general statistical methods. Given the small size of our dataset and the high variance among children, we face a risk of overfitting to spurious patterns in the data if we aren't careful. Therefore, we begin by assessing overall patterns through statistical analysis. If statistical analysis fails to find significant differences between annotations, it is highly unlikely that machine learning, which also relies on statistics, will uncover any patterns in the data.

We initiate our analysis with a Multivariate ANalysis Of VAriance (MANOVA) test, a single test that assesses the variance of multiple variables across multiple groups [11]. Although an ANOVA is typically used in similar situations, we have many dependent variables (i.e., features) in our dataset. Performing multiple ANOVAs (one for each feature) on the same dataset raises the risk of Type I errors (false positives) [12]. Therefore, we opt to conduct an omnibus MANOVA test to determine if there are any significant differences among the groups. In our context, the groups correspond to three of the four annotated states: low aroused, positive aroused, and negative aroused. We use only three of the four states because the data has already been baseline corrected, as outlined in sections 2.2 and 2.3. As a result, the baseline class serves as a reference and does not provide any new information.

In this section, we will employ the feature set extracted using the dynamic segmentation approach for all tables and figures. However, we will also apply the described process to the statically segmented feature set at the end of this section, reporting only the final results to maintain conciseness in the report.

3.1 MANOVA

3.1.1 Univariate assumption 1: Independence

MANOVA has several assumptions about the data that must be met before we can perform the test [11]. First, like a traditional ANOVA, a MANOVA expects all observations to be independent. This means that each observation must come from a unique participant, and each affective state can only be observed once per participant. Normally, a Repeated Measures MANOVA would be performed when there are multiple observations for each participant. However, since we only have 2 participants with two recordings each, this approach does not suit our data. As a result, we remove the two duplicate recordings (SL1_2 and LT2_2) from our dataset. We also remove the shorter of the two positive aroused state measurements for SL1_1 and LT5_1. This reduces our dataset to a total of 16 observations: 4 low, 4 negative, and 8 positive.

Table 6: Schematic depiction of the dataset used for the statistical analysis. In total there are 16 affective states observed at 8 participants. Each row corresponds to one affective state observed at one participant. Each observation has 26 features.

#	Recording	Label	Features
1	SL1_1	**Positive**	[26]
2	SL2_2	**Positive**	[26]
3	SL5_1	**Positive**	[26]
4	SL7_1	**Positive**	[26]
5	LT2_1	**Positive**	[26]
6	LT3_1	**Positive**	[26]
7	LT4_1	**Positive**	[26]
8	LT5_1	**Positive**	[26]
9	SL1_1	**Negative**	[26]
10	SL7_1	**Negative**	[26]
11	LT4_1	**Negative**	[26]
12	LT5_1	**Negative**	[26]
13	SL5_1	**Low**	[26]
14	LT1_1	**Low**	[26]

15	LT2_1	Low	[26]
16	LT3_1	Low	[26]

The assumptions for MANOVA are an extension of those for the univariate ANOVA. Since a multivariate assumption can only be valid if all separate univariate assumptions are met, we first check the univariate assumptions before moving on to the multivariate ones. This approach is particularly useful given the size of our dataset. With only 16 observations, and the smallest group containing just 4, the rule of thumb suggests that we can include a maximum of 3-4 features. By first testing the univariate assumptions, we can identify and exclude any poorly-behaved features before making our final selection of 3 features.

3.1.2 Univariate assumption 2: Normality

In a MANOVA, ANOVA's assumption of normality is extended to an assumption of normality of the multivariate distribution. Consequently, we first test the normality of each dependent variable. From a visual perspective, the univariate assumption appears to be met, as can be seen in Figures 6, 7, and 8. However, for thoroughness, we performed a Shapiro-Wilk normality test for all features across all groups. The results are listed in Table 7. From this table, we can see that several features cannot be used in our statistical analysis, as the null hypothesis of normality cannot be rejected for them.

Table 7: Normality assumption check per feature per group. A value below 0.05 means that we cannot reject the null hypothesis (e.g., it deviates from normality) and hence, should remove the feature from the analysis.

Feature name	Low	Negative	Positive	Accepted
HR	0,72	0,51	0,58	Yes
SDNN	0,76	**0,04**	0,34	NO
RMSSD	0,62	0,44	0,77	Yes
SDNNRMSSD	0,15	0,96	0,69	Yes
HF	0,88	0,46	0,12	Yes
HFn	0,07	0,99	0,99	Yes
LF	**0,04**	0,92	0,87	NO
LFn	0,56	0,63	0,18	Yes
LnHF	**0,02**	0,99	0,9	NO
LFHF	0,89	0,76	0,09	Yes
MeanTonic	0,76	0,24	**0,02**	NO
SDTonic	0,45	0,79	0,17	Yes

SkewTonic	0,72	0,36	0,34	Yes
KurtTonic	0,06	0,83	0,82	Yes
TFMean	0,58	0,45	0,08	Yes
TFSD	0,41	0,9	0,92	Yes
TFSkew	0,28	0,75	0,07	Yes
TFKurt	0,81	0,95	0,88	Yes
TFLF	0,44	0,79	0,06	Yes
TFHF	0,89	0,67	0,81	Yes
TFLFHF	0,07	0,74	0,09	Yes
TFLFn	0,08	0,69	0,09	Yes
TFHFn	0,95	0,46	0,07	Yes
TFSampEn	0,09	0,39	0,61	Yes
TFApEn	0,34	0,66	0,58	Yes
TFCD	0,38	0,94	0,39	Yes

As we said before, the multivariate assumption is checked after final feature selection. We first proceed with the other MANOVA assumptions in univariate form.

3.1.3 Univariate assumption 3: Outliers

Like an ANOVA, a MANOVA expects no outliers. As observed in Figures 6, 7, and 8, there are a few outliers present in the data. In one specific instance, child LT2_2 exhibited a heart rate of 130 BPM during a supposed low-arousal state. Since the state lasted only 30 seconds, and the heart rate showed an increase relative to the baseline, we decided to exclude this data point from the dataset. Other outliers, upon examination, were not found to be particularly outside of the expected distribution and, as such, were retained in the dataset.

3.1.4 Univariate assumption 4: Homogeneity of variances

In a MANOVA, the univariate assumption of homogeneity of variances is extended to the equality of the covariance matrices of all dependent variables. As mentioned before, we start by checking the univariate assumption that the variances should be equal between the different groups for each variable (i.e., a variable should have approximately the same variance across all groups). The univariate assumption is assessed using Levene's test. A significant result on Levene's test indicates that variances are heterogeneous among the different groups, and thus the assumption is violated. Consequently, we need a p-value > 0.05 for all variables. Table 8 displays the results of Levene's test for all variables. Several variables should be excluded due to their p-value < 0.05, indicating that the assumption of homogeneity of variances is not met for these variables.

Table 8: Testing for homogeneity of variances per feature across all four groups using Levene's test.

Feature name	p-value	Accepted
HR	0,82	Yes
RMSSD	0,4	Yes
SDNNRMSSD	**0,02**	NO
HF	0,5	Yes
HFn	0,53	Yes
LFn	0,27	Yes
LFHF	0,76	Yes
SDTonic	0,68	Yes
SkewTonic	0,35	Yes
KurtTonic	0,11	Yes
TFMean	0,53	Yes
TFSD	0,24	Yes
TFSkew	**0,04**	NO
TFKurt	0,33	Yes
TFLF	0,82	Yes
TFHF	0,12	Yes
TFLFHF	0,36	Yes
TFLFn	0,36	Yes
TFHFn	0,36	Yes
TFSampEn	0,26	Yes
TFApEn	0,48	Yes
TFCD	**0,01**	NO

After testing for normality and homogeneity of variances, 19 features remain available for further analysis. From these remaining features, HR (Heart Rate), RMSSD (Root Mean Square of Successive Differences), and SkewTonic are selected for the multivariate assumption tests in the MANOVA. HR is selected because it reflects a fundamental aspect of arousal - elevation in heart rate, which is typically associated with increased arousal. RMSSD is chosen due to its known correlation with vagal tone [5]. Vagal tone, which is the activity of the vagus nerve, increases when arousal is low [1]. Hence, a higher RMSSD value would indicate a more relaxed

state. Lastly, SkewTonic is chosen as it has been identified as the best predictor of arousal among all the tonic-related features in previous studies [13].

In this case, we perform univariate assumption tests on all features before choosing a final selection of features. Assumption tests only check the validity of a test. As such, we do not get information about their influence on the outcome variable or their effect on significance. As such, it is ok to first do assumption checks before feature selection. Moreover, due to the small datasets there is a high risk of a feature being excluded. If we would have chosen three features first, we could have been at risk that one or two of the features could not be included in the analysis. In that case, you have to determine other features and do a re-check or do the analysis using only one or two features. As we have a very limited dataset, this would affect statistical power tremendously. Hence, this would have been very unfavourable.

3.1.5 Multivariate assumption 1: Normality of multivariate distribution

After feature selection, the assumption of multivariate normality on the set of the three selected features is checked using the Henze-Zirkler test [14]. Results are listed in Table 9. As all p-values are higher than 0.05, multivariate normality can be assumed across all groups.

Table 9: P-values of the Henze-Zirkler test for multivariate normality. Across all groups, the null-hypothesis can be rejected and, hence, multivariate normality can be assumed.

Low arousal	Positive arousal	Negative arousal	Accepted
0.39	0.66	0.39	Yes

3.1.6 Multivariate assumption 2: Homogeneity of covariance matrices

Next, the multivariate assumption of equality of covariance matrices is checked using Box's M test. The null hypothesis (covariance matrices are not equal) is rejected if p>0.001 [15]. The result ($\chi^2(12)=28.8$, $p=0.0049$) allowed us, therefore, to reject the null hypothesis and assume equality of the covariance matrices. As such, the fourth and last assumption of the MANOVA is met.

3.1.7 Results

All multivariate assumptions are met. Hence, we are allowed to proceed with the MANOVA on the dynamically segmented dataset now containing 16 observations and 3 features. Figure 9 shows the distribution of the features across the three annotated affective states.

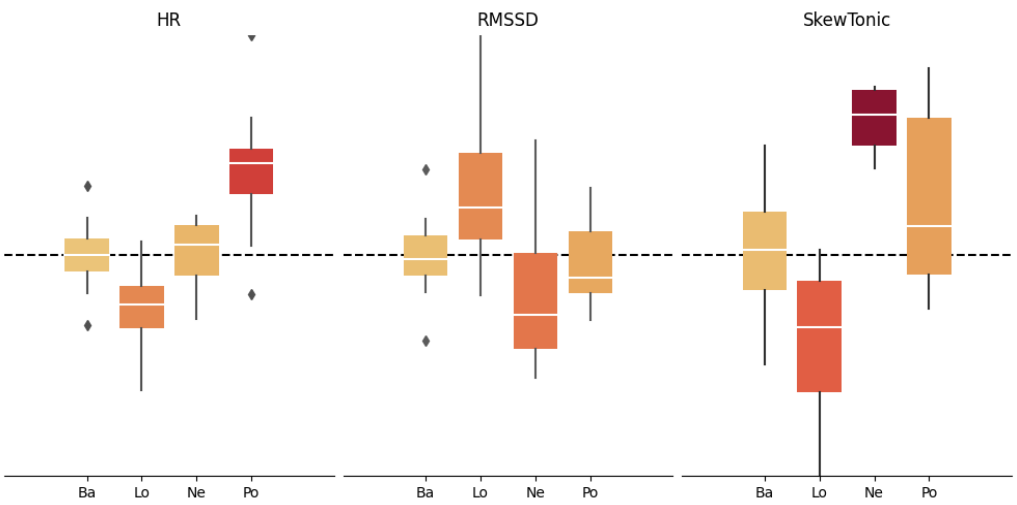

Figure 9: Boxplots showing the distribution of the final feature selection across the three annotated affective states and the reference baseline state.

Using a MANOVA without intercept and Pillai's trace, there is a significant effect of affective state on the HR, RMSSD, and EDA tonic skewness using the dynamic segmentation strategy, **V = 1.27, F(9,39) = 3.18, *p* = .006**.

The statically segmented dataset meets the requirements for both the univariate and multivariate assumptions, just like the dynamically segmented dataset. However, there are fewer positive observations in the statically segmented dataset, with only four compared to eight in the dynamically segmented dataset. The same three features (HR, RMSSD, and SkewTonic) were chosen for the multivariate analysis, which yielded a dataset containing 12 observations and 3 features. In contrast to the dynamically segmented dataset, using a MANOVA without intercept and Pillai's trace, there was no a significant effect of affective state on the HR, RMSSD, and EDA tonic skewness, **V = 1.01, F(9,27) = 1.51, *p* > .05**.

There are several possible explanations for this outcome. First, the reduced number of observations in the statically segmented dataset (especially for the positive state) likely reduced the power of the statistical analysis. Secondly, the static segmentation approach only considers a 2-minute window of each state, centred around the middle of the state. When the states are shorter than 2 minutes, the window often overlaps with other affective states, which may introduce noise into the data and impact the analysis. Overall, these factors could have contributed to the lack of significance in the results for the statically segmented dataset.

3.2 Follow-up analysis

Upon finding a significant omnibus MANOVA test result for the dynamically segmented dataset, we decided to proceed with a subsequent analysis. While a typical approach would involve

conducting multiple ANOVAs to assess the predictive power of each feature, we instead chose to perform a discriminant analysis. Our goal was not to examine the explained variance of a single feature, as no single feature has been identified that can simultaneously differentiate between arousal and valence. Instead, we employed Linear Discriminant Analysis (LDA) to investigate the combined effects of HR, RMSSD, and tonic skewness. However, due to the small size of our dataset and the reliance of LDA on labels for its analysis, we also utilised Principal Component Analysis (PCA), a dimensionality reduction technique that explores underlying correlations without prior knowledge of the labels. Figure 10 presents the dataset transformed using both PCA and LDA.

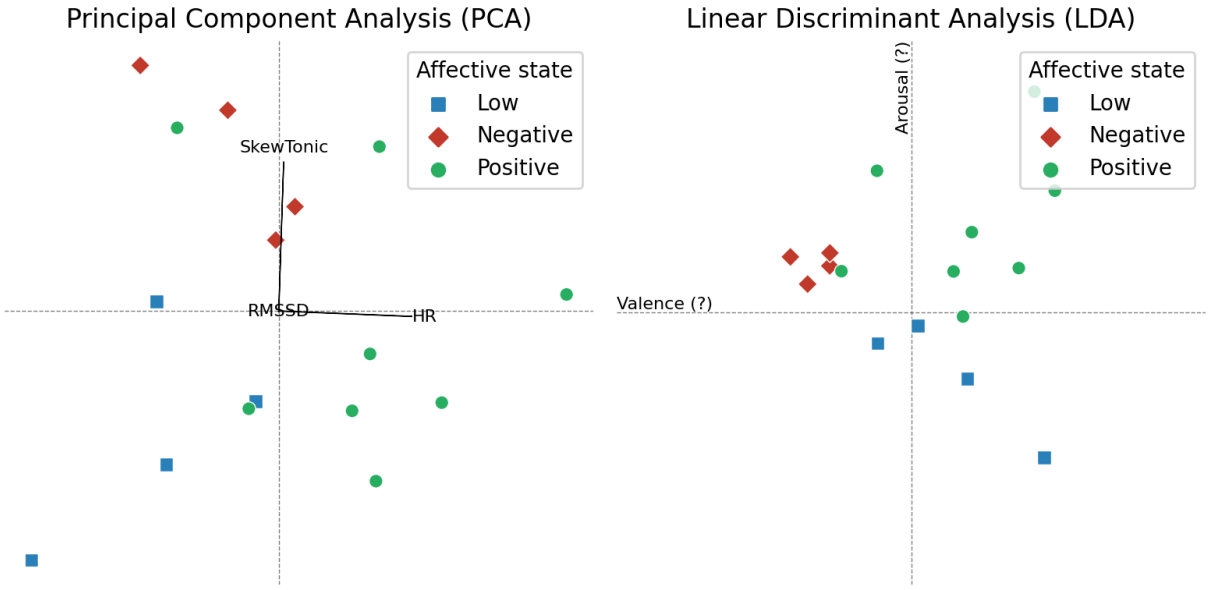

Figure 10: Three features (HR, RMSSD, and SkewTonic) of the dynamically segmented dataset are analysed using PCA and LDA on their underlying correlations and their predictive power with respect to the affective state. Where PCA maps HR and SkewTonic to the x and y-axis, respectively, LDA uses the data's labels to find the optimal separation of the data in a 2 dimensional plane. Surprisingly, LDA's separation hints towards a near-perfect valence-arousal separation.

The separation results of the PCA and LDA analyses, although impressive, have to be taken in with caution. The dataset only features 16 data points making generalizability impossible to assess. That said, both techniques can separate the datapoints very good. LDA's dimensions even seem to map on the theoretical valence and arousal axes and even without labels, PCA finds tonic skewness and HR to be two perpendicular dimensions. Hence, despite the small number of data points and the limited information about generalizability, this should be seen as a first success.

3.3 Classification

The promising separation achieved by PCA and LDA in the previous section prompted us to assess the results using a classifier in a real-world study design. In this section, we'll conduct LDA classification based on all data points extracted through dynamic segmentation. In other words, we aim to estimate affect every 30 seconds and compare it to the annotated affect. Accordingly, we will employ LDA to learn optimal separation from unaggregated data points. We will use all four annotated classes since a patient is not expected to be in an affective state 100% of the time. Therefore, a real-world system should be capable of detecting the baseline state as well. Additionally, we will include the repeated measurements for the positive states of SL5_1 and LT1_1, as they no longer pose any analytical problems. Furthermore, to prevent overfitting, we will divide the dataset into a training set and a test set. We will perform the split based on participants, as this reflects real-world usage. In practice, the algorithm would be trained/tuned on a fixed set of participants, and the system should maintain its performance even when introduced to new, unseen children. To maximise the validity of our benchmark, we will allocate 50% of the data for training (D1) and 50% for testing (D2). We will use a stratified, grouped k-fold split with k=2 to ensure a balanced number of data points per affective state in both the training and testing sets. Table 10 provides details of the splits and their associated recordings.

Table 10: Train-test split of the full dataset based on participants. The number of data points per affective state and their totals per set are listed. Stratified group splitting was used to balance both sets based on the amount of data points per state.

Recording	**Baseline**	**Low**	**Positive**	**Negative**
D1 (training)				
SL1_1	12		8	4
LT1_1	6	7		
LT3_1	13	2	8	
LT4_1	7		14	2
Total	**38**	**9**	**30**	**6**
D2 (testing)				
SL5_1	3	1	9	
SL7_1	9		5	3
LT2_1	13	3	8	

LT5_1	22		4	3
SL2_2	10		7	
Total	**57**	**4**	**33**	**6**

We will conduct validation in two ways: first by training on dataset D1 and testing on D2, and then by training on D2 and testing on D1. For both validation rounds, we will compute the precision, recall, and F1-score. Given that the dataset is heavily unbalanced (with approximately 20 times more baseline points than low-aroused observations), we will primarily focus on the F1-score, as relying on accuracy would give a biassed view. For instance, predicting only the baseline state would be correct about 60% of the time, which would result in a misleading 60% accuracy. The F1-score accounts for this imbalance. The classification results for both validation rounds are presented as confusion matrices in Figure 11, while Table 11 reports the performance metrics.

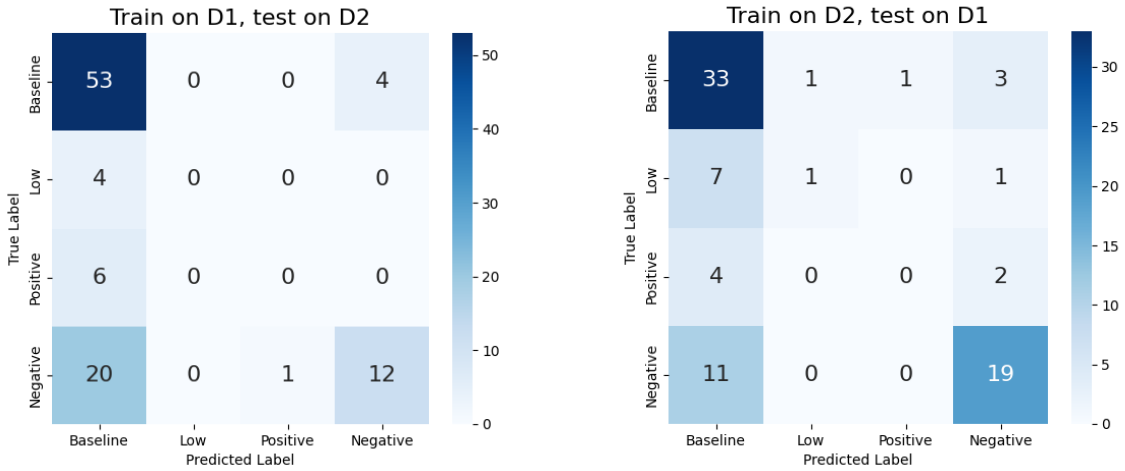

Figure 11: Confusion matrices of both validation folds. Once of the LDA trained on D1 and tested on D2, and once trained on D2 and tested on D1.

Table 11: LDA's precision, recall, and F1-score for each affective state in each fold.

Affective state	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**Support**
Fold 1				
Baseline	64.0	93.0	76.0	57 points
Low	0	0	0	4 points
Negative	0	0	0	6 points

Positive	75.0	36.0	49.0	33 points
Subtotal	**35.0**	**32.0**	**31.0**	
Fold 2				
Baseline	60.0	87.0	71.0	38 points
Low	50.0	11.0	18.0	9 points
Negative	0	0	0	6 points
Positive	76.0	63.0	69.0	30 points
Subtotal	**47.0**	**40.0**	**40.0**	
Total	**41.0**	**36.0**	**35.5**	

Although an overall F1-score of 35.5% renders the system in its present form impractical for real-world use, there are some encouraging findings. The second fold, which benefits from greater patient diversity in the training data, yields superior results compared to the first fold, particularly in classifying positive affect. While low and negative affective states receive minimal support due to the imbalance of the data, the LDA classifier achieves respectable F1-scores for both baseline and positive states. In fact, the classifier detects a positive state in 63% of instances when a child is genuinely experiencing high arousal positive emotions. More importantly, when the classifier identifies a positive state, it is accurate 76% of the time. This level of precision is promising, as it bolsters the system's credibility; it is preferable to remain uncertain and default to the baseline state rather than incorrectly identifying an affective state when the child isn't experiencing it.

We employed a rigorous, though demanding, train-test split approach. In contrast to standard procedures that randomly partition data with an 80-20 split, we implemented a 50-50 patient-based split. As a result, the variance between the train and test set is maximised to closely reflect real-world conditions. Therefore, it is not surprising that the classifier achieves modest average performance compared to other studies in the literature [16,17]. However, these results are valid and highlight the unfortunate reality that current machine learning techniques struggle to identify patterns associated with affective states in the IM-TWIN dataset.

4. Conclusion and future developments

In this study, we extracted 26 features from HR, HRV, and EDA data using two different segmentation strategies. We then conducted an omnibus MANOVA test to assess the learnability of the resulting dataset. During the univariate and multivariate assumption checking

phase, seven features were excluded. Of the remaining 19 features, three were selected to mitigate the curse of dimensionality, as the smallest groups had only four observations. The MANOVA test showed significant results (p=.006) using the dynamic segmentation strategy and the HR, RMSSD, and EDA tonic skewness features. As a result, we performed follow-up analyses using PCA and LDA. LDA was highly successful in separating low, negative, and positive arousal states, with the two dimensions found to map directly to the theoretical valence and arousal dimensions. Because of this finding, an LDA classifier was built to assess the validity of these results in IM-TWIN's real-world application—predicting affective state every 30s for new, unseen children. We split the dataset into test and train sets using a challenging 50-50 inter-patient stratified K-fold split, where K=2. Unfortunately, the classifier could not detect two of the four affective states and only achieved a 35% F1-score. As a result, we conclude that machine learning cannot extract meaningful patterns correlated to affective states using IM-TWIN's dataset in its current form.

Although the LDA classifier is impractical for real-world use, the significant MANOVA test and successful LDA separation show the potential of IM-TWIN once the dataset quality is improved. Specifically:

1. The number of observations should be increased by at least a factor of 5-10, especially for the low and negative arousal states.
2. The dataset should be less skewed or large enough to address skewness through undersampling.
3. The study design should follow a predefined, balanced order of stimuli to avoid unnecessary data removal. Recordings should always begin with a baseline measurement, and all recordings should capture at least 2 consecutive minutes of each affective state.
4. Affective states should not be annotated in excessive detail. Although affective states can affect facial expressions within seconds, biosignals like EDA adapt more slowly. The difference between a 10-second positive arousal state followed by a baseline annotation is likely negligible. Meeting requirement 3 would likely address this issue.

The main takeaway from this report is one of optimism and potential. Assessing children's emotional responses using biosensors in noisy, ambulatory settings is challenging. Children are constantly in motion—jumping, rolling, running—while sensors continuously track their biosignals, trying to detect the microsecond changes in heart rhythm or the microsiemens variations in skin conductance level. Despite the enormity of this challenge, we obtained a small yet high-quality dataset through extensive preprocessing and achieved statistically significant results, which is noteworthy. Future work should prioritise improving data collection, guided by the recommendations outlined above. Nonetheless, the progress we've made should be viewed, albeit cautiously, as the first glimmers of life for IM-TWIN.

5. References

1. Cacioppo, J. T., Tassinary, L. G., & Berntson, G. (Eds.). (2007). Handbook of psychophysiology. Cambridge university press.
2. Camm, A. J., Malik, M., Bigger, J. T., Breithardt, G., Cerutti, S., Cohen, R. J., ... & Singer, D. H. (1996). Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Circulation, 93(5), 1043-1065.
3. Laborde, S., Mosley, E., & Thayer, J. F. (2017). Heart rate variability and cardiac vagal tone in psychophysiological research–recommendations for experiment planning, data analysis, and data reporting. Frontiers in psychology, 8, 213.
4. Munoz, M. L., Van Roon, A., Riese, H., Thio, C., Oostenbroek, E., Westrik, I., ... & Snieder, H. (2015). Validity of (ultra-) short recordings for heart rate variability measurements. PloS one, 10(9), e0138921.
5. Thayer, J. F., & Lane, R. D. (2000). A model of neurovisceral integration in emotion regulation and dysregulation. Journal of affective disorders, 61(3), 201-216.
6. Boucsein, W. (2012). Electrodermal activity. Springer Science & Business Media.
7. Munoz, M. L., Van Roon, A., Riese, H., Thio, C., Oostenbroek, E., Westrik, I., ... & Snieder, H. (2015). Validity of (ultra-) short recordings for heart rate variability measurements. PloS one, 10(9), e0138921.
8. Healey, J., & Picard, R. (1998, May). Digital processing of affective signals. In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181) (Vol. 6, pp. 3749-3752). IEEE.
9. Tronstad, C., Staal, O. M., Sælid, S., & Martinsen, Ø. G. (2015, August). Model-based filtering for artifact and noise suppression with state estimation for electrodermal activity measurements in real time. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 2750-2753). IEEE.
10. Posada-Quintero, H. F., Florian, J. P., Orjuela-Cañón, Á. D., & Chon, K. H. (2016). Highly sensitive index of sympathetic activity based on time-frequency spectral analysis of electrodermal activity. American Journal of Physiology-Regulatory, Integrative and Comparative Physiology, 311(3), R582-R591.
11. Warne, R. T. (2014). A primer on multivariate analysis of variance (MANOVA) for behavioral scientists. Practical Assessment, Research & Evaluation
12. Frane, A. V. (2015). Are per-family type I error rates relevant in social and behavioral science?. Journal of Modern Applied Statistical Methods, 14(1), 5.
13. Cecchi, S., Piersanti, A., Poli, A., & Spinsante, S. (2020, September). Physical stimuli and emotions: EDA features analysis from a wrist-worn measurement sensor. In 2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD) (pp. 1-6). IEEE.
14. Henze, N., & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. Communications in Statistics-Theory and Methods, 19(10), 3595-3617.
15. Hahs-Vaughn, D. (2016). Applied Multivariate Statistical Concepts. Taylor & Francis.
16. Yu, D., & Sun, S. (2020). A systematic exploration of deep neural networks for EDA-based emotion recognition. Information, 11(4), 212.
17. Sarkar, P., & Etemad, A. (2020). Self-supervised ECG representation learning for emotion recognition. IEEE Transactions on Affective Computing, 13(3), 1541-1554.