Nuances of Interrater Agreement on Automatic Affect Prediction from Physiological Signals

A Systematic Review of Datasets Presenting Various Agreement Measures and Affect Representation Schemes

Oana Madalina Fron[1]
Supervisor and Responsible Professor: Bernd Dudzik[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Oana Madalina Fron
Final project course: CSE3000 Research Project
Thesis committee: Bernd Dudzik, Catharine Oertel

Abstract

This study explores the influence of interrater agreement measures and affect representation schemes in automatic affect prediction systems using physiological signals. These systems often use supervised learning and require unambiguous and objective labeling, a challenge when multiple human annotators are involved, which can affect model performance. The research involved the first part of a two-stage process: systematically reviewing datasets and their characteristics concerning interrater agreement on the affective interpretation of physiological signals. This stage established a reliable foundation for the second step: a future analysis of model performance reported in technical papers utilizing these datasets. The main takeaways were that the number of raters varies significantly over datasets and the complexity introduced by combining affect representation schemes can negatively affect interrater agreement.

1 Introduction

Automatic affect prediction systems aim to recognize and analyse human emotions by training machine learning models on datasets with various types of signals (text, speech, images, and physiological signals) from study participants [1]. In order to train the prediction model for supervised learning, the datasets used as input need to be labeled. However, the humans which develop these systems are subjective by nature [2]. Therefore, they interpret and label the emotions provoked by these signals differently based on direct factors, such as previous personal experience, as well as indirect factors, like media interaction. Currently, we lack a standard way to remove this bias when creating labeled datasets for model training [3].

This study is focused on physiological signals, such as heart rate variability, skin conductance, and measurements of autonomous nervous system activity. People think physiological data could provide insights about emotional phenomena because the autonomic nervous system (ANS) plays a crucial role in regulating bodily responses to emotional stimuli. For instance, physiological signals such as heart rate variability and skin conductance reflect the ANS's activity, which changes with different emotional states. By monitoring these signals, researchers can gain a deeper understanding of how emotions manifest in the body [4]. Training affect prediction systems on this type of signals has gained significant attention due to its potential application in various fields, including mental and physical healthcare and human-computer interaction; it could impact even governmental policy creation criteria concerning the case of personal data usage in intelligent agents training or in large language models [5, 6, 7].

The aim of this research is to explore the extent to which IRA influences the performance of automatic affect prediction systems in the context of physiological datasets. By understanding this relationship, we can improve the reliability and validity of evaluating affect prediction systems and enhance their practical utility in real-world applications, such as healthcare and human-computer interaction. This type of research implies a two-step process. First, a systematic identification and review of the existing literature describing datasets and their characteristics concerning interrater agreement on the affective state interpretation from physiological signals. This stage established a reliable foundation for the second step: analysing the model performance reported in technical papers utilizing these datasets. Due to a nine-week time limitation, the second step was left for future research. Therefore, only the first five subquestions from the following list are fully addressed in this paper, while the sixth is left for future research:

Q1: *What types of affective states have been targeted by datasets?* Emotion, mood and reaction are a few examples of state types under the category of "affect"; the annotation task depends on their exact definition.

Q2: *What different affect representation schemes have been used in these datasets and what is the motivation for using them?* Different representation schemes result in different labels; identifying a trend in scheme usage would provide insight into the most targeted emotions in physiological datasets.

Q3: *What data is available regarding interrater agreement (presence, number of raters, types of measures, level of agreement)?* If these characteristics tend to differ between datasets, there might exist a bias towards analysing the model performance, since evaluation depends on the labeling process.

Q4: *Is there a change in how datasets measure interrater agreement over time?* Seeing a change of measure over time while the other factors stay the same would be an argument in favour of the correlation between the chosen interrater agreement measure and model performance.

Q5: *Is there a relationship between the affect representation scheme used by datasets and their interrater agreement?* The level of agreement might be influenced by the vagueness of the affect representation scheme, since annotators have varying number of labels to choose from.

Q6: *Is there a relationship between the interrater agreement in datasets and the empirical performance of affect prediction systems using them for training and evaluation?* This directly expresses the importance of objectivity in labeling processes on human data. Due to time limitations, this falls out of the scope of this project, but some indications on how to proceed in future work will be provided.

Starting with section 2, more details about the subtopics of interest are discussed. In section 3, the paper describes the methodology used to conduct the study and how it is divided in the dataset extraction step and the literature review step. Section 4 refers to results per subtopic building up to the main research statement, while section 5 explains how the study was conducted in a responsible manner. Following this, section 6 discusses the results for different subquestions. Last but not least, section 7 provides a conclusion for the research project and how it provides ground findings for possible future work. For extra information, an appendix is available as well.

2 Background

In affective computing, affect representation schemes are methodologies used to encode and describe human emotions, typically through dimensional models (e.g., valence-arousal)[8], categorical models (e.g., Ekman's basic emotions)[9], appraisal models [10], facial action coding systems [11], or multimodal approaches, allowing for accurate emotion recognition and response in computational systems [12]. These diverse representation schemes significantly

impact the analysis of automatic affect prediction models by influencing the granularity, interpretability, and complexity of the predictions, thereby affecting the model's performance and applicability in real-world scenarios [13].

The assessment of these systems' performance often relies on interrater agreement (IRA) measures, which quantify the level of agreement among human annotators in labeling emotional states. IRA measures are statistical tools used to evaluate the consistency or consensus among different raters who observe and categorize the same phenomena. Common measures include Cohen's kappa, which adjusts for chance agreement [14], and Krippendorff's alpha, which is applicable for various data types and handles missing data effectively [15]. For physiological datasets, the annotators can have both internal and external perspectives: they can either be study participants themselves and label their own signals, or they can be third-party annotators, labeling signals produced by other people; this difference might influence affect labeling trends. Other aspects which require some inspection are the frequency of agreement measures inclusion in the datasets and the variability of agreement measures (to see which are most often chosen for physiological datasets). Despite their importance, the impact of IRA measures on the evaluation and interpretation of affect prediction systems remains unclear [16, 17].

3 Methodology

This study employs a systematic review methodology following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to comprehensively survey existing literature on datasets used for automatic affect prediction. A systematic review is a rigorous and structured approach to identifying, evaluating, and synthesizing all available evidence relevant to a research question [18, 19, 20]. PRISMA provides a structured approach to ensure transparency and completeness in reporting systematic reviews, enhancing the reproducibility and reliability of the study findings [21]. The study involves the following key steps, illustrated in Figure 1:

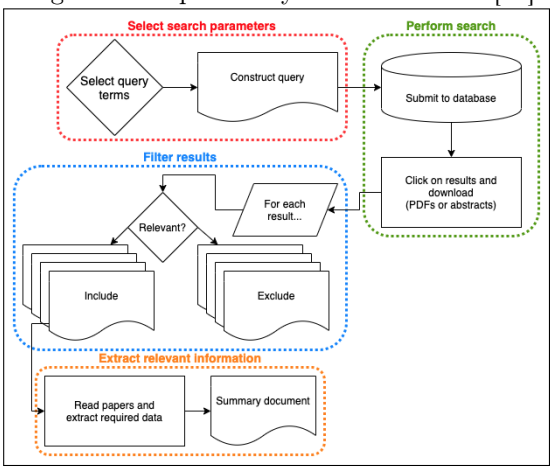
1. **Searching:** Define the research question and develop an exhaustive search strategy to identify relevant studies.

2. **Filtering:** Create a protocol for extracting papers describing datasets used for automatic affect prediction which respect a set of inclusion and exclusion criteria. Screen studies based on the protocol to select those meeting the research objectives.

3. **Extraction:** Extract data and relevant information from selected studies in a systematic manner: create a dataframe which summarizes the dataset papers and allocates them to matching subtopics. This step leads to the analysis of the main research statement.

With these three aspects in regard, subsection 3.1 presents the search strategy employed including the choice of search engines and the query-based process of paper selection, subsection 3.2 discusses the eligibility criteria, as well as the feasibility criteria, while subsection 3.3 shows search results.

Figure 1: Steps in a Systematic Review [19]



3.1 Search Strategy

The search strategy was designed to retrieve all potentially relevant studies. It includes defining search terms, selecting appropriate databases, and detailing the search syntax and operators used.

It was opted for search engines that are avaliable to students, especially the ones for which TU Delft offers institutional access, and that are well known for providing accurate results in the main field of this research, namely Computer Science, as well as interconnected areas which might help understand the physiological and psychological aspects of the subtopics [22]. Therefore, taken into account were Scopus [1], IEEE Xplore [2] and Web Of Science [3]. Other search engines which could have been taken into consideration are ACM Digital Library [4] and Google Scholar [5], but they were disregarded from this study because of their limitations; Google Scholar has few options to limit or narrow search results, users cannot for example limit results to peer reviewed, full text materials or subject, while ACM's coverage is wide-ranging but not comprehensive (the chosen engines provide more reliability).

The complete paper selection process involved the following steps:

1. Formulation of tailored search queries for each engine, incorporating the keywords listed in Table 1, with according wildcards for derivations.
2. Collection of non-duplicate papers retrieved from the searches.
3. Screening of titles to determine relevance based on the eligibility criteria.
4. Screening of abstracts of retained papers to assess suitability for inclusion based on eligibility criteria.
5. Full-text assessment of the remaining papers to determine final inclusion or exclusion based on eligibility criteria.

[1] https://www.scopus.com/search/form.uri?display=basic
[2] https://ieeexplore.ieee.org/Xplore/home.jsp
[3] https://www.webofscience.com/wos/author/search
[4] https://dl.acm.org
[5] https://scholar.google.com

Queries for each search engine were formulated based on a combination of relevant keywords related to physiological emotion recognition. These keywords were identified through an iterative process and are listed in Table 1. The combination of keywords consists of disjunctions applied to different terms describing each topic (applied per column) and conjunctions between topics (applied between resulting disjunctions). The keywords are then modified by adding wildcards, such that all similar forms of the initial word are found.

Table 1: Keywords used in search query for each concept

Physiological	**Emotion**	**Recognition**	**Dataset**	**Raters**
physiological	emotion	recognition	dataset	rater
biosignal	affect	detection	database	interrater
biometric	feeling	identification	corpus	inter-rater
nervous system	mood	classification		inter rater
anatomic	reaction	analysis		
somatic		prediction		

Queries were constructed to ensure that papers mentioning these concepts in titles, abstracts, or keywords were captured. The only part which is set to be searched for in the entire paper body is the one related to interrater agreement, since it is relevant to this research, but might not be the main point of discussion in the papers describing the datasets. The process of finding a query which captures the desired subjects in integrity, while accounting for the feasibility factor of the study, was logged in Appendix C. It can be observed that, over time, the motivation behind each update on the query became more specific in order to increase reproducibility and understanding of the literature review process (eg. removing "sentiment analysis" since it would provide vague results compared to the other types of data and removing "affected" from the affect family, since it introduces unwanted papers which are disease related instead of affect related). The queries were adapted for the particularities of each engine, while trying to maintain the balance between broadness and accuracy of search terms, which is also visible in Appendix C.

By excluding the terms related to "rater", papers about datasets which do not include IRA measures were obtained, thus improving this study according to the motivation of the inclusion criteria in Table 3 (more details in subsection 3.2). The search process of non-IRA datasets was performed in parallel to the one focusing on IRA for the following reason: stricter feasibility criteria was necessary mostly in the former case, where approximately 24.000 results matched the query, exceeding the time limited capability of scanning (more details in subsection 3.2.1).

Therefore, the resulting papers from each search engine are presented in table 2. The last two columns represent the simplest form of filtering which could already be done before screening, namely applying the feasibility restriction of "data papers" to the non-IRA papers and removing non-English papers. English papers describing non-English datasets were kept, since understanding the aggregated details from the presentation paper matters more to this survey than the data itself. These restrictions were applied by checking the "data type" and "language" filtering options provided by search engines, which consider the tags and keywords of the papers. All papers were then imported to Mendeley [6] to follow a deduplication process, resulting in a total of 90 (54 with IRA, 36 without) candidates for screening.

[6] https://www.mendeley.com/search/

Table 2: Summary of Dataset Papers from Different Sources

Source	**Query With IRA**	**Query Without IRA**	**Filtered Query With IRA**	**Filtered Query Without IRA**
Scopus	52	17075	50	26
Web of Science	8	6354	8	26
IEEE Xplore	1	577	1	0

3.2 Filtering Process

This step involves defining eligibility criteria and screening retrieved studies against the criteria set. These criteria may include publication type, relevance to the research question, and dataset characteristics.

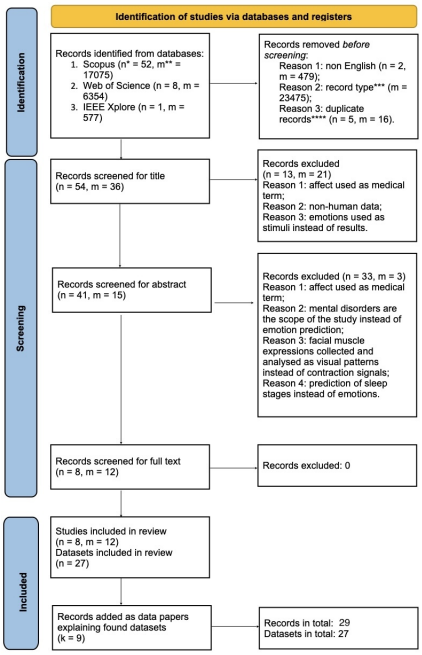
The eligibility criteria consists of inclusion and exclusion factors, as mentioned in the renewed 2020 PRISMA guidelines [21]. This ensures that the scope of the literature review is narrowed down to subjects mentioned in the research statement (and subtopics) through the inclusion of desired topics, while also maintaining reproducibility through the exclusion of papers which are not relevant at the moment of this research. The inclusion and exclusion criteria do not have to be met at once for two main reasons: first, some of them only cover information which could be used for a subset of the subtopics, but not for all (eg. a paper discusses the correlation between different affect representation schemas and the model performance, but does not include interrater agreement methods). Second, the main interest is in papers about datasets which include information about interrater agreement measures, but the papers which do not include it could also provide interesting insights about the prevalence of the method in this field. For this reason, both categories were considered, keeping in mind the limitations in subsection 3.2.1.

A brief motivation for each criterion is mentioned in Table 3. Besides this, papers focusing on interrater agreement measures and guidelines for systematic review were read in order to define a proper review protocol which is clear and reproducible and to ensure that responsible research is conducted. These papers are not part of the survey itself and therefore are not mentioned as inclusion criteria.

The following screening steps are described in more detail in Figure 2, according to the PRISMA 2020 flow diagram template [21]. Some specific reasons why papers got excluded from the study during screening which were not part of the eligibility criteria initially are related to the medical interpretation of the word "affect" instead of the psychological one and the presence of emotions used as stimuli instead of results.

The screening stages resulted in 20 records and 27 datasets to be used for data extraction, which is presented in section 3.3. Nine additional data papers were manually added to improve the reliability of results, since some of the most valuable datasets for this study were mentioned only as secondary data sources in the systematically retreived papers. A possible reason why they did not appear as results of the search query is that they were not suitable for all topics required to analyse the main research statement, but are still proper candidates for specific subquestions. This indicates that in future research, where the time limitation is not an issue, the study can be improved by considering more niched queries for each subtopic separately to make sure this type of papers is not omitted.

Figure 2: Adapted PRISMA 2020 flow diagram for systematic reviews which included searches of databases and registers only [21]



*n refers to papers mentioning IRA
**m refers to papers not mentioning IRA
***record type filtering applied only to non-IRA to papers (see section 3.2)
****duplicates removed through Mendeley (see section 3.2)

3.2.1 Feasibility Limitations

Due to the 9 week time limitation to this research project, a few feasibility criteria were derived in order to achieve the goal of answering the main question in a reproducible manner, while leaving space for improvements in future work. The main choice was to leave the second thread of the two-stage plan mentioned in section 1 as material for future work and focus only on the first stage: the systematic identification and review of the relevant datasets.
Another choice was to split the search in two parallel threads, one for papers which mention IRA and one for the opposite, leading to a slightly adapted PRISMA flow diagram in figure 2. The main goal of this research involves the presence of IRA in the datasets and its connection to the model performance, while assessing the overall presence of IRA in datasets in the first step of the two-stage plan is merely an insight for the research practices in this area of study. Also, the search for non-IRA datasets resulted in approximately 24,000 matches, exceeding the time-limited capacity for detailed scanning. Thus, stricter feasibility criteria were necessary, leading to the application of the "data type" filter of the search engines to narrow results to "data papers." This step ensured a manageable subset for analysis.

Table 3: Inclusion and Exclusion Criteria

Inclusion Criteria	**Motivation**	**Application Method**
Papers focusing on physiological emotion recognition which use different affect representation schemes.	To ensure relevance to the research topic and understand the types of data.	Title and abstract screening.
Papers presenting datasets with physiological signals used for emotion detection, which include interrater agreement measures.	Useful for interrater agreement analysis given certain dataset features.	Query modification: conjugate with rater related terms.
Papers presenting datasets with physiological signals used for emotion detection, which fail to include interrater agreement measures.	Useful to assess the impact of missing agreement measures and to comment about the prevalence of using these measures in the field of Computer Science.	Query modification: do not use rater related terms.
Exclusion Criteria	**Motivation**	**Application Method**
Papers not written in English.	Ensures language consistency for review comprehensibility.	Search engine language filter option.
Papers reliant solely on secondary data sources (datasets that are mentioned, but not created by the authors of the paper).	To prioritize primary research studies (created by the authors of the paper) which explain the reasoning behind collecting the data.	Abstract and full text screening.
Papers published after April 2024.	Ensures relevance within the designated timeframe of the review.	Search engine date filter option.
In the case where IRA is present, papers about datasets which **do not** mention the type of affect annotations (self-reported or third-party).	Data about both experienced and perceived affect is needed to see how study participants describe the feeling themselves and how external parties interpret the feeling of the participants.	Abstract and full text screening.
Studies which do not provide information on the participants or subjects involved in the dataset production.	Evaluation of sample characteristics and generalizability is difficult/impossible without this information, as well as potential biases in the data (predominant genetics, conditions which do not characterize the general population).	Abstract and full text screening.
Studies targeting non-human datasets.	Would be interesting to analyse, but the comparison between human and non-human data is infeasible given the time limitation.	Abstract and full text screening.

Table 4: Summary of Datasets and Related Information

Dataset	Data Paper	Related Papers	Year	Affect Representation Scheme	Interrater Agreement Measure	No. Raters	Self-reported vs. Third Party Annotations	Physiological Signal Covered	Stimuli
EMOEEG	[23]	-	2017	VA	Cohen's kappa	8	Self-reported	EEG, EOG, EMG, ECG, EDA	Audiovisual
NAA	[24]	-	2017	Aggression levels, fear (VAD)	Krippendorff's α	15	3rd Party	ECG, EDA, EMG	Dyadic interaction
RAMAS	[25]	-	2018	Anger, Sadness, Disgust, Happiness, Fear, Surprise (VAD)	Krippendorff's α	21	3rd Party	EDA	Dyadic interaction
RECOLA	[26]	[27]	2013	Positive, Negative, Neutral (VA)	Cronbach's α	6	3rd Party	EDA, ECG	Remote Dyadic interaction
Proprietary	[28]	-	2010	Anger, Disgust, Fear, Anxiety, Sadness, Desire, Calm and Happiness	Fleiss'kappa	3	3rd Party	EDA, BVP	Game environments
PhyMER	[29]	-	2018	Angry, Happy, Sad, Surprised, Neutral, Fear, Disgusting (VA)	Cronbach's α, $Fleiss'kappa$	28	Self-reported	BVP, EDA, ECG, skin temp	Video
I DARE	[30]	-	2014	VA	Fleiss'kappa	63	Self-reported	EEG, SC, PPG, EMG, and ET	Images
AMIGOS	[31]	[30]	2021	control, familiarity, liking, Angry, Happy, Sad, Surprised, Neutral, Fear, Disgusting (VA)	Cronbach's α	44	Self-reported (emotion labels, 40), 3rd Party (VA, 4)	EEG, ECG, EDA	Video
ASCERTAIN	[32]	[30]	2018	Linking, Engagement, Familiarity, VA	Cohen's kappa	58	Self-reported	ECG, EEG, EDA	Video
BIO-VID-EMO	[33]	[30, 13]	2016	amusement, sadness, anger, disgust, fear	-	-	Self-reported	SC, ECG, tEMG	Video
DEAP	[34]	[30, 13]	2012	Familiarity, Like or Dislike, VAD	Fisher's method	32	Self-reported	EEG, Peripheral	Music Videos
DREAMER	[35]	[30]	2018	amusement, excitement, happiness, calmness, anger, disgust, fear, sadness, surprise, VAD	coefficient of variation (CV)	25	Self-reported	EEG, ECG	Audiovisual
MAHNOB-HCI	[36]	[30, 13]	2012	Neutral, Anxiety, Amusement, Sadness, Joy, Disgust, Anger, Surprise, and Fear (VADP)	Cohen's kappa	27	Self-reported	EEG, EDA, respiration, amplitude, ECG, Eye Gaze	Audiovisual
MPED	[37]	[30]	2018	Joy, funny, disgust, anger, fear, sad, neutrality	-	23	Self-reported	EEG, ECG, EDA, RSP	Video
SEED	[38]	[30, 13]	2016	Positive, Negative, Neutral (Absolute)	-	-	-	EEG, Eye Gaze	Video
DAPPER	[39]	-	2021	upset, hostile, ashamed, nervous, afraid, alert, inspired, determined, attentive, active	-	-	-	ECG, EDA	daily environment
K-EmoCon	[40]	-	2020	20 categories	Krippendorff's α	37	Self-reported, 3rd Party and **Partner**	EEG, ECG, BVP, EDA, skin temp	Dyadic interaction
Proprietary	[41]	-	2018	pleasant, energetic, tense, angry, fearful, happy, sad, tenderness	-	27	-	EEG, fMRI	Music
BIRAFFE2	[42]	-	2022	VA	-	102	Self-reported	ECG, EDA, accelerometer	Game environments
Proprietary	[43]	-	2021	VA	-	29	Self-reported	ST, EDA, BVP, acceleration	Indoor environment
Proprietary	[44]	-	2023	Positive, negative, neutral	-	346	Self-reported	ECG, EDA, RESP	Driving
ECSMP	[45]	-	2021	neutral, fear, sad, happy, anger, and disgust	-	89	Self-reported	EEG, ECG, EDA, TEMP, accelerometer	Video
Emognition	[46]	-	2022	amusement, awe, enthusiasm, liking, surprise, anger, disgust, fear, sadness, VA	ANOVA	43	Self-reported	EEG, BVP, HR, EDA, SKT, ACC, GYRO	Video
G-REx	[47]	-	2024	VA, uncertainty	-	190	Self-reported	EDA, PPG	Video
NeuroBioSense	[48]	-	2024	Joy, Surprise, anger, disgust, sadness, fear, and neutral	-	58	Self-reported	PPG, EDA, TEMP	Advertising
AKTIVES	[49]	-	2023	Reaction / no reaction	Majority	3	3rd Party	BVP, EDA, ST	Video
DECEiVeR	[50]	-	2024	neutral, calm, tiredness, tension, excitement, VA	-	11	Self-reported	ECG, EDA, RESP, EMG	Acting

Legend: V - Valence, A - Arousal, D - Domination, P - Potency, ECG - electrocardiogram, EEG - electroencephalogram, EOG - electrooculogram, EMG - electromyography, EDA - electrodermal activity, BVP - blood volume pulse, skin temp or ST or SKT - skin temperature, PPG - photoplethysmography, ET - endothelin (vasoconstrictor peptide), SC - skin conductance, RSP - respiration, fMRI - functional magnetic resonance imaging, HR - heart rate, ACC - accelerometer, GYRO - internal gyroscope

3.3 Data Extraction

In order to analyse the particularities needed to answer each subquestion, data was extracted from the final set of 29 records and 27 datasets included in review and stored in Mendeley. An Excel [7] dataframe was created in order to keep track of relevant information from each dataset paper, as shown in table 4, which is the main source of information for all the results that follow in section 4.

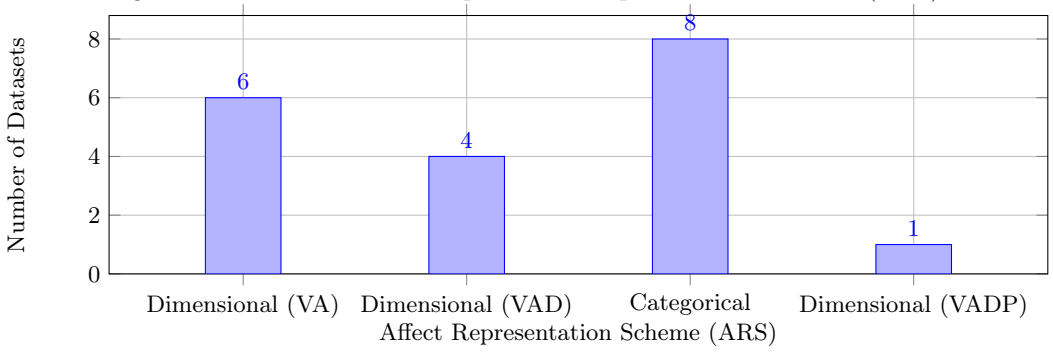
4 Results

This section explains the main findings based on the datasets which fit each of the subquestions and shows why only subsets of the datasets are fit for each subquestion. Subsection 4.1 handles the first two subquestions, subsection 4.2 handles the third, subsection 4.3 looks at subquestion four, while subsection 4.4 answers the fifth subquestion.

4.1 Targeted Affect and Affect Representation Schemes (ARS)

The datasets summarized in Table 4 primarily focus on various emotions, with some also considering mood and reactions. For instance, the DEAP [34], MAHNOB-HCI [36], and AMIGOS [31] datasets target emotions such as happiness, sadness, anger, and fear. The EMOEEG [23] and PhyMER [29] datasets, on the other hand, target broader emotional categories, including neutral states. Moods, typically represented as positive, negative, or neutral states, are less frequently targeted but are integral in datasets like AMIGOS [31] and ASCERTAIN [32]. These datasets link emotions to personality traits, providing a richer context for understanding affective responses. For example, AMIGOS [31] takes into account both mood and personality, although only emotions are measured in the experiments. Similarly, ASCERTAIN [32] connects emotions to personality, highlighting the interplay between transient emotional states and enduring personality traits. Reactions, such as those captured in the AKTIVES [49] and DAPPER [39] datasets, are another form of affective state. These reactions are often context-specific, reflecting immediate responses to particular stimuli, such as games or daily environments.

Figure 3: Number of Datasets per Affect Representation Scheme (ARS)



Note: V - Valence, A - Arousal, D - Dominance, P - Potency

[7] https://www.microsoft365.com/launch/excel

The ARS used across these datasets vary widely, as seen in Figure 3, reflecting different theoretical approaches to categorizing emotions. The most prevalent scheme is the circumplex model, which represents emotions along the dimensions of valence (pleasantness) and arousal (activation level). This model is used in datasets like EMOEEG [23], DEAP [34], and MAHNOB-HCI [36], providing a continuous representation of affective states that is useful for capturing the nuances of emotional experiences. Other common ARS include the Six Basic Emotions model [9], which categorizes emotions into fear, anger, disgust, sadness, happiness, and surprise, as seen in the Proprietary and RAMAS [25] datasets. This categorical approach is straightforward and aligns with foundational theories in psychology. Additionally, some datasets utilize adapted models, such as the one used in the K-EmoCon dataset [40], which incorporates 20 categories of emotions and considers both self-reported and third-party annotations. This comprehensive approach allows for a more detailed and multifaceted understanding of affective states.

The motivation for using these different ARS lies in their ability to capture various aspects of emotional experiences. Dimensional models like the circumplex model are favored for their ability to represent the continuous and interrelated nature of emotions [51], while categorical models are valued for their simplicity and ease of interpretation [9]. The choice of ARS often depends on the specific goals of the study and the nature of the stimuli used.

4.2 Quantifying Interrater Agreement (IRA)

Interrater agreement (IRA) is a crucial factor in the reliability and validity of affective computing datasets, as it measures the consistency among different raters'annotations. The datasets mentioned in Table 4 offer comprehensive data regarding the presence, number of raters, types of measures, and levels of agreement across various datasets.

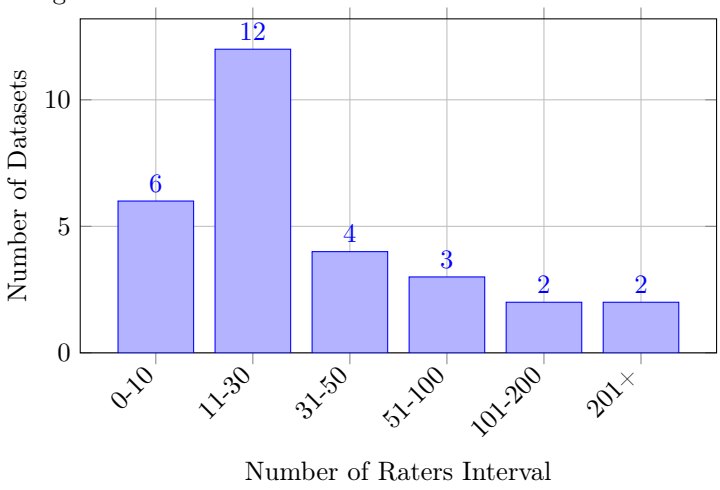
The majority of datasets include an IRA measure, indicating the importance placed on the reliability of annotations. Commonly used IRA measures in these datasets include Cohen's kappa [14], Krippendorff's alpha [15], Fleiss'kappa [52], Cronbach's alpha [53], and ANOVA [54]. Each of these measures is suited to different types of data and agreement scenarios. For instance, Cohen's kappa is used for binary or categorical data, while Cronbach's alpha is typically employed for continuous labels, providing insight into the internal consistency of the annotations.

The number of raters varies significantly across datasets, ranging from as few as 3 to as many as 346, as seen in Figure 4.2. This variation can impact the reliability of the IRA measure, as a higher number of raters can provide a more robust estimate of agreement. For example, the PhyMER dataset [29] utilized 28 evaluators to label stimulus videos, with subsequent labeling by participants. This two-stage labeling process helps verify the induced emotions'accuracy and the participants'felt emotions, providing a comprehensive IRA measure. In PhyMER [29], Cronbach's alpha for valence and arousal was exceptionally high at 0.97 and 0.96, respectively, while Fleiss'kappa for categorical annotations was moderate at 0.40.

Notably, datasets such as AMIGOS [31] and ASCERTAIN [32], which incorporate both self-reported and third-party annotations, use measures like Cronbach's alpha and Cohen's kappa to ensure the reliability of both types of annotations. This dual approach helps to cross-validate the data, enhancing the overall reliability of the dataset.

The use of different IRA measures and the number of raters highlights the diversity in dataset annotation methodologies. For example, the RECOLA dataset [26], which focuses on remote dyadic interactions, uses Cronbach's alpha with 6 raters, ensuring a high level

Figure 4: Number of Datasets in Each Interval of Raters



of agreement for continuous labels. On the other hand, the DEAP dataset [34] employs Fisher's method [55] with 32 raters, providing a robust measure of agreement for self-reported annotations on music videos.

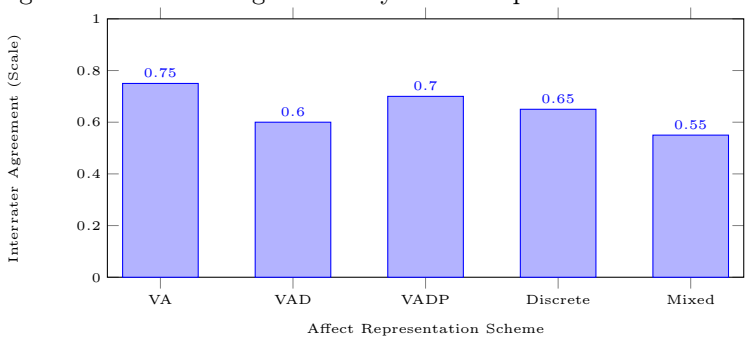
4.3 IRA Over Time

Examining the datasets summarized in Table 4 reveals trends and changes in the IRA measures over time, offering a perspective on their evolution and the factors influencing their adoption. In the early 2010s, the IRA measures predominantly used were classical statistical methods such as Cohen's kappa and Fleiss'kappa. For instance, the RECOLA dataset (2013) [26] employed Cronbach's alpha, a measure typically used to assess the reliability of psychometric instruments. The Proprietary dataset from 2010 [28] also used Fleiss'kappa, another robust measure for multiple raters. These early choices reflect a preference for established, well-validated methods in measuring agreement among annotators.

As the field progressed, a diversification in IRA measures became apparent. By the mid to late 2010s, datasets like NAA (2017) [24] and RAMAS (2018) [25] adopted Krippendorff's alpha, known for its ability to handle various data types and missing values, making it more versatile for complex annotation tasks. This shift suggests an increasing awareness and need for flexibility in IRA measures to accommodate diverse annotation schemes and data characteristics. In recent years, there has been a notable shift towards more dynamic and comprehensive approaches. The AMIGOS dataset (2021) [31] uses both Cronbach's alpha and Fleiss'kappa, indicating a hybrid approach to capture different aspects of agreement. Similarly, the PhyMER dataset (2018) [29] incorporates both Cronbach's alpha and Fleiss'kappa, reflecting a nuanced understanding of IRA that blends multiple measures to enhance reliability assessments.

Moreover, the introduction of dynamic annotations, where continuous or time-variant data is annotated, necessitates more sophisticated IRA measures. Datasets like DEAP (2012) [34] and DREAMER (2018) [35] use methods like Fisher's method and the coefficient of variation (CV), which cater to the dynamic nature of the data. This trend underscores a growing recognition of the limitations of static IRA measures in capturing the complexities

Figure 5: Interrater Agreement by Affect Representation Scheme



Note: V - Valence, A - Arousal, D - Dominance, P - Potency, Discrete: Discrete Emotion Categories, Mixed: Combination.

of temporal and dynamic annotations, leading to the adoption of more advanced statistical methods.

4.4 Relationship Between IRA and ARS

The investigation into the relationship between ARS and IRA measures reveals some trends across various datasets presented in Table 4 and summarised in Figure 4.4. A clear pattern emerges when examining datasets utilizing the VA scheme. Datasets such as EMOEEG [23], which employ VA, show a moderate level of interrater agreement with Cohen's kappa as the measure (k = 0.61-0.80 generally considered substantial agreement). Similarly, the RECOLA dataset [26], also using VA but with Cronbach's alpha, exhibits a substantial agreement among raters. These findings suggest that the VA scheme, despite its simplicity, allows for consistent annotation, likely because of the reduced cognitive load on annotators, who only need to assess two dimensions. In contrast, datasets employing more detailed affect representation schemes, like RAMAS [25] and AMIGOS [31], show varied levels of interrater agreement. RAMAS, with its complex VAD scheme, utilizes Krippendorff's alpha, achieving moderate agreement. The increased complexity and number of dimensions in the VAD scheme can introduce vagueness, as annotators may interpret the dimensions differently, reducing consistency. Datasets like MAHNOB-HCI [36], which use an extended VADP (Valence-Arousal-Dominance-Potency) scheme, still maintain substantial interrater agreement (Cohen's kappa). However, this dataset benefits from a high number of raters (27), which can mitigate individual bias and variability, enhancing overall agreement. This suggests that while more detailed schemes can lower agreement due to their complexity, the impact can be counterbalanced by increasing the number of raters. Datasets employing discrete emotion categories present varied results. The PhyMER dataset [29], encompassing multiple discrete emotions, shows substantial agreement with Cronbach's alpha and Fleiss'kappa. In this case, the discrete categories may provide clear, distinct labels, reducing ambiguity compared to continuous dimensions. However, datasets like DREAMER [35], which also use a combination of discrete emotions and VAD, report lower consistency (coefficient of variation), highlighting that the complexity introduced by combining schemes can affect agreement negatively.

5 Responsible Research

Responsible research practices are crucial for maintaining the integrity and ethical standards of scientific investigations. This section examines both methodological rigor and ethical considerations in this project.

5.1 Reflection upon Methodology

The systematic literature review aimed for thoroughness and reproducibility, adhering to established guidelines. Detailed documentation was maintained at every step, yet conducting the review with one student may have introduced potential errors, particularly during data extraction and analysis. To minimize bias, the data extraction protocol was followed strictly, although human error remains a possibility. The selection criteria for datasets might also introduce bias, despite efforts to be comprehensive. Future research should involve multiple reviewers to cross-verify data, enhancing the reliability and validity of findings and providing a more robust understanding of the relationship between affect representation schemes and interrater agreement.

5.2 Ethical viewpoint of physiological affect prediction

Physiological affect prediction poses significant ethical concerns, including potential misuse in surveillance or behavior manipulation without consent. In general, this raises serious privacy issues, as individuals may be unaware of their physiological data being collected and analyzed, but all datasets chosen for this project mentioned that participants agreed on their data being processed in this scope.

The subjectivity in labeling physiological responses can lead to biased datasets, exacerbating fairness and equity issues in affective computing systems. Such biases may arise from cultural, gender, age, or other demographic differences, leading to inaccurate affective state representations. Ensuring transparency and respect for user privacy is essential. The chosen datasets were inclusive and representative to mitigate biases, and informed consent was obtained from participants with clear information on data use. Researchers should consider the broader implications of their work and design systems that promote fairness and equity. Addressing these ethical considerations will help develop reliable and ethical physiological affect prediction systems that respect individual rights and positively contribute to society.

6 Discussion

This study's use of a systematic review methodology adhering to PRISMA[21] guidelines effectively ensures a comprehensive and transparent examination of datasets for automatic affect prediction. By systematically searching, filtering, and extracting relevant studies, the methodology allows for a rigorous assessment of existing literature. PRISMA enhances reproducibility and reliability, crucial for validating the study's findings. The detailed search strategy, careful application of eligibility criteria, and structured data extraction collectively support the study's aim to identify key trends and gaps in current research, ultimately contributing to the advancement of affective computing. The approach is thorough, yet adaptable for future refinements.

IRA is a critical factor in the reliability of annotations, with common measures including Cohen's kappa[14], Krippendorff's alpha[15], Fleiss' kappa[52], and Cronbach's alpha[53].

The number of raters varies widely across datasets, impacting the robustness of IRA measures. For instance, PhyMER[29] employs a two-stage labeling process with 28 evaluators, achieving high Cronbach's alpha values. Datasets like AMIGOS[31] and ASCERTAIN[32], which incorporate both self-reported and third-party annotations, use multiple IRA measures to ensure data reliability. The evolution of IRA measures over time shows a shift from classical methods like Cohen's kappa to more versatile measures like Krippendorff's alpha. This transition reflects an increasing need for flexibility in handling diverse annotation schemes. Recent datasets adopt hybrid approaches to capture different aspects of agreement, as seen in AMIGOS and PhyMER.

A clear pattern emerges: simpler ARS like VA tend to yield higher interrater agreement due to reduced ambiguity. For example, datasets such as EMOEEG[23] and RECOLA[26], which use VA, demonstrate substantial agreement. In contrast, more complex schemes, such as VAD and VADP, show varied levels of agreement, which can be mitigated by increasing the number of raters, as observed in MAHNOB-HCI[36]. Datasets using discrete emotion categories, like PhyMER, also achieve substantial agreement, suggesting that clear, distinct labels can enhance consistency.

7 Conclusion and Future Work

This analysis reveals diverse approaches in affect representation schemes (ARS) and interrater agreement (IRA) across datasets. Simpler ARS like VA yield higher IRA, as seen in datasets like EMOEEG[23] and RECOLA[26], while more complex schemes like VAD and VADP show varied agreement levels, mitigated by a higher number of raters. This trend underscores the importance of ARS selection in ensuring reliable annotations. For future work, examining the impact of ARS and IRA on model performance could provide deeper insights into optimizing affective computing models. Additionally, replicating this study without current limitations, such as restricted dataset access and varying annotation protocols, would offer more robust conclusions. Ensuring a standardized approach in future studies could further clarify the relationship between ARS complexity, IRA consistency, and model accuracy, ultimately enhancing the reliability of affective computing systems.

A Acknowledgements

I would like to express my sincere gratitude to my supervisor and my group peers for their invaluable guidance and support throughout this project. Their insightful feedback and encouragement have been instrumental in the successful completion of this work.

B Large Language Model Usage

ChatGPT [8] was involved only to improve the paper writing process, but was not involved in the actual research procedures. The main usage case was styling figures and tables in LaTeX [9] and summarizing the big dataset table into smaller, more concise ones.

Prompt example:

[8] https://openai.com/chatgpt/
[9] https://latex.org/forum/

Make a plot in latex with the binned number of raters from Table A (the dataset table) on the X axis and the number of datasets corresponding to those bins on the Y axis.

Answer:

```
\begin{figure}
\centering
\caption{Number of Datasets in Each Interval of Raters}
\begin{tikzpicture}
\begin{axis}[
    ybar,
    symbolic x coords={0-10, 11-30, 31-50, 51-100, 101-200, 201+},
    xtick=data,
    x tick label style={rotate=45, anchor=east},
    nodes near coords,
    ymin=0,
    xlabel={Number of Raters Interval},
    ylabel={Number of Datasets},
    width=10cm,
    height=6cm,
    bar width=1cm,
    enlarge x limits={abs=1cm},
    ylabel near ticks,
    xlabel near ticks,
    grid=major,
    ymajorgrids=true,
    xmajorgrids=true
]

\pgfplotstableread{
interval datasets
0-10 6
11-30 12
31-50 4
51-100 3
101-200 2
201+ 2
}\datatable

\addplot table[x=interval,y=datasets] \datatable;

\end{axis}
\end{tikzpicture}
\end{figure}
```

C Query Building Process

Query	**Results on Scopus**	**Notes**
(physiolog* OR biosignal* OR biometric* OR nervous system OR anatomic* OR somatic*) AND (emotion* OR affect* OR feel* OR mood* OR sentiment* OR react*) AND (recogni* OR detect* OR identif* OR classif* OR analys* OR predict*)	580,806	Most general, no filtering applied.
(physiolog* OR biosignal* OR biometric* OR nervous system OR anatomic* OR somatic*) AND (emotion* OR affect* OR feel* OR mood* OR sentiment* OR react*) AND (recogni* OR detect* OR identif* OR classif* OR analys* OR predict*)	20,849	Basic filtering applied: Computer Science articles only, English only.
(physiolog* OR biosignal* OR biometric* OR nervous system OR anatomic* OR somatic*) AND (emotion* OR affect* OR feel* OR mood* OR react*) AND (recogni* OR detect* OR identif* OR classif* OR predict*)	16,025	Removed sentiment analysis since it would provide vague results compared to the other types of data. If limited to data papers, gives **22 results**.
(dataset* OR database*) AND (physiolog* OR biosignal* OR biometric* OR nervous system OR anatomic* OR somatic*) AND (emotion* OR affect* OR feel* OR mood* OR react*) AND (recogni* OR detect* OR identif* OR classif* OR predict*)	3,279	Instead of limiting the search to "data papers", include dataset related terms –> focus on purpose of the papers instead of the labels scopus gives them.
annotat* AND (dataset* OR database*) AND (physiolog* OR biosignal* OR biometric* OR nervous system OR anatomic* OR somatic*) AND (emotion* OR affect* OR feel* OR mood* OR react*) AND (recogni* OR detect* OR identif* OR classif* OR predict*)	194	Include annotation related terms –> focus only on supervised (labeled) learning models, ignore unsupervised since they don't give any information on the human perspective.

Query	Results on Scopus	Notes
annotat* AND (dataset* OR database*) AND (physiolog* OR biosignal* OR biometric* OR nervous system OR anatomic* OR somatic*) AND (emotion* OR affect* OR feel* OR mood* OR react*) AND (recogni* OR detect* OR identif* OR classif* OR predict*) AND NOT affected	172	Remove "affected" from the affect family –> it introduces unwanted papers which are disease related instead of affect related.
rater AND annotat AND (dataset* OR database*) AND (physiolog* OR biosignal* OR biometric* OR nervous system OR anatomic* OR somatic*) AND (emotion* OR affect* OR feel* OR mood* OR react*) AND (recogni* OR detect* OR identif* OR classif* OR predict*) AND NOT affected	5	Check how many papers include rater information. Since the number is very low, broadening of other subjects is needed.
rater AND (annotat OR label* OR supervised) AND (dataset* OR database*) AND (physiolog* OR biosignal* OR biometric* OR nervous system OR anatomic* OR somatic*) AND (emotion* OR affect* OR feel* OR mood* OR react*) AND (recogni* OR detect* OR identif* OR classif* OR predict*) AND NOT affected	6	Extended domain for annotation.
rater AND (annotat OR label* OR supervised) AND (dataset* OR database*) AND (physiolog* OR biosignal* OR biometric* OR nervous system OR anatomic* OR somatic*) AND (emotion* OR affect* OR feel* OR mood* OR react*) AND (recogni* OR detect* OR identif* OR classif* OR predict*) AND NOT affected	9	Remove limitation to Computer Science –> include papers that are published in journals for other disciplines while still providing the information needed.
FINAL QUERIES BEFORE SCANNING		

Query	Results on Scopus	Notes
TITLE-ABS-KEY((dataset* OR database* OR corpus) AND (physiolog* OR biosignal* OR biometric* OR nervous system OR anatomic* OR somatic*) AND (emotion* OR affect* OR feel* OR mood* OR react*) AND (recogni* OR detect* OR identif* OR classif* OR predict*) AND NOT affected) AND ALL((interrater OR inter-rater OR rater* OR inter rater) AND agreement)	**Scopus: 52 including interrater section, 17.075 excluding it. Web of Science: 8 including interrater section, 6.354 excluding it.**	**No filters applied. Add "corpus" to database section. Look for "interrater" related terms in full paper body because they may not be the main point of the paper and still exist. Add "agreement" to all data to make the interrater search more specific. Remove "annotated" related terms because they are included in the interrater family implicitly.**
ALL(emotion recognition OR affective computing OR emotion classification OR emotion prediction OR affect prediction) AND physiological signal* AND (dataset* OR database* OR corpus) AND (interrater OR inter-rater OR inter rater OR rater*)	**IEEE Xplore: 1 including interrater section, 577 excluding it**	**No filters applied. Keep the wildcard number limited due to engine restrictions.**

References

[1] R. W. Picard, *Affective computing*. MIT press, 2000.

[2] R. Cheruvalath, "Artificial intelligent systems and ethical agency," *Journal of Human Values*, vol. 29, no. 1, pp. 33–47, 2023.

[3] F. Cabitza, A. Campagner, and M. Mattioli, "The unbearable (technical) unreliability of automated facial emotion recognition," *Big Data & Society*, vol. 9, no. 2, 2022.

[4] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biological Psychology*, vol. 84, no. 3, pp. 394–421, 2010.

[5] S. De Nadai, M. D'Incà, F. Parodi, M. Benza, A. Trotta, E. Zero, L. Zero, and R. Sacile, "Enhancing safety of transport by road by on-line monitoring of driver emotions," in *Proceedings of the 2016 11th System of Systems Engineering Conference (SoSE)*, (Kongsberg, Norway), pp. 1–4, 2016.

[6] R. Guo, S. Li, L. He, W. Gao, H. Qi, and G. Owens, "Pervasive and unobtrusive emotion sensing for human mental health," in *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare*, (Venice, Italy), pp. 436–439, 2013.

[7] B. Verschuere, G. Crombez, E. Koster, and K. Uzieblo, "Psychopathy and physiological detection of concealed information: A review," *Psychologica Belgica*, vol. 46, pp. 99–116, 2006.

[8] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, p. 261 – 292, 1996. Cited by: 1025.

[9] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, p. 169 – 200, 1992. Cited by: 6082.

[10] K. Scherer, *Appraisal Considered as a Process of Multilevel Sequential Checking*, vol. 92, pp. 92–120. 05 2001.

[11] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[12] G. Smith and J. Carette *What lies beneath—a survey of affective theory use in computational models of emotion*, Jun 2022.

[13] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, p. 2074, Jun 2018.

[14] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[15] K. Krippendorff, *Content analysis: An introduction to its methodology*. Sage Publications, 2004.

[16] K. A. Hallgren, "Computing inter-rater reliability for observational data: An overview and tutorial," *Tutorials in Quantitative Methods for Psychology*, vol. 8, no. 1, pp. 23–34, 2012.

[17] K. L. Gwet, *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[18] M. G. Cherry, A. Boland, and R. Dickson, *Doing a systematic review: A student's guide*. SAGE Publications, 2024.

[19] S. Goldfarb-Tarrant, A. Robertson, J. Lazić, T. Tsouloufi, L. Donnison, and K. Smyth, "Scaling systematic literature reviews with machine learning pipelines," 10 2020.

[20] J. Smith and A. Brown, *Systematic Reviews in Psychology: A Practical Guide*. Oxford University Press, 2019.

[21] C. Sohrabi, T. Franchi, G. Mathew, A. Kerwan, M. Nicola, M. Griffin, M. Agha, and R. Agha, "Prisma 2020 statement: What's new and the importance of reporting guidelines," *International Journal of Surgery*, vol. 88, p. 105918, Apr 2021.

[22] "Tudelft databses homepage." `https://databases.tudl.tudelft.nl/?f=EEMCS&d=CS&t=&q=&y0=research%20data&y1=reference&y2=reports&y3=articles&y4=standards&y5=e-books&y6=patent%20information&y7=statistics&y8=educational%20resource&y9=theses&y10=e-journals&y11=catalogue`. [Accessed 20-06-2024].

[23] A.-C. Conneau, A. Hajlaoui, M. Chetouani, and S. Essid, "Emoeeg: A new multimodal dataset for dynamic eeg-based emotion recognition with audiovisual elicitation," in *25th European Signal Processing Conference (EUSIPCO)*, (Kos, Greece), pp. 738–742, IEEE, August 28-September 2 2017.

[24] I. Lefter, C. Jonker, S. Klein Tuente, W. Veling, and S. Bogaerts, "Naa: A multimodal database of negative affect and aggression," in *Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII2017)*, (United States), pp. 21–27, IEEE, 2017. ACIIW 2017 : 7th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos ; Conference date: 23-10-2017 Through 26-10-2017.

[25] M. Konstantinova, V. Ivanova, D. Vlasov, V. Filippov, S. Malykh, and A. Karpov, "Ramas: The russian acted multimodal affective set for affective computing and emotion recognition studies," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, (San Antonio, TX, USA), pp. 204–210, IEEE, September 3-6 2018.

[26] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, (Shanghai, China), pp. 1–8, IEEE, April 22-26 2013.

[27] L. A. Bugnon, R. A. Calvo, and D. H. Milone, "Dimensional affect recognition from hrv: An approach based on supervised som and elm," *IEEE Transactions on Affective Computing*, vol. 11, pp. 32–44, January-March 2020.

[28] A. Drachen, S. N. Yannakakis, J. H. Smith, and A. Kokkinakis, "An automated approach to estimate player experience in game events from psychophysiological data," in *Proceedings of the 6th International Conference on the Foundations of Digital Games*, (Monterey, CA, USA), pp. 270–277, ACM, June 19-21 2010.

[29] H. Liu, Z. Zhang, B. Schuller, and J. Liu, "Phymer: Physiological dataset for multimodal emotion recognition with personality as a context," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI)*, (Boulder, CO, USA), pp. 211–218, ACM, October 16-20 2018.

[30] M. Bilucaglia, M. Zito, A. Fici, C. Casiraghi, F. Rivetti, M. Bellati, and V. Russo, "I dare: Iulm dataset of affective responses," *Frontiers in Human Neuroscience*, vol. 18, 2024.

[31] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 479–493, 2021.

[32] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "Ascertain: Emotion and personality recognition using commercial sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2018.

[33] L. Zhang, S. Walter, X. Ma, P. Werner, A. Al-Hamadi, H. C. Traue, and S. Gruss, ""biovid emo db": A multimodal database for emotion analyses validated by subjective ratings," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–6, 2016.

[34] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis ;using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[35] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, 2018.

[36] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[37] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, and Z. Cui, "Mped: A multi-modal physiological emotion database for discrete emotion recognition," *IEEE Access*, vol. 7, pp. 12177–12191, 2019.

[38] W.-L. Zheng and B.-L. Lu, "A multimodal approach to estimating vigilance using eeg and forehead eog," *Journal of Neural Engineering*, vol. 14, 11 2016.

[39] X. Shui, M. Zhang, Z. Li, X. Hu, F. Wang, and D. Zhang, "A dataset of daily ambulatory psychological and physiological recording for emotion research," *Scientific Data*, vol. 8, 06 2021.

[40] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Scientific Data*, vol. 7, Sept. 2020.

[41] I. Daly, N. Nicolaou, D. Williams, F. Hwang, A. Kirke, E. Miranda, and S. Nasuto, "Neural and physiological data from participants listening to affective music," *Scientific Data*, vol. 7, 06 2020.

[42] K. Kutt, D. Drążyk, L. Żuchowska, M. Szelążek, S. Bobek, and G. Nalepa, "Biraffe2, a multimodal dataset for emotion-based personalization in rich affective game environments," *Scientific Data*, vol. 9, p. 274, 06 2022.

[43] N. Gao, M. Marschall, J. Burry, S. Watkins, and F. Salim, "Understanding occupants' behaviour, engagement, emotion, and comfort indoors with heterogeneous sensors and wearables," *Scientific Data*, vol. 9, 06 2022.

[44] Q. Meteier, M. Capallera, E. Salis, L. Angelini, S. Carrino, M. Widmer, O. Abou Khaled, E. Mugellini, and A. Sonderegger, "A dataset on the physiological state and behavior of drivers in conditionally automated driving," *Data in Brief*, vol. 47, p. 109027, 03 2023.

[45] Z. Gao, X. Cui, W. Wan, W. Zheng, and Z. Gu, "Ecsmp: A dataset on emotion, cognition, sleep, and multi-model physiological signals," *Data in Brief*, vol. 39, p. 107660, 12 2021.

[46] S. Saganowski, J. Komoszyńska, M. Behnke, B. Perz, D. Kunc, B. Klich, L. Kaczmarek, and P. Kazienko, "Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables," *Scientific Data*, vol. 9, 04 2022.

[47] P. Bota, J. Brito, A. Fred, P. Cesar, and H. Plácido da Silva, "A real-world dataset of group emotion experiences based on physiological data," *Scientific Data*, vol. 11, 01 2024.

[48] B. Kocacinar, P. İnan, E. Zamur, B. Çalşimşek, F. Patlar Akbulut, and C. Catal, "Neurobiosense: A multidimensional dataset for neuromarketing analysis," *Data in Brief*, vol. 53, p. 110235, 02 2024.

[49] B. Coskun, S. Ay, D. Erol Barkana, H. Bostanci, I. Uzun, A. Oktay, B. Tuncel, and D. Tarakci, "A physiological signal database of children with different special needs for stress recognition," *Scientific Data*, vol. 10, 06 2023.

[50] L. Aly, L. Godinho, P. Bota, G. Bernardes, and H. Plácido da Silva, "Acting emotions: A comprehensive dataset of elicited emotions," *Scientific Data*, vol. 11, p. 147, 01 2024.

[51] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 12 1980.

[52] F. Moons and E. Vandervieren, "Measuring agreement among several raters classifying subjects into one-or-more (hierarchical) nominal categories. a generalisation of fleiss' kappa," 2023.

[53] M. Tavakol and R. Dennick, "Making sense of cronbach's alpha," *International Journal of Medical Education*, vol. 2, pp. 53–55, 06 2011.

[54] L. St»hle and S. Wold, "Analysis of variance (anova)," *Chemometrics and Intelligent Laboratory Systems*, vol. 6, no. 4, pp. 259–272, 1989.

[55] R. Elston, "On fisher's method of combining p-values," *Biometrical Journal*, vol. 33, pp. 339 – 345, 01 1991.