Exploring the Underlying Emotional Models in Emotion Recognition Systems with Electrodermal Activity

Tomas A. D'Amelio[1,2] *, Lorenzo A. Galán[2] *, Emmanuel A. Maldonado[2], Agustín A. Díaz Barquinero[2], Enzo Tagliazucchi[1,3]†, Denis A. Engemann[4]†

[1] Departamento de Física, Universidad de Buenos Aires and Instituto de Física Interdisciplinaria y Aplicada (INFINA-CONICET), CABA, Argentina

[2] Facultad de Psicología, Universidad de Buenos Aires, Buenos Aires, Argentina.

[3] Latin American Brain Health Institute (BrainLat), Universidad Adolfo Ibanez, Santiago, Chile

[4] Roche Pharma Research and Early Development, Neuroscience and Rare Diseases, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Basel, Switzerland

Author Note

Tomás A. D'Amelio http://orcid.org/0000-0001-7947-2024

Lorenzo A. Galán https://orcid.org/0009-0002-4652-5411

Emmanuel A. Maldonado https://orcid.org/0009-0004-9170-2127

Agustín A. Díaz Barquinero

Enzo Tagliazucchi https://orcid.org/0000-0003-0421-9993

Denis A. Engemann https://orcid.org/0000-0002-7223-1014

Corresponding author: Tomás A. D´Amelio, Departamento de Física, Universidad de Buenos Aires and Instituto de Física Interdisciplinaria y Aplicada (INFINA-CONICET), Facultad de Ciencias Exactas y Naturales, Pabellón I, Ciudad Universitaria, Buenos Aires 1428, Argentina.

Email: dameliotomas@gmail.com

* These authors contributed equally to this work.

† These authors also contributed equally to this work.

Abstract

Affective computing is an interdisciplinary field that aims to automatically recognize and interpret emotions. Recent research has focused on using physiological signals (e.g., electrodermal activity) to improve emotion recognition. However, little attention has been paid to the theoretical emotion models underlying these systems. Here, we conducted a systematic review and meta-analysis of the literature on automatic emotion recognition systems using electrodermal activity. We found that models predicting arousal generally outperformed those predicting valence, which is consistent with our pre-registered hypothesis. This finding aligns well with the conceptual framework that views arousal as a psychological and physiological state linked to autonomic nervous system activity, making it more directly related to electrodermal activity. Furthermore, we observed a discrepancy between the types of machine learning models used, mainly classification models, and the emotional models adopted, often of a dimensional nature. Specifically, despite the increased use of dimensional affective models, there has been no corresponding increase in the use of regression models, which would be consistent with the continuous nature of these data. We conclude that a comprehensive understanding of affective states requires consideration of both psychological and computational perspectives in affective computing research.

Keywords: affective computing, emotion recognition, electrodermal activity, emotion models, systematic review, meta-analysis

Introduction

Affective computing, which emerged in the 1990s, is a growing interdisciplinary field aiming at incorporating emotions into artificial intelligence (Picard, 1999). Researchers in this field combine affective science, computer science and engineering methods not only to advance the scientific understanding of human emotions but also to develop technologies that can effectively operate in emotionally nuanced contexts (Brigham, 2017; Calvo et al., 2015; Calvo, 2010). A particular interest lies in the development of methods and algorithms for automatically recognizing and interpreting affective states (Picard, 2000). These advancements in affective computing hold significant promise, opening up transformative applications across a range of fields such as healthcare (Yannakakis, 2018), education (Aylett & Paiva, 2012; Yadegaridehkordi et al., 2019), and self-driving cars (Alharbi et al., 2020).

From an affective science perspective, the study of emotions in affective computing is based on various theoretical models, which can be broadly divided into two types: categorical and dimensional models. A prominent historical example is Ekman's (1992, 2002) categorical model, which includes six basic emotions (i.e. happiness, sadness, fear, anger, disgust, and surprise). In recent years, dimensional models based primarily on Russell's (1980, 2003) circumplex model of affect have received increased attention. This model represents affective states as a sequence of numbers representing the intensity of different dimensions that characterize these states, such as valence (positive to negative affective states) and arousal (physiological level of activation). Unlike categorical models, in which participants choose an emotion label from a set of options, dimensional models typically require participants to rate emotions on dimensional scales (e.g., valence on a 1-9 scale).

The application of machine learning in affective computing has opened new avenues for understanding human emotions. Not only does it enable the creation of sophisticated and accurate emotion recognition systems, but it also paves the way for more nuanced analysis

(Lei & Gratch, 2023). Such advances have profound implications for a variety of applications in psychology and affective science, enriching our understanding of emotional experiences (Calvo & D'Mello, 2010; D'Mello et al., 2018). For example, machine learning has been successfully applied to the complex problem of predicting affective states (Shu et al., 2018; Zeng et al., 2007). Most of the publications focused on supervised learning, which operates by training on a set of known data and corresponding labels, essentially learning the underlying relationships between them (Alpaydin, 2020). The resulting models are practical for summarizing complex psychophysiological datasets into scores and can be used for making predictions on new data points.

Supervised methods can be subdivided into classification and regression techniques. A classification problem is a type of predictive modeling task where the objective is to assign a label (or class) from a predefined set to a new observation based on its features (or attributes). Classification becomes particularly salient when applying categorical models of emotion, like Ekman's. For example, emotion recognition from facial expressions often operates as a classification task, where models are trained on datasets of facial expressions labeled with their corresponding emotions—happiness, sadness, anger, and so forth. These trained models can then classify emotions in new, previously unseen facial images.

Alternatively, a regression problem is a type of predictive modeling task where the goal is to predict a continuous output variable. These are especially useful when leveraging dimensional models of emotion, such as Russell's circumplex model. Here, the goal is to predict the intensity or degree of an affective dimension; for instance, a prediction model might estimate a continuous value for arousal intensity based on certain input parameters (see Figure 1).

[Figure 1 here]

However, it is worth noting that the use of classification techniques is not strictly confined to categorical models of emotion. Often, classification is employed even when the underlying scales originate from dimensional models, such as classifying emotions into 'high valence' and 'low valence' categories. While this approach may offer some practical advantages, it essentially discretizes what is fundamentally a continuous variable, potentially leading to a loss of information. Therefore, when the underlying emotional variables are originally understood as continuous, leveraging regression techniques could provide a more nuanced understanding and make full use of the available data.

Equally critical as the choice between classification and regression is the methodological decision regarding which types of input data to use for training affective computing models. Affective computing models have been trained to predict human emotions using various inputs, including facial expressions (Said & Barr, 2021), posture (Huang et al., 2021), speech (Atmaja et al., 2022), thermography (Clay-Warner & Robinson, 2015), and text (Guo, 2022). Recently, there has been a growing interest in utilizing physiological signals because they are readily available, believed to be less sensitive to social and cultural variability (Jang et al., 2014), and allow for continuous sampling of participants' physiological activity (Sharma et al. 2019).

Physiological signals can be broadly divided into central and peripheral measures. Central measures, such as electroencephalograms (EEG), focus on capturing brain activity, while peripheral measures, including electrocardiography (ECG), electromyography (EMG), and electrodermal activity (EDA), record activity from the peripheral nervous system. Importantly, peripheral measures are closely linked to emotional responses because they record interoceptive feedback from various systems-including the stress, neuroendocrine, immune, and gastrointestinal systems-that strongly influence our subjective emotional experiences (Pace-Schott et al., 2019).

EDA holds unique value as a peripheral measure in emotion recognition. Renowned for its ease of use and accessibility (Babaei et al., 2021), EDA is widely regarded as a robust proxy for emotional arousal, as it is innervated by the sympathetic branch of the autonomic nervous system (Boucsein, 2012). Formerly known as the galvanic skin response, EDA measures changes in skin's electrical properties triggered by the activity of sweat glands, specifically the sudomotor neurons (Boucsein, 2012). These glands, while primarily functioning in thermoregulation, are especially active in the palms and soles during states of high emotional arousal, regardless of valence (Sato et al., 2020). This includes both negative emotions such as fear or stress, and positive emotions like excitement or joy. However, the relationship between EDA and valence—whether the emotional experience is positive or negative—is more complex and requires further exploration. Therefore, EDA serves as a reliable indicator primarily of arousal in affective science (Boucsein, 2012) and for detecting affective dimensions in affective computing (Sánchez-Reolid et al., 2022).

Recently, substantial efforts have been directed towards enhancing the predictive capabilities of emotion recognition models using Electrodermal Activity (EDA) signals, as highlighted in reviews by Posada-Quintero & Chon (2020) and Shukla et al., (2019). These advancements have primarily focused on feature extraction techniques involving various domains such as time, frequency, and time-frequency (Shukla et al., 2019). However, despite these methodological improvements, there has been a notable gap in the literature concerning the underlying emotional models employed in EDA-based emotion recognition systems.

Addressing a critical but often overlooked aspect, our research conducts an in-depth systematic review and meta-analysis with a primary focus on investigating the underlying emotion models used in EDA-based emotion recognition systems. Our review also examines the characteristics of self-reported emotion models, the sample, and the techniques used for emotion stimulation. In addition, we evaluate machine learning models and various EDA parameters, such as equipment and locations, commonly used in automatic emotion recognition tasks. Finally, we aim to establish a link between the affective science and affective computing literatures by comparing the performance of arousal and valence prediction models using EDA. Based on the established relationship between arousal, autonomic nervous system activity, and EDA, we hypothesize that arousal prediction models would typically outperform valence prediction models.

Methods

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocol (Liberati et al., 2009, Page et al., 2022) is used in this systematic review to ensure the replicability of the study (http://prisma-statement.org).

Preregistration

In recent years, there has been a significant increase in interest and concern for reproducibility in science (Munafò et al., 2017). To follow good open science practices and avoid publication bias, the planned study was registered in the Open Science Framework (OSF) web repository prior to implementation (see https://osf.io/zbqm6). For example, the hypothesis that arousal prediction models would perform better than valence prediction models was preregistered in this document.

Eligibility Criteria

In our review, we included studies that used machine learning techniques for emotion recognition using EDA. For the purposes of this review, a machine learning model is defined as the use of statistical models to estimate functions algorithmically for making predictions or decisions. This estimation aims to minimize generalization error for better out-of-sample prediction. We incorporate simple linear models, such as penalized logistic regression, as well as advanced techniques like deep learning models, which possess the added capability of learning features automatically. Each unique model from every article was included as a separate instance for evaluation if they differed in any of the variables of interest in the review, such as dataset, algorithm, or type of output. Conversely, if models differed only in feature construction or selection, we considered only the best-performing model, as feature manipulation is outside the scope of this review. However, to avoid biased selection, all models from the selected papers were included into the meta-analysis. Finally, our study was restricted to English-written articles that examined non-clinical human samples.

Exclusion criteria included master's and doctoral theses, book chapters as non-peer-reviewed literature, reviews, meta-analyses, commentaries, workshops, descriptions and abstracts. Studies that included only models trained on multimodal signals (e.g., EDA and heart rate variability) were also excluded. However, if different models were trained in the same paper, and one or more used only EDA features, they were included for further analysis. Finally, papers that created software tools (unless they were emotion recognition software and tested according to the criteria established in this review) were not included in this review.

Search Strategy and Selection Process

Scopus and PubMed were used to identify journal articles, conferences, and preprints published between January 1, 2010 and December 31, 2020. If articles exist in both versions, preference was given to publications that have been peer-reviewed after preprint. The flowchart from this review can be found in Figure 2.

[Figure 2 here]

A comprehensive search strategy was employed in the Pubmed and Scopus databases to exhaustively retrieve all relevant papers that satisfy our inclusion criteria. The following keywords and Boolean operators were used: “EDA” OR “Electrodermal Activity” OR “GSR” OR “Galvanic Skin Response” OR “Skin conductance” OR “SCR” OR “SCL” AND “Emotions” OR “Emotion” OR “Affective” AND “Recognition” OR “Decoding” OR “Detection” OR “Classification” OR “Regression”. The search terms were applied to the title, abstract, and keywords. The advanced search criteria included: “between 2010 - 2020; journal, conference proceeding as source type; article, conference paper as document type; and English as language”. This search was conducted on February 21, 2021.

To ensure the comprehensiveness of our search, an iterative process was adopted whereby the search strategy was refined and re-run until all known relevant papers were identified in the search results. Additionally, we manually screened the reference lists of included papers to identify further potential studies not captured by our search strategy, reducing the likelihood of missing relevant research.

Once the papers were identified, they were exported to the Rayyan tool (Ouzzani et al., 2016) to streamline the initial selection process. The investigators (cf. section *Author Contributions*) divided the total number of papers equally among themselves and eliminated duplicates. Each of these authors checked the titles, abstracts, and keywords for discrepancies based on the inclusion and exclusion criteria, and double-checked the classifications made by the other authors, which were randomly assigned. The number of excluded papers and the reasons for their exclusion were documented in Rayyan.

The remaining papers were then exported to Zotero (Idri, 2015), a free open-source software and web management tool. The investigating author read the full papers, discarding those that did not meet the inclusion criteria. During multiple readings, specific dimensions were identified, and papers that did not fulfill these dimensions were discarded at each step.

Data Collection Process

For the papers that met the criteria, a subset of the authors was designated to conduct the data extraction. Each designated author retrieved the data from all the recognition models found in their assigned articles. These models were classified into nine categories for further analysis (see Data Elements section). Subsequently, a second designated author, different from the first, reviewed the classifications to ensure consistency and accuracy. Any disagreements or conflicts in the classifications were addressed in a joint meeting with a third author, who served as a mediator if necessary.

Data Elements

A shared database was created among the authors, where each paper was classified according to the following criteria: metadata (data on the authors, title, year of publication, journal in which it was published, type of article, country of affiliation of the first author); type of data (original or from database, if the latter, specify which); participants (relevant characteristics of the sample: size, gender, age range, country of origin, etc.); affective stimulation technique (type of affective stimulus used, exposure time); self-report (if used, what type used, emotion model used); EDA (equipment used, location of sensors); statistical learning models (output of model used, type of model); emotion model and performance of emotion recognition models with EDA. This database is openly available in a GitHub repository (see the Supplementary Material section).

Study Risk of Bias Assessment

Each paper in each round was double-checked by two different authors for discrepancies in classification and inclusion using the Rayyan and Zotero tools. If discrepancies existed, they were noted and resolved with a third member of the team.

Meta-analysis

In conducting a meta-analysis to compare the performance of valence and arousal recognition models in emotion recognition, we followed a comprehensive, systematic protocol consisting of several key steps.

Our first step was to define the scope of our meta-analysis by establishing strict inclusion and exclusion criteria. A critical requirement was that studies use accuracy as a measure of performance-an indicator that is widely used across models. We prioritized studies that used binary classification models that produced distinct output categories (i.e., low vs. high valence, and low vs. high arousal). Because our goal was to conduct a comparative analysis of arousal versus valence,, we only included studies that examined models of both dimensions, hence, excluding those that were limited to a single affective dimension. All models trained, tested and reported in each of the publications were included for further analysis.

We extracted data from the selected studies. Details such as the machine learning model used and performance scores were carefully transcribed to ensure that data for arousal and valence models were clearly separated.

Permutation and bootstrap tests were used for null-hypothesis significance testing and uncertainty estimation, respectively, for the observed differences in accuracy between arousal and valence. These tests were chosen for their simplicity and robustness to assumptions about the underlying data distribution. The number of resamples for both permutation and bootstrap tests was kept at the software's default of 9999.

Finally, we conducted multiple regression analyses using both ordinary least squares (OLS) and robust regression with Huber's T norm. These analyses were conducted to explore the influence of sample size, publication year and mean accuracy on the differences in accuracy.

Software

Quantitative analyses were conducted using the Python programming language (Van Rossum y Drake, 2009), supplemented by specialized libraries such as NumPy (van der Walt, Colbert, & Varoquaux, 2011), SciPy (Virtanen et al., 2020), Matplotlib (Hunter, 2007), pandas (McKinney, 2010), and statsmodels (Seabold & Perktold, 2010). Results will be summarized in narrative, tabular, or graphical form.

Results

Screening and Selection

Our systematic review procedure, outlined in Figure 2, includes the detailed enumeration of discarded articles, the respective reasons for their exclusion, and the final selection of articles retained for further analysis. Through rigorous evaluation, we refined our corpus to a set of 99 studies, providing a comprehensive landscape of 499 different emotion prediction models for our subsequent in-depth analysis.

Data Analysis

Geographical Distribution and Source of Publications

The predominance in the distribution of articles lies with China, contributing 16 articles. This is followed by researchers in Germany, India, and the USA, each providing 7 articles. Thes distribution of articles from research teams in Germany is noteworthy for its breadth, with papers sourced from multiple groups, signaling a robust variety in research interests. Subsequently, 5 articles were contributed by researchers in Turkey, Italy, and Malaysia. Within Italy and Malaysia, a significant portion of the articles came from specific research groups; for example, a single group in Italy produced 4 out of 5 papers, demonstrating this particular group's leading position within the Italian scientific community. Similarly, in Malaysia, the entire body of 5 articles came from one research group, indicating an even greater level of focus.

Researchers in Spain and Iran each produced 4 articles, while Switzerland, Romania, Pakistan, Taiwan, and Greece are represented by 3 articles each. Countries including Japan, Austria, Tunisia, Macedonia, Portugal, Korea, and Finland are represented with two articles each. A variety of other countries complete the distribution(see Figure 3).

[Figure 3 here]

On the continental scale, Asia stands at the forefront, contributing the majority of articles, totaling 49. This is followed by Europe, with a contribution of 39 articles. The Americas offered 9 articles, and Africa and Australia, though less prolific, rounded off the diversity with their contributions of 3 and 1 articles respectively.

We also analyzed the sources of these academic papers. The total number of papers was almost equally divided between journals and congresses (50 and 48, respectively), with journals taking a slight lead. One preprint was also included in our study.

Of the journal-based papers, most were published in engineering and computer science journals. The journal "Sensors" published the most of these articles (n=7), followed by "IEEE Access" and "IEEE Transactions on Affective Computing" with 4 and 3 articles, respectively. "Scientific Reports, Studies in Health Technology and Informatics", and “Frontiers in ICT” each published two of the included articles. The remaining 30 journal articles were evenly distributed among several other journals in related fields.

Data type

We categorized the databases used in the studies into two types: "restricted" databases, which were limited to the use of a specific study, and "open" databases, which were accessible to the public. Out of the 100 recorded database usages (with one study utilizing two databases), 61% utilized restricted databases while the remaining 39% used open databases. Nonetheless, a trend towards open databases has been observed over time, as illustrated in Figure 4A.

Regarding the open databases, the majority (87%) were pre-established databases, 10% were newly created databases publicly accessible, while the remaining 3% were available upon request. Among the publicly accessible databases, 27% were derived from DEAP (Koelstra et al., 2011), while both AMIGOS (Miranda-Correa et al., 2021) and MAHNOB (Soleymani et al., 2011) represented 21%. Other databases were used less frequently, as shown in Figure 4B. Please see Table 1 for further database details.

[Figure 4 here]

Participants

For the studies reviewed, we observed that the average sample size was 49 participants, ranging from a minimum of 4 to a maximum of 457. It's noteworthy that 10% of the studies did not report their sample size, and they were excluded from meta-analysis consequently.

When considering gender representation in these samples, we noted that 30% of the papers did not report the female portion of their sample. For the papers that did, on average, 26 females were included, with a range from 1 to 236.

Age of participants was unreported in 46% of the papers. When reported, the average participant age was 28 years, with the range spanning from 1 to 65 years old. Of the papers that reported average participant age, 41% did not provide the age range. In those that did, the mean age was 26 years, with a range from 19.44 to 36.1 years.

The geographical diversity of the participants was also considered, though only 22% of the papers (n= 22) specified the country of origin of the sample. In these cases, Malaysia (n=5), China (n=3), and Iran (n=3) were the most frequently reported.

Self report

In relation to affective questionnaires utilized in the studies, we found that 49% did not use or specify any affective questionnaires. On the other hand, 46% utilized and specified affective questionnaires, and 5% used questionnaires from other studies.

Within the studies that specified affective questionnaires, Self-Assessment Manikin (SAM; Bradley & Lang, 1994) was the most commonly used, applied in 70% of these studies. The Positive and Negative Affect Schedule (PANAS; Watson, Clark y Tellegen, 1988) was employed in 23% of these cases. Less frequently used questionnaires included the Affective grid (Russell, Weiss, Mendelsohn, 1989), Perceived Stress Scale (PSS; Cohen, Kamarch, & Mermelstein, 1983), and Differential Emotion Scale (DES; Gross & Levenson, 1995).

The large amounts of missing information in the studies, particularly concerning participant demographics and the use of affective questionnaires, are concerns that we will further address in the discussion section.

Relationships between Emotional Categories and Dimensions

Emotional categorization and dimensional representation are two primary perspectives employed in affective research. In our review, we found a variety of emotional categories represented in studies using self-report measures. Importantly, the language used to describe these emotional categories mirrors the terminology employed in the original studies we examined. These terms ranged from specific emotional states like 'Disgust' and 'Fear' to more general experiences or conditions such as 'Pleasant' and 'Stress.' The most frequently cited categories included emotional states such as 'Disgust,' 'Fear,' and 'Sadness,' with 'Neutral' and 'Surprise' also being commonly used, as depicted in Figure 5A. Other categories appeared less frequently in the literature. use.

[Figure 5 here]

Emotional dimensions were also observed in studies incorporating self-report measures. The six dimensions identified were: ‘Valence’, ‘Arousal’,’Dominance’, ‘Predictability’, ‘Preference’, and ‘Familiarity’. Arousal and valence dominated the field, with dominance following closely. See Figure 5B for a visual representation.

To provide a more granular perspective, we generated two graphs showcasing relationships between various categories and dimensions, illustrated in Figures 5C and 5D. It is crucial to understand these relationships as they shape the comprehensive view of emotional states in the reviewed studies.

Emotion Elicitation Techniques

Emotion elicitation plays a central role in affective research. Our findings suggest that standardized elicitation techniques were used in 19% of the studies reviewed. Only six standardized techniques were found, in order of most frequent use: IAPS (International Affective Picture System; Huang & Chiang, 2014), TSST (Trier Social Stress Test; Allen et al., 2017), Stroop color-word interference test (SCWT; Stroop, 1992), Rapid-ABC play protocol (Ousley et. al., 2012), International Affective Digitized Sound (IADS; Bradley & Lang, 2007), and Robin (Morreale et al., 2012).

We also found a significant preference for multimodal strategies, with 62% of studies using such methods. Among specific techniques, video was the most common (45%), followed by music (18%) and pictures (11%). Active participant involvement was reported in 78% of studies.

Electrodermal Activity (EDA)

Among the studies we reviewed, we identified 25 distinct EDA devices, with 2% of studies employing custom-made devices. Biosemi ActiveTwo, Shimmer, and BIOPAC were the most commonly used devices (see Figure 6).

The placement of electrodes was reported in 36% of cases. When reported, the left side was preferred for electrode placement (62%). The hands were the most common site for electrode placement (83%), specifically in the middle (43%) and index (36%) fingers. Figure 7 provides a detailed overview of electrode placement. It is important to note that specifying the location of electrode placement is crucial for methodological rigor and facilitates the replication of studies, ensuring consistent and comparable results across research.

Model performance

Our analysis tracked the progression and distribution of affective models (dimensional or categorical) and algorithm types (classification or regression) used across studies. It is worth noting that affective models are used to conceptualize human emotions in either discrete categories (e.g., happiness, sadness, anger) or continuous dimensions (e.g., arousal, valence). Similarly, the types of algorithms can vary, with classification algorithms predicting discrete outcomes and regression algorithms predicting continuous outcomes.

Over time, we observed an initial preference for categorical models during the first five years (excluding 2012), followed by a shift towards a dominance of dimensional models (refer to Figure 8A). Additionally, classification analysis clearly predominated the publication landscape, accounting for approximately 90% of the total number of models, whereas regression analysis was performed in a minority of publications (Figure 8B).

An intriguing aspect we observed was the extent to which studies ventured into a psychological or physiological interpretation of their results, rather than limiting their discussion to the performance of the model. Out of 99 papers, only 23 offered such insights.

Meta-analysis

Our meta-analysis, conducted according to predefined inclusion and exclusion criteria, analyzed 76 arousal models and an equal number of valence models from 12 different studies. The analysis revealed a statistically significant advantage of arousal models over valence models (*Mean difference* = 3.56%, 95% CI [2.30%, 5.47%]; $p < 0.001$), as determined by permutation and bootstrap analyses.

In addition, the Bland-Altman plot (see Figure 9) showed a negative correlation between mean accuracy and the interdimensional performance difference (i.e., arousal minus valence). This implies that as mean accuracy increases, the performance difference between arousal and valence tends to decrease.

Our statistical models, both ordinary least squares (OLS) and robust regression, were significant, yielding adjusted R-squared values of 0.392 and 0.416, respectively. These models identified sample size, publication year, and mean accuracy as salient predictors of the observed difference in accuracy between the arousal and valence models.

Discussion

In the present work, our objectives were twofold, as we conducted a systematic review and meta-analysis focusing on the confluence of affective science and affective computing in the context of emotion recognition systems using EDA. First, we sought to systematically review methodological aspects of 99 studies published between 2010 and 2020, such as emotion elicitation techniques and affective annotation tools. Second, we selected a subset of 76 pairs of arousal and valence models from 12 studies for meta-analysis according to our pre-specified inclusion and exclusion criteria. In the following sections, we discuss methodological considerations and qualitatively integrate these findings.

Discrepancy between Emotion Modeling and Machine Learning Approaches

This section aims to highlight a notable discrepancy in emotion recognition systems with EDA: although there is an increasing trend toward the use of dimensional models for affective variables like valence and arousal, the machine learning approaches employed in the field have not adapted in parallel to capture the continuous nature of these emotional dimensions.

Our analysis reveals a surge in the use of dimensional emotion models in the affective computing literature. This shift signifies a growing recognition of emotions as complex, multidimensional constructs, rather than discrete, categorical entities (Barrett, 2017). Simultaneously, our review sought to investigate whether this trend towards dimensional emotion modeling was mirrored by an increase in the adoption of regression models, an approach well-suited for capturing the continuous nature of emotional data. However, contrary to our expectations, we found that the use of regression models has not increased alongside the growing adoption of dimensional emotion models. Interestingly, this pattern seems consistent with findings from other domains of emotion recognition research; Saganowski et al. (2022) also observed in their comprehensive review of emotion recognition systems from physiological signals that there is a scarcity of studies tackling regression tasks, indicating that the preference for classification over regression extends beyond EDA signals to other peripheral physiological signals as well.

This discrepancy suggests a bias in the field towards simplifying dimensional emotional data into categories (e.g. 'positive' or 'negative'). Such simplification may not only be a performance optimization strategy, but may also be driven by the specific application goals of the technologies being developed. For example, a virtual agent designed to detect low valence in a user in order to generate a predefined response may require simplifying emotional data into more actionable categories.

However, if we want to use affective computing to better understand human affect, this tendency to categorize risks not only oversimplifies the nuanced nature of dimensional emotion constructs, but also fails to capture critical information useful for more accurate and representative modeling. By reducing dimensional data to categorical representations, we may inadvertently neglect important variations and subtleties inherent in emotional experiences, therefore limiting the generalization of these models.

Our findings underscore the need for future research in affective computing to consciously balance the practical necessities of machine learning approaches with the conceptual richness of dimensional emotion models. As the field continues to evolve, it is critical that we strive for methods that not only optimize performance but also aim to understand the complex nature of affective states accurately.

The Role of Detailed Methodologies and Open Databases in Achieving Replication

The fields of psychology and neuroscience, as documented in recent studies (Button et al., 2013; Open Science Collaboration, 2015), confront substantial challenges in achieving replication, a fundamental aspect of scientific rigor. While these obstacles are not unique to these fields, they have received considerable attention in the broader scientific literature (Egger et al., 2019). Our findings suggest that impediments to replicability are also present in the field of affective computing.

Firstly, a considerable number of the studies reviewed did not provide detailed information regarding their samples. For example, less than a quarter of the papers disclosed the country of origin of the sample, a critical omission given the well-established influence of culture on emotional experience (Campos et al., 1994; Lim, 2016; Matsumoto, 1989, 1991). In addition, important demographic data, such as sample size and age of participants, were often missing. Most notably, the absence of gender distribution is particularly troubling given existing research highlighting gender differences in emotional responses (Chaplin, 2015; Abbruzzese et al., 2019; García-Fernández et al., 2021; Fischer et al., 2004; Manstead, 1992).

Secondly, despite a modest uptick in the adoption of open databases, the majority of studies continue to utilize private databases. This practice severely diminishes the transparency and reproducibility foundational to scientific advancement (Nosek et al., 2015; Diener & Biswas-Diener, 2016). Our observations align with Roy et al. (2019), who pointed out the challenges in reproducing study findings due to the unavailability of data and code.

For those studies that do utilize open databases, there is a discernible concentration on a few select resources, with 87% of studies relying on a small set of databases. This is especially evident with frequently cited databases such as DEAP and AMIGOS, which suggests a constriction in the variety of data sources leveraged in research.

Thirdly, a significant lack of reporting concerns the use of physiological measures. Only 36% of studies reported the placement of EDA sensors, leaving the location of the hemibody (i.e., right or left) unknown in most cases. This omission complicates replication and comparative analysis across studies, particularly in light of studies reporting lateralization of EDA responses (Banks et al., 2012; Costanzo et al., 2015; Kasos et al., 2018; Picard et al., 2016).

Finally, there is a lack of standardization in the methods used to elicit emotions in most of the literature reviewed. While some variation in methods can be attributed to concerns about ecological validity (Paylor, 2009), the lack of detailed reporting on elicitation techniques remains an area for improvement. Furthermore, approximately half of the studies either did not mention or did not utilize self-report questionnaires, which are essential for precise emotion prediction (Zhang et al., 2016). Without uniform elicitation protocols and transparency regarding the use of self-report methods, the effectiveness of emotion induction in subjects cannot be reliably assessed.

Limited use of Theoretical Frameworks

A significant issue encountered during our review can be described as the adoption of a theoretically "neutral" stance or "agnostic approach" in the field of affective computing. This phenomenon refers to studies on emotions conducted without explicit attention or consideration to the theoretical foundations that underpin emotion elicitation and recognition.

In particular, we found that most articles originate from technical and engineering-oriented journals rather than those focused on psychology or neuroscience. This disciplinary divide is significant because it suggests that studies targeting a technical audience might not fully account for the psychological and physiological underpinnings of emotion recognition. The predominance of technical discourse raises crucial questions about the synergy between predictive models and traditional explanatory frameworks in psychology and neuroscience. While explanatory models aim to illuminate the causal dynamics within data, predictive models excel at discerning patterns within complex and high-dimensional datasets, thereby offering a complementary perspective that enhances generalizability and the discovery of subtle, underlying relationships (Yarkoni & Westfall, 2017; Bzdok, Engemann, Thirion, 2021).

To take full advantage of affective computing research, it will be crucial to adopt an integrated methodology that combines the predictive power of computational models with the rich theoretical insights of explanatory frameworks. However, for predictive models to be truly informative about emotional experience, they should be interpretable. Our review reveals a notable deficiency in this area: of 99 papers reviewed, only 23 attempted to interpret the results in terms of physiological or psychological implications, suggesting a significant gap in the literature. A serious effort to synthesize predictive power with the interpretability necessary for physiological and psychological analysis will enable us to better understand and replicate the complexity of human emotion through advanced artificial intelligence models.

Another important concern relates to the selection of emotional categories and dimensions analyzed. Many of the studies reviewed omitted the use of self-report measures and delegated the responsibility for categorizing affective states solely to the researcher, depending on the chosen emotion elicitation methods, which were often ad hoc. This raises critical questions about the validity of emotional labeling; for example, can we reliably distinguish whether an emotion labeled "amusement" is really amusement and not joy or happiness? Beyond potential distortions of emotion categories, certain categories were entirely missing in the reviewed studies, e.g. anger. It is imperative that future studies not only use self-report measures to validate emotional states, but also adopt a more rigorous and theoretically informed approach to the selection of emotional categories and dimensions. Without addressing these fundamental issues, we risk perpetuating uncertainties that undermine the development of the field, possibly due to a lack of closer cross-disciplinary collaboration between psychology and computer science disciplines (Behnke et al., 2023).

Arousal Models Outperform Valence in EDA-Based Emotion Recognition

Our meta-analysis of 76 arousal models and an equal number of valence models across 12 unique studies provides valuable insights into their comparative performance in emotion recognition: our data suggest that arousal models outperform valence models with EDA. The relatively higher accuracy of arousal models could be attributed to several factors. One possibility is that arousal, which is by definition associated with physiological activation, may manifest more noticeably in autonomic signals, thereby facilitating its detection by machine learning models using EDA (Kreibig, 2010). Another possible explanation relates to the type of stimuli used in the studies included in our analysis. Certain types of stimuli, such as images or sounds, may be more likely to elicit strong arousal responses than distinct valence responses, leading to greater accuracy in arousal models (Bradley and Lang, 2000). Future research would benefit from investigating the potential effects of different stimulus types, including their low and high-level properties, on the performance of arousal and valence models.

The observed negative correlation between mean accuracy and the difference in accuracy (arousal - valence) is striking, especially at lower levels of mean accuracy. This significant performance gap at lower levels could indicate that arousal models are performing significantly better than chance levels, while valence models are struggling to do so at a lower level of combined performance (close to 50% accuracy). This could indicate that one dimension is being modeled effectively while the other is not. The more pronounced difference may be due to the inherent characteristics of the arousal dimension, which may be more easily captured by physiological signals such as EDA. For example, the convex optimization approach to EDA processing (cvxEDA) proposed by Greco et al. (2016) has become a common preprocessing step in machine learning pipelines dealing with EDA, which allows estimation of autonomic nervous system activity from the EDA signal. Phasic component peaks obtained from cvxEDA have been found to correlate with different levels of arousal, providing a robust method for quantifying arousal (Greco et al., 2016). Such preprocessing methods enhance the detection capabilities of arousal models by refining the input data for subsequent machine learning algorithms. Alternatively, performance discrepancies between arousal and valence models could arise from other methodological choices, such as feature engineering and selection, that inadvertently favor the detection of one emotional dimension over the other, again highlighting the importance of theoretically informed methodological choices in affective computing research.

As models improve in overall accuracy, the performance gap appears to narrow. However, models with extremely high mean accuracy seem to be outliers and should be interpreted with caution. Given the exploratory nature of our study and its reliance on a limited dataset, these observations should be considered preliminary findings that highlight the need for more extensive research to understand the factors contributing to these trends.

The superior performance of arousal models, as suggested by our meta-analysis, points to areas for further research to enhance the accuracy of valence models. This could involve refining feature selection methods, testing more suitable machine learning algorithms, or incorporating multimodal data sources that may provide a richer context for predicting valence.

In conclusion, our meta-analysis offers valuable empirical evidence to the field of affective computing, illuminating the comparative efficacy of arousal and valence models in emotion recognition. As the field progresses, delving into the specific complexities and nuanced differences of various affective dimensions will be pivotal to further advance the predictive capabilities and ecological validity of emotion recognition models.

Generalizability of Emotion Recognition Research Requires Cultural and Geographical Representation

Generalizability is a cornerstone of scientific research, vital for ensuring reproducibility and broad applicability. In emotion recognition research, particularly using EDA, the geographical and cultural diversity of study samples is crucial (Boucsein, 2012). Our systematic review reveals a concentration of research within Asian countries, notably China, India, and Turkey, with significant contributions from the United States and Europe. However, regions like Latin America and Africa are markedly underrepresented.

This geographical and cultural bias extends to dataset compositions, often overlooking diverse racial and ethnic groups. Such limitations are not unique to our field but reflect a broader trend in human-centric research. The predominance of data from Asian contexts could impair the applicability of emotion recognition models across varied cultural and racial backgrounds. This concern is echoed in recent studies (Verhoef & Fosch-Villaronga, 2023), highlighting the need for more inclusive research practices.

Expanding datasets geographically is vital for enhancing research generalizability.

Recent trends show an increase in testing machine learning models' transferability across diverse datasets (He et al., 2022; Engemann et al., 2018; Rayatdoost & Soleymani, 2018), emphasizing the need for more inclusive research in underrepresented regions. Future research that embraces this diversity will support the development of conclusions that are more generalizable and methodologically robust.

Limitations

The scope of this work is primarily influenced by the nature of our sample. We acknowledge that certain conditions of our review may limit the breadth of our conclusions. It is important to clarify that our decision to exclude articles using a multimodal approach to emotion prediction (Al Osman & Falk, 2017; Poria et al., 2017; Song et. al., 2008) was driven by our focus on the EDA literature, given the close relationship between EDA and arousal as an affective dimension. While this focus allows for a deep dive into the EDA literature, it necessarily neglects other potentially informative signals, such as ECG or pupillometry, that may provide complementary or even superior insights into arousal. Thus, future iterations of this work might consider expanding the review to include these other unimodal signals or even multimodal approaches. Such an expansion could provide a more nuanced understanding of the relationship between different branches of peripheral nervous system activity and affective components such as arousal. For readers interested in the broader domain of physiological signals, especially in the context of wearable technology, the work of Saganowski et al. (2022) offers comprehensive insights that could greatly complement the findings of this review.

Second, articles studying clinical populations were intentionally excluded, as our goal was to explore the basic study of affective states at the event level—specifically, the prediction of different affective states within the same subject. In contrast, clinical studies typically aim to detect mood alterations, such as anxiety or depression, relative to healthy control subjects, thus, shifting the focus of analysis to the subject level (between subjects) rather than the event level (within subjects). While subject-level studies of arousal and emotion are critical for developing biomarkers in support of novel therapeutics , they did not fit in the scope of this work. However, given the evidence on the influence of mental states on electrodermal activity (Sarchiapone et al., 2018; Öhman, 1981; Vahey & Becerra, 2015), we would expect that modeling stimulus-induced modulation of arousal and emotion may hold a key to clinical applications by, potentially, uncovering differences in processing affective stimuli that are characteristic for certain groups of patients.

Finally, our study is limited to papers published by the end of 2020. The time limit was set according to the start of our review process. Given the rapid progress and exponential increase in the number of papers on affective computing over time (Guo et al., 2020), the fact that this review has only included papers up to 2020 leaves out numerous papers between now and the publication of this paper. Future reviews can use this paper as a guide for future work that incorporates the latest advances in the field.

Recommendations

Through our exploration of affective computing models, we have identified several critical practices that can improve the transparency, reproducibility, and overall quality of research in this field. These recommendations cover various aspects of the research process, including data transparency and accessibility, sample characteristics, methodological clarity, cultural and theoretical considerations, performance metrics, data analysis sharing, and study pre-registration. We have summarized these key recommendations in Table 2: "Recommendations for Effective and Transparent Affective Computing Research:.

Following these recommendations can help researchers ensure that their work is conducted with the utmost rigor, thereby enhancing its credibility and facilitating meaningful advances in the field. For additional guidance, particularly on physiological measures, researchers should consult Behnke et al. (2022), who provide a checklist for responsible wearable use in affective research. Taken together, these guidelines can serve as a valuable resource for those already engaged in affective computing research and for those considering entering this burgeoning field.

Conclusions

This study set out to explore affective and machine learning models and EDA measures used in the affective computing literature. Our findings highlighted key areas for improvement in scientific practice, particularly in relation to data transparency and reporting. A recurring observation was the lack of a comprehensive psychophysiological interpretation of the findings.

Thus, we emphasize the importance of a psychological perspective in the study of emotions, in addition to the computational and statistical methods employed in affective computing. While the latter offers exciting new avenues for emotion research, a concerted effort to synergize affective science with affective computing is essential. Affective computing has already benefited from the affect-elicitation techniques employed in affective science, and conversely, affective science has adopted the measurement approaches of affective computing (D'Mello et al., 2018). However, while the field of affective computing is emerging as a distinctly interdisciplinary field, recent evidence suggests that collaboration and cross-fertilization between scientists from different disciplines (e.g., psychologists and computer scientists) remains the exception rather than the rule in affective computing and affective science (Behnke et al., 2023)

By fostering interdisciplinarity, we can significantly enhance the quality and robustness of our models, leading to a more holistic and nuanced understanding of affective phenomena. Furthermore, by generating models that are more physiologically valid, we can achieve superior developments in the area of affective computing (Kappas & Gratch, 2023). Artificial intelligence models that better represent the generative mechanisms of affective states will thus generalize more effectively when implemented in practical applications. This integrated approach also lays the groundwork for future research, as it will be pivotal in addressing identified gaps and advancing the field

Declaration of Conflicting Interests

D.A.E is a full-time employee of F. Hoffmann-La Roche Ltd.

Funding

E.T. is supported by PICT-2019-02294 (Agencia I+D+i,Argentina) and ANID/FONDECYT Regular 1220995 (Chile)

Supplementary Material

Supplementary material for this article is available online in a GitHub repository (https://github.com/EmmAMaldonado/review-emotion-recognition-eda).

Author Contributions

Conceived and designed the systematic review and meta-analysis: T.A.D.

Performed the systematic search and data extraction: L.A.G, E.A.M., A.A.D.

Reviewed the systematic search and data extraction: T.A.D., L.A.G, E.A.M., A.A.D.

Analyzed the data: T.A.D., L.A.G, E.A.M., A.A.D.

Contributed to the development of the meta-analytic approach and provided analysis tools: T.A.D., D.A.E.

Wrote the initial draft of the manuscript: T.A.D., L.A.G, E.A.M., A.A.D.

Contributed to the writing of the review and editing of the manuscript for critical intellectual content: T.A.D., D.A.E., E.T

Critically revised the manuscript: D.A.E., E.T

References

Abbruzzese, L., Magnani, N., Robertson, I. H., & Mancuso, M. (2019). Age and gender differences in emotion recognition. *Frontiers in psychology*, *10*, 2371.doi: 10.3389/fpsyg.2019.02371

Alharbi, M., Karimi, H.A. (2020). PROBE: preparing for roads in advance of barriers and errors. In: Arai, K., Bhatia, R., Kapoor, S. (eds) Proceedings of the Future Technologies Conference (FTC) 2019. FTC 2019. *Advances in Intelligent Systems and Computing, vol 1069. Springer, Cham.* https://doi.org/10.1007/978-3-030-32520-6_67

Allen, A. P., Kennedy, P. J., Dockray, S., Cryan, J. F., Dinan, T. G., & Clarke, G. (2017). The trier social stress test: principles and practice. *Neurobiology of stress*, *6*, 113-126.https://doi.org/10.1016/j.ynstr.2016.11.001

Al Osman, H., & Falk, T. H. (2017). Multimodal affect recognition: current approaches and challenges. *Emotion and Attention Recognition Based on Biological Signals and Images*, 59-86.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Atmaja, B. T., Sasou, A., & Akagi, M. (2022). Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Communication*.https://doi.org/10.1016/j.specom.2022.03.002

Aylett, R., & Paiva, A. (2012). Computational modelling of culture and affect. *Emotion Review, 4*(3), 253-263.https://doi.org/10.1177/1754073912439766

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience, 12*(1), 1-23.https://doi.org/10.1093/scan/nsx060

Babaei, E., Tag, B., Dingler, T., & Velloso, E. (2021, May). A critique of electrodermal activity practices at chi. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).https://doi.org/10.1145/3411764.3445370

Banks, S. J., Bellerose, J., Douglas, D., & Jones-Gotman, M. (2012). Bilateral skin conductance responses to emotional faces. *Applied psychophysiology and biofeedback*, *37*, 145-152.https://doi.org/10.1007/s10484-011-9177-7

Behnke, M., Saganowski, S., Kunc, D., & Kazienko, P. (2022). Ethical considerations and checklist for affective research with wearables. *IEEE Transactions on Affective Computing.*

Behnke, M., Saganowski, S., Kaczmarek, Ł. D., & Kazienko, P. (2023, March). Emotions Studied by Computer Scientists and Psychologists—A Complementary Perspective. In *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events* (PerCom Workshops) (pp. 206-211). IEEE.

Bzdok, D., Engemann, D., & Thirion, B. (2020). Inference and prediction diverge in biomedicine. *Patterns*, *1*(8).

Bishop, C. M. (2006). Machine learning. *Machine learning*, *128*(9).

Boucsein, W. (2012). *Electrodermal activity*. Springer Science & Business Media.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, *25*(1), 49-59.https://doi.org/10.1016/0005-7916(94)90063-9

Bradley, M. M., & Lang, P. J. (2007). The international affective digitized sounds. *IADS-2: Stimuli, instruction manual and affective ratings. 2nd ed. Gainesville, FL: The Center for Research in Psychophysiology, University of Florida*.

Brigham, T. J. (2017). Merging technology and emotions: introduction to affective computing. *Medical reference services quarterly*, *36*(4), 399-407.https://doi.org/10.1080/02763869.2017.1369289

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, *14*(5), 365-376.

Calvo, R. A. (2010). Latent and emergent models in affective computing. *Emotion Review*, *2*(3), 288-289.https://doi.org/10.1177/1754073910368735

Calvo, R. A., & D'Mello, S. (2010). Affect detection: an interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, *1*(1), 18-37.https://doi.org/10.1109/T-AFFC.2010.1

Calvo, R. A., D'Mello, S., Gratch, J. M., & Kappas, A. (Eds.). (2015). *The Oxford handbook of affective computing*. Oxford Library of Psychology.

Campos, J. J., Mumme, D., Kermoian, R., & Campos, R. G. (1994). A functionalist perspective on the nature of emotion. *Japanese Journal of Research on Emotions*, *2*(1), 1-20.https://doi.org/10.4092/jsre.2.1

Chan, D. (2010). So why ask me? Are self-report data really that bad?. In *Statistical and methodological myths and urban legends* (pp. 329-356). Routledge.

Chaplin, T. M. (2015). Gender and emotion expression: a developmental contextual perspective. *Emotion Review*, *7*(1), 14-21.https://doi.org/10.1177/1754073914544408

Clay-Warner, J., & Robinson, D. T. (2015). Infrared thermography as a measure of emotion response. *Emotion Review*, *7*(2), 157-162.https://doi.org/10.1177/1754073914554783

Cockburn, A., Gutwin, C., & Dix, A. (2018, April). Hark no more: on the preregistration of chi experiments. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1-12).https://doi.org/10.1145/3173574.3173715

Cockburn, A., Dragicevic, P., Besançon, L., & Gutwin, C. (2020). Threats of a replication crisis in empirical computer science. *Communications of the ACM*, *63*(8), 70-79.https://doi.org/10.1145/3360311

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of health and social behavior*, 385-396.https://doi.org/10.2307/2136404

Costanzo, E. Y., Villarreal, M., Drucaroff, L. J., Ortiz-Villafañe, M., Castro, M. N., Goldschmidt, M., ... & Guinjoan, S. M. (2015). Hemispheric specialization in affective responses, cerebral dominance for language, and handedness: Lateralization of emotion, language, and dexterity. *Behavioural brain research*, *288*, 11-19.https://doi.org/10.1016/j.bbr.2015.04.006

Davidson, R. J., Sherer, K. R., & Goldsmith, H. H. (Eds.). (2009). *Handbook of affective sciences*. Oxford University Press.

Diener, E., Biswass-Diener, R. (2020). The Replication Crisis in Psychology. Available online: http://noba.to/q4cvydeh .

D'Mello, S., Kappas, A., & Gratch, J. (2018). The affective computing approach to affect measurement. *Emotion Review*, *10*(2), 174-183.https://doi.org/10.1177/1754073917696583

Egger, M., Ley, M., & Hanke, S. (2019). Emotion recognition from physiological signal analysis: a review. *Electronic Notes in Theoretical Computer Science*, *343*, 35-55.https://doi.org/10.1016/j.entcs.2019.04.009

Ekman, P. (1992). Facial expressions of emotion: an old controversy and new findings. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *335*(1273), 63-69.https://doi.org/10.1098/rstb.1992.0008

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *Facial action coding system: facial action coding system: the manual: on CD-ROM*. Research Nexus.

Engemann, D. A., Raimondo, F., King, J. R., Rohaut, B., Louppe, G., Faugeras, F., ... & Sitt, J. D. (2018). Robust EEG-based cross-site and cross-protocol classification of states of consciousness. *Brain, 141*(11), 3179-3192.

García‐Fernández, L., Romero‐Ferreiro, V., Padilla, S., David López‐Roldán, P., Monzó‐García, M., & Rodriguez‐Jimenez, R. (2021). Gender differences in emotional response to the COVID‐19 outbreak in Spain. *Brain and behavior*, *11*(1), e01934.https://doi.org/10.1002/brb3.1934

Greco, A., Valenza, G., Lanata, A., Scilingo, E. P., & Citi, L. (2015). cvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE transactions on biomedical engineering, 63*(4), 797-804.

Gross, J. J., & Levenson, R. W. (1995). Emotion elicitation using films. *Cognition & emotion*, *9*(1), 87-108.https://doi.org/10.1080/02699939508408966

Guo, F., Li, F., Lv, W., Liu, L., & Duffy, V. G. (2020). Bibliometric analysis of affective computing researches during 1999~ 2018. *International Journal of Human–Computer Interaction*, *36*(9), 801-814.https://doi.org/10.1080/10447318.2019.1688985

Guo, L., Wang, L., Dang, J., Chng, E. S., & Nakagawa, S. (2022). Learning affective representations based on magnitude and dynamic relative phase information for speech emotion recognition. *Speech Communication*, *136*, 118-127.https://doi.org/10.1016/j.specom.2021.11.005

He, Z., Zhong, Y., & Pan, J. (2022). An adversarial discriminative temporal convolutional network for EEG-based cross-domain emotion recognition. *Computers in biology and medicine, 141*, 105048.

Handayani, D., Wahab, A., & Yaacob, H. (2015). Recognition of emotions in video clips: the self-assessment manikin validation. *TELKOMNIKA (Telecommunication Computing*

Electronics and Control), *13*(4), 1343-1351.http://doi.org/10.12928/telkomnika.v13i4.2735

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29-29.https://doi.org/10.1038/466029a

Henrich, J. (2020). *The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous*. Penguin UK.

Huang, W., & Chiang, S. (2014). The international affective picture system (IAPS). *Behavior Research Methods*, 46(2), 518-526. https://doi.org/10.3758/s13428-013-0422-5

Huang, Y., Wen, H., Qing, L., Jin, R., & Xiao, L. (2021). Emotion recognition based on body and context fusion in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3609-3617).

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering*, *9*(03), 90-95.

Idri, N. (2015). Zotero software: a means of bibliographic research and data organisation; teaching bibliographic research. *Arab World English Journal (AWEJ) Special Issue on CALL*, (2).http://dx.doi.org/10.2139/ssrn.2843984

Jang, E. H., Park, B. J., Kim, S. H., Chung, M. A., Park, M. S., & Sohn, J. H. (2014, April). Emotion classification based on bio-signals emotion recognition using machine learning algorithms. In *2014 International Conference on Information Science, Electronics and Electrical Engineering* (Vol. 3, pp. 1373-1376). IEEE.https://doi.org/10.1109/InfoSEEE.2014.6946144

Jones, E., Oliphant, T., & Peterson, P. (2001). SciPy: Open source tools for Python.

Kappas, A., & Gratch, J. (2023). These Aren’t The Droids You Are Looking for: Promises and Challenges for the Intersection of Affective Science and Robotics/AI. *Affective Science*, 1-6.

Kasos, K., Zimonyi, S., Kasos, E., Lifshitz, A., Varga, K., & Szekely, A. (2018). Does the electrodermal system "take sides" when it comes to emotions?. *Applied psychophysiology and biofeedback*, *43*, 203-210.https://doi.org/10.1007/s10484-018-9398-0

Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., ... & Patras, I. (2011). Deap: a database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, *3*(1), 18-31.https://doi.org/10.1109/T-AFFC.2011.15

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: a review. *Biological psychology, 84*(3), 394-421.https://doi.org/10.1016/j.biopsycho.2010.03.010

Lei, S., & Gratch, J. (2023). Emotional Expressivity is a Reliable Signal of Surprise. *IEEE Transactions on Affective Computing.*

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., ... & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Annals of internal medicine, 151*(4), W-65.https://doi.org/10.7326/0003-4819-151-4-200908180-00136

Lim, N. (2016). Cultural differences in emotion: differences in emotional arousal level between the east and the west. *Integrative medicine research*, *5*(2), 105-109.https://doi.org/10.1016/j.imr.2016.03.004

Levenson, R. W. (2011). Basic emotion questions. *Emotion review*, *3*(4), 379-386.https://doi.org/10.1177/1754073911410743

Lobachev, S. (2009). Top languages in global information production. *Digital Voices: An Open Access Practice Journal*, *1*(1).https://doi.org/10.21083/partnership.v3i2.826

Manstead, A. S. R. (1992). Gender differences in emotion. In A. Gale & M. W. Eysenck (Eds.), *Handbook of individual differences: Biological perspectives* (pp. 355–387). John Wiley & Sons.

Matsumoto, D. (1989). Cultural influences on the perception of emotion. *Journal of Cross-Cultural Psychology*, *20*(1), 92-105.https://doi.org/10.1177/0022022189201006

Matsumoto, D. (1991). Cultural influences on facial expressions of emotion. *Southern Journal of Communication*, *56*(2), 128-137.https://doi.org/10.1080/10417949109372824

McKinney, W. (2010, June). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, No. 1, pp. 51-56).

Mede, N. G., Schäfer, M. S., Ziegler, R., & Weißkopf, M. (2021). The "replication crisis" in the public eye: germans' awareness and perceptions of the (ir) reproducibility of scientific research. *Public Understanding of Science*, *30*(1), 91-102.https://doi.org/10.1177/0963662520954370

Miranda-Correa, J. A., Abadi, M. K., Sebe, N., & Patras, I. (2021). Amigos: a dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, *12*(2), 479-493.https://doi.org/10.1109/TAFFC.2018.2884461

Morreale, F., Masu, R., & De Angeli, A. (2013). Robin: an algorithmic composer for interactive scenarios. *Sound and Music Computing*. Available: https://www.researchgate.net/profile/Fabio-Morreale/publication/253650269_ROBIN_AN_ALGORITHMIC_COMPOSER_FOR_INTERACTIVE_SCENARIOS/links/00b4951fa33aadeb40000000/ROBIN-AN-ALGORITHMIC-COMPOSER-FOR-INTERACTIVE-SCENARIOS.pdf

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., ... & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature human behaviour*, *1*(1), 1-9.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., & Breckler, S. J. & Contestabile, M.(2015). Promoting an open research culture. *Science*, 348(6242), 1422-1425. https://doi.org/10.1126/science.aab2374

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Ousley O. Y., Arriaga R., Abowd G. D., Morrier M. (2012). Rapid assessment of social-communicative abilities in infants at risk of autism. *Technical Report CBI-100, Center for Behavior Imaging*.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic reviews*, *5*, 1-10.https://doi.org/10.1186/s13643-016-0384-4

Öhman, A. (1981). Electrodermal activity and vulnerability to schizophrenia: a review. *Biological Psychology*, *12*(2-3), 87-145.https://doi.org/10.1016/0301-0511(81)90008-9

Pace-Schott, E. F., Amole, M. C., Aue, T., Balconi, M., Bylsma, L. M., Critchley, H., ... & VanElzakker, M. B. (2019). Physiological feelings. *Neuroscience & Biobehavioral Reviews, 103*, 267-304.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Alonso-Fernández, S. (2021). Declaración PRISMA 2020: una guía actualizada para la publicación de revisiones sistemáticas. Revista española de cardiología, 74(9), 790-799.https://doi.org/10.1016/j.recesp.2021.06.016

Panksepp, J., & Watt, D. (2011). What is basic about basic emotions? lasting lessons from affective neuroscience. *Emotion review*, *3*(4), 387-396.https://doi.org/10.1177/1754073911410741

Paylor, R. (2009). *Questioning standardization in science. Nature Methods, 6(4), 253–254.*

Picard, R. W. (1999, August). Affective computing for HCI. In *HCI (1)* (pp. 829-833).

Picard, R. W. (2000). *Affective Computing*. MIT press.

Picard, R. W. (2010). Emotion research by the people, for the people. *Emotion Review*, *2*(3), 250-254.https://doi.org/10.1177/1754073910364256

Picard, R. W., Fedor, S., & Ayzenberg, Y. (2016). Multiple arousal theory and daily-life electrodermal activity asymmetry. *Emotion review*, *8*(1), 62-75.https://doi.org/10.1177/1754073914565517

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: from unimodal analysis to multimodal fusion. *Information Fusion*, 37 , 98-125.https://doi.org/10.1016/j.inffus.2017.02.003

Posada-Quintero, H. F., & Chon, K. H. (2020). Innovations in electrodermal activity data collection and signal processing: a systematic review. *Sensors*, *20*(2), 479.https://doi.org/10.3390/s20020479

Rayatdoost, S., & Soleymani, M. (2018, September). Cross-corpus EEG-based emotion recognition. In *2018 IEEE 28th international workshop on machine learning for signal processing (MLSP)* (pp. 1-6). IEEE.

Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013, April). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (pp. 1-8). IEEE.https://doi.org/10.1109/FG.2013.6553805

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., & Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, *16*(5), 051001.10.1088/1741-2552/ab260c

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.https://psycnet.apa.org/doi/10.1037/h0077714

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect grid: a single-item scale of pleasure and arousal. *Journal of personality and social psychology*, *57*(3), 493.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, *110*(1), 145.https://psycnet.apa.org/doi/10.1037/0033-295X.110.1.145

Saganowski, S., Perz, B., Polak, A., & Kazienko, P. (2022). Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review. *IEEE Transactions on Affective Computing.*

Said, Y., & Barr, M. (2021). Human emotion recognition based on facial expressions via deep learning on high-resolution images. *Multimedia Tools and Applications*, *80*(16), 25241-25253.https://doi.org/10.1007/s11042-021-10918-9

Sánchez-Reolid, R., López de la Rosa, F., Sánchez-Reolid, D., López, M. T., & Fernández-Caballero, A. (2022). Machine learning techniques for arousal classification from electrodermal activity: a systematic review. *Sensors*, *22*(22), 8886.https://doi.org/10.3390/s22228886

Sarchiapone, M., Gramaglia, C., Iosue, M., Carli, V., Mandelli, L., Serretti, A., ... & Zeppegno, P. (2018). The association between electrodermal activity (EDA), depression and suicidal behaviour: a systematic review and narrative synthesis. *BMC psychiatry*, *18*(1), 1-27.https://doi.org/10.1186/s12888-017-1551-4

Sato, W., Kochiyama, T., & Yoshikawa, S. (2020). Physiological correlates of subjective emotional valence and arousal dynamics while viewing films. *Biological Psychology, 157*, 107974.https://doi.org/10.1016/j.biopsycho.2020.107974

Seabold, S., & Perktold, J. (2010, June). Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference* (Vol. 57, No. 61, pp. 10-25080).

Sharma, K., Castellini, C., van den Broek, E. L., Albu-Schaeffer, A., & Schwenker, F. (2019). A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific data, 6*(1), 196. https://doi.org/10.1038/s41597-019-0209-0

Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., ... & Yang, X. (2018). A review of emotion recognition using physiological signals. *Sensors, 18*(7), 2074.https://doi.org/10.3390/s18072074

Shukla, J., Barreda-Angeles, M., Oliver, J., Nandi, G. C., & Puig, D. (2019). Feature extraction and selection for emotion recognition from electrodermal activity. *IEEE Transactions on Affective Computing*, *12*(4), 857-869.https://doi.org/10.1109/TAFFC.2019.2901673

Siedlecka, E., & Denson, T. F. (2019). Experimental methods for inducing basic emotions: a qualitative review. *Emotion Review*, *11*(1), 87-97.https://doi.org/10.1177/1754073917749016

Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2011). A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, *3*(1), 42-55.https://doi.org/10.1109/T-AFFC.2011.25

Soleymani, M., Pantic, M., & Pun, T. (2011). Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing*, *3*(2), 211-223.https://doi.org/10.1109/T-AFFC.2011.37

Song, M., You, M., Li, N., & Chen, C. (2008). A robust multimodal approach for emotion recognition. *Neurocomputing*, *71*(10-12), 1913-1920. https://doi.org/10.1016/j.neucom.2007.07.041

Stroop, J. R. (1992). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: General*, *121*(1), 15.https://psycnet.apa.org/doi/10.1037/h0054651

Sturm, C., Oh, A., Linxen, S., Abdelnour Nocera, J., Dray, S., & Reinecke, K. (2015, April). How WEIRD is HCI? Extending HCI principles to other countries and cultures. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2425-2428). https://doi.org/10.1145/2702613.2702656

Subramanian, R., Wache, J., Abadi, M. K., Vieriu, R. L., Winkler, S., & Sebe, N. (2016). ASCERTAIN: emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, *9*(2), 147-160.https://doi.org/10.1109/TAFFC.2016.2625250

Tracy, J. L., & Randles, D. (2011). Four models of basic emotions: a review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion review*, *3*(4), 397-405.https://doi.org/10.1177/1754073911410747

Vahey, R., & Becerra, R. (2015). Galvanic skin response in mood disorders: a critical review. *International Journal of Psychology & Psychological Therapy*, 15(2), 275–304.

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. CreateSpace.

van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering, 13*(2), 22.

Verhoef, T., & Fosch-Villaronga, E. (2023). Towards affective computing that works for everyone. *arXiv preprint arXiv:2309.10780.*

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & Van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, *17*(3), 261-272.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology*, *54*(6), 1063.https://psycnet.apa.org/doi/10.1037/0022-3514.54.6.1063

Yadegaridehkordi, E., Noor, N. F. B. M., Ayub, M. N. B., Affal, H. B., & Hussin, N. B. (2019). Affective computing in education: a systematic review and future research. *Computers & Education*, *142*, 103649.https://doi.org/10.1016/j.compedu.2019.103649

Yannakakis, G. N. (2018). Enhancing health care via affective computing. *Malta Journal of Health Sciences*, 5(1), 38-42. Available: https://www.um.edu.mt/library/oar//handle/123456789/31730

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122.https://doi.org/10.1177/1745691617693393

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2007, November). A survey of affect recognition methods: audio, visual and spontaneous expressions. In *Proceedings of the 9th international conference on Multimodal interfaces* (pp. 126-133).https://doi.org/10.1145/1322192.1322216

Zhang, K., Zhang, H., Li, S., Yang, C., & Sun, L. (2018, June). The PMEmo dataset for music emotion recognition. In *Proceedings of the 2018 acm on international conference on multimedia retrieval* (pp. 135-142).https://doi.org/10.1145/3206025.3206037

Zhang, B., Essl, G., & Mower Provost, E. (2016, October). Automatic recognition of self-reported and perceived emotion: does joint modeling help?. In *Proceedings of the*

18th ACM International Conference on Multimodal Interaction (pp. 217-224).https://doi.org/10.1145/2993148.2993173

Table 1. Overview of datasets used for emotion recognition studies using EDA

Dataset	Citation	N	Purpose	EDA device	Other signals	Affective annotation	Stimuli	Link
DEAP	Koelstra et al. (2012)	32	To facilitate the advancement of emotion analysis methods using a multimodal dataset incorporating physiological signals.	Biosemi Active Two	EEG, EOG, EMG, BPG, Resp, Temp, Video	SAM with dimensions: valence, arousal, dominance. Other self-assessments: like/dislike, and familiarity.	Fourteen music videos (one-minute length).	https://www.eecs.qmul.ac.uk/mmv/datasets/deap/
AMIGOS	Miranda-Correa et al. (2018)	40	To enable the understanding of emotions, opinions, and social dynamics in real-world scenarios through the analysis of multimodal social interaction data.	Shimmer 2R	EEG, ECG, Video, Audio	PANAS and self-assessments with dimensions: valence, arousal, dominance, like/dislike, familiarity. Basic emotions: anger, disgust, fear, sadness, surprise, happiness, and neutral.	Sixteen short videos and four long videos	http://www.eecs.qmul.ac.uk/mmv/datasets/amigos/index.html
MAHNOB	Soleymani et al. (2012)	27	Promote the exploration and understanding of non-verbal dimensions in human communication, emotion recognition and implicit tagging	Biosemi Active Two	EEG,ECG, Video, Audio, Eye-gaze, Resp, Temp	Self-assessment with dimensions: valence, arousal, dominance and predictability	Twenty videos between 34.9 and 117 s long	https://mahnob-db.eu/
ASCERTAIN	Subramanian et al. (2016)	58	To analyze stress and anxiety symptoms.	Commercial Bluetooth sensor	EEG, ECG, Video	Dimensions: valence, arousal, engagement, liking and familiarity	Thirty-six movie clips. The average length is 80 seconds	N/A
MMDB	Rehg et al. (2013)	121	To support the study and modeling of social behavior, emotion, and nonverbal communication in dyadic interactions	Affectiva Q sensor	Video, Audio	Valence-arousal emotional rating scales	Five-minute social interaction scenarios.	https://cbs.ic.gatech.edu/mmdb/

RECOLA	Ringeval et al. (2013)	46	To enable real-time analysis of emotional responses in social interaction scenarios	Biopac mp36	ECG, Video, Audio	SAM with dimensions: valence and arousa. And PANAS	Twenty-three teams were given a collaborative task that elicited social and emotional behaviors (~ 15 minutes length)	https://qualinet.github.io/databases/audiovisual/recola/
PMEmo	Zhang et al. (2018)	457	To capture and analyze individual emotional expressions using a multimodal data set collected from multiple sources.	Biopac mp150	-	SAM with dimensions: valence and arousal.	794 music clips.	https://github.com/HuiZhangDB/PMEmo
Hazumi1911	Katada et al. (2020)	26	To improve the understanding of human-computer interaction by studying communication methods and interpreting multimodal signals and expressions	Empatica e4	HR, Video, Audio	Self-annotation: enjoy-not enjoy: External annotation: enjoy-bored	Human-agent interaction dialogues.	https://github.com/ouktlab/Hazumi1911/
BioVid Emo DB	Zhang et al. (2016)	86	To facilitate research and development of emotion analysis techniques applied to video data	Nexus-32	EMG, ECG, Video	Self-assessment: valence, arousal, amusement, sadness, anger, disgust and fear.	Films clips (length: between 32 and 245 seconds)	https://emotionslab.org/about/
CONTINUOUS LIRIS ACCEDE	Baveye et al. (2015)	10	To facilite researchers and practitioners with a comprehensive and diverse collection of video data for various computer vision and multimedia tasks.	Bodymedia armband	-	Continuous arousal annotations	30 movies (6 different genres and 5 languages). Total duration: ~7 hours	https://liris-accede.ec-lyon.fr/database.php
WESAD	Schmidt et al. (2018)	15	To provide a high quality multimodal database that allows the detection and analysis of stress and affective states	Empatica e4 and RespiBAN Professional	BPG, ECG, EMG, Resp, Temp,	PANAS, SAM, SSSQ and STAI	Three conditions: amusement with funny video clips (total duration	https://github.com/WJMatthew/WESAD

through wearable devices.	ACC	392 seconds); stress induced by Trier Social Stress Test with a self-presentation task; and a seven-minute guided meditation session

Note. ACC = axis accelerometer, BPG = Blood Pressure, BVP= blood volume pulse, ECG= Electrocardiography, EDA = Electrodermal Activity, EEG= Electroencephalography, EMG = Electromyography, EOG= Electrooculography, HR = Heart Rate, N = Number of Participants, N/A = Not Available, Resp= Respiration, Temp = Temperature, SAM = Self-Assessment Manikin, SSSQ= Short Stress State Questionnaire,STAI = State-Trait Anxiety Inventory, PANAS= Positive Affective and Negative Affective Scale

Table 2. Recommendations for effective and transparent Affective Computing research

	Recommendation	Description
1	Data Transparency and Accessibility	Provide open access to the study data whenever possible to facilitate replication. Ensure a clear description of the data used, specifying if the entire dataset or a subset was used and which signals were incorporated. For newly collected data, strive to make this public, and detail all relevant sample information, such as gender and nationality.
2	Sample Properties	Providing a thorough account of the sample properties, such as age range, gender distribution, and geographical origin, is essential. Recognize and explicitly state the source of your sample to circumvent WEIRD (Western, Educated, Industrialized, Rich, Democratic) bias and improve the generalizability of your results. Highlight the influence of demographic details on the study outcomes and advocate for including a diverse demographic for improved validity.
3	Methodology Clarity	Clear and precise methodology is key. Offer detailed descriptions of the emotion elicitation techniques and measurement methods utilized in your study for accurate replication. Discuss any limitations or potential issues related to the methods used.
4	Cultural and Theoretical Considerations	Account for the significant influence cultural differences and social constructs have on emotion studies. Explain how these variations have been managed in your study. Provide clarity on the theoretical framework guiding your study, specifying which emotions were examined and the rationale behind this choice.
5	Baseline and Performance Metrics	To provide context and a reference point for your study's results, it is beneficial to include a state-of-the-art baseline or compare findings with dummy models, which are basic prediction models that do not learn from data but use simple strategies, such as predicting the most frequent class or the mean of the target variable. Avoid reporting a singular performance metric; instead, share multiple metrics such as accuracy, precision, recall, F1 score, or AUC-ROC, as each offers unique insights into your model's performance. If ranking is involved, including a confusion matrix is beneficial for its detailed breakdown of correct and incorrect model predictions.
6	Data Analysis Sharing for Reproducibility	For full reproducibility, make the data analysis publicly accessible. This includes all scripts, code, and detailed descriptions of the computational procedures used to analyze the data. Use version control tools like Git to track changes and developments in scripts and codes. Publish these in a public repository with detailed documentation to enable others to replicate the data processing and analysis if desired.
7	Interpretability and Physiological Implications	The results of affective computing studies can be effectively related to the broader body of affective science literature. Analyzing these results from psychological and/or physiological perspectives can create meaningful connections, enriching our understanding of affective phenomena.
8	Study Preregistration	If your study is hypothesis-driven, it's recommended to preregister your research. Preregistration offers assurance to the scientific community of the absence of Hypothesizing After the Results are Known (HARKing), thereby enhancing the validity of your findings and the integrity of the research process. Consider using preregistration platforms that offer varying levels of access control to accommodate any concerns about early disclosure of your work.

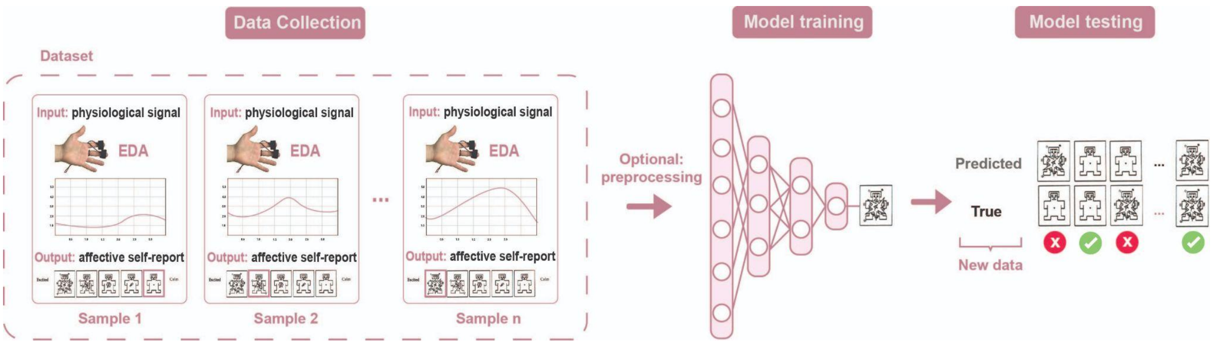

Figure 1. Emotion prediction pipeline using EDA. *Note*. EDA = electrodermal activity

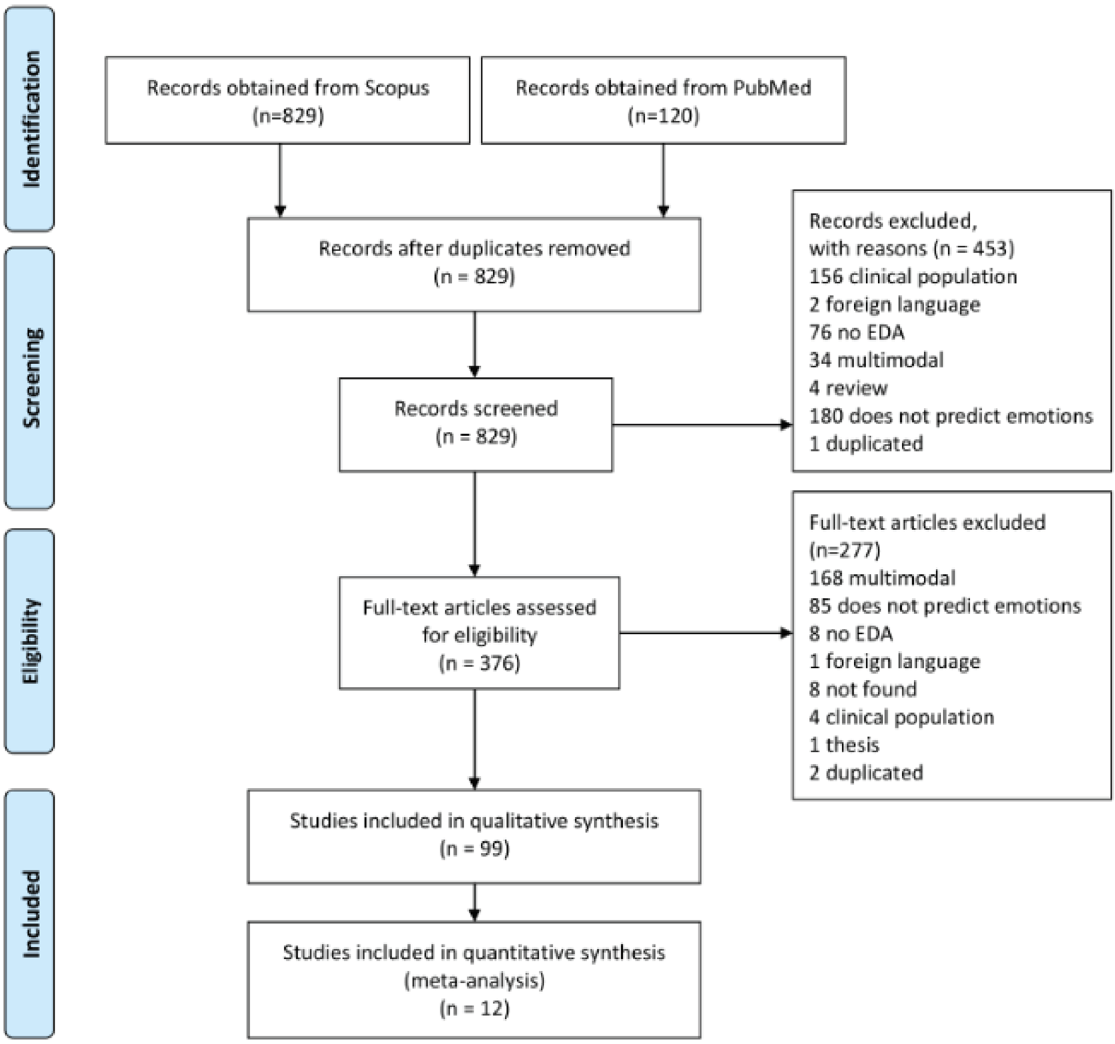

Figure 2. Systematic review flowchart

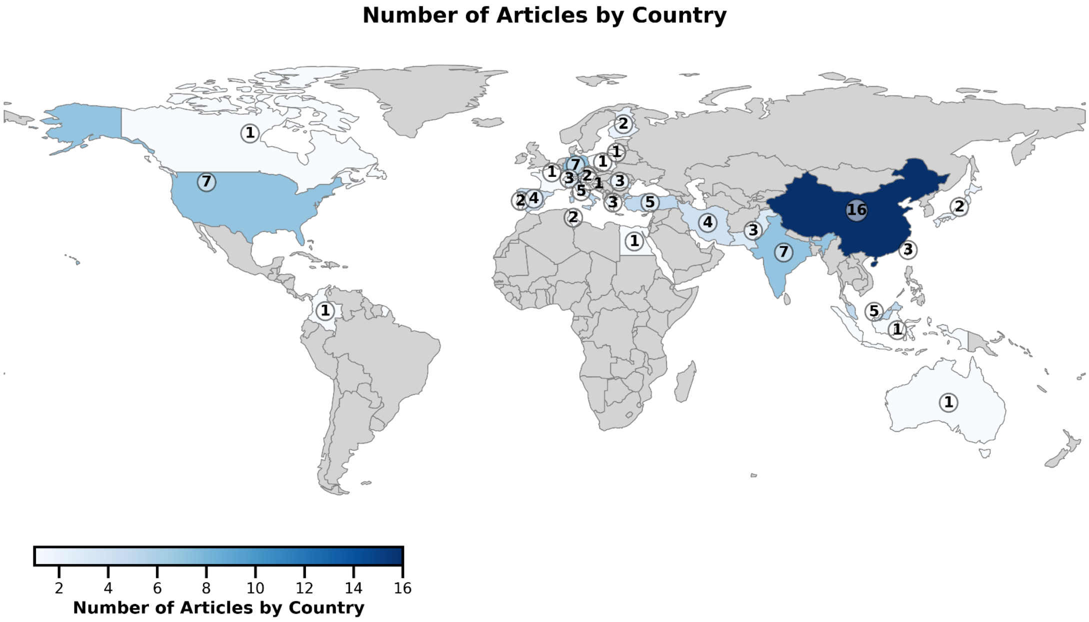

Figure 3. Distribution of papers in "Emotion Recognition and EDA" based on the country of residence of the first author. *Note.* The color spectrum in the lower left corner indicates the number of papers contributed by each country, ranging from light blue (one paper) to dark blue (most papers, with China leading with 16 papers). The numbers inside the circles represent the actual count of papers contributed by each country. Countries with no contributions in this area are shown in gray.

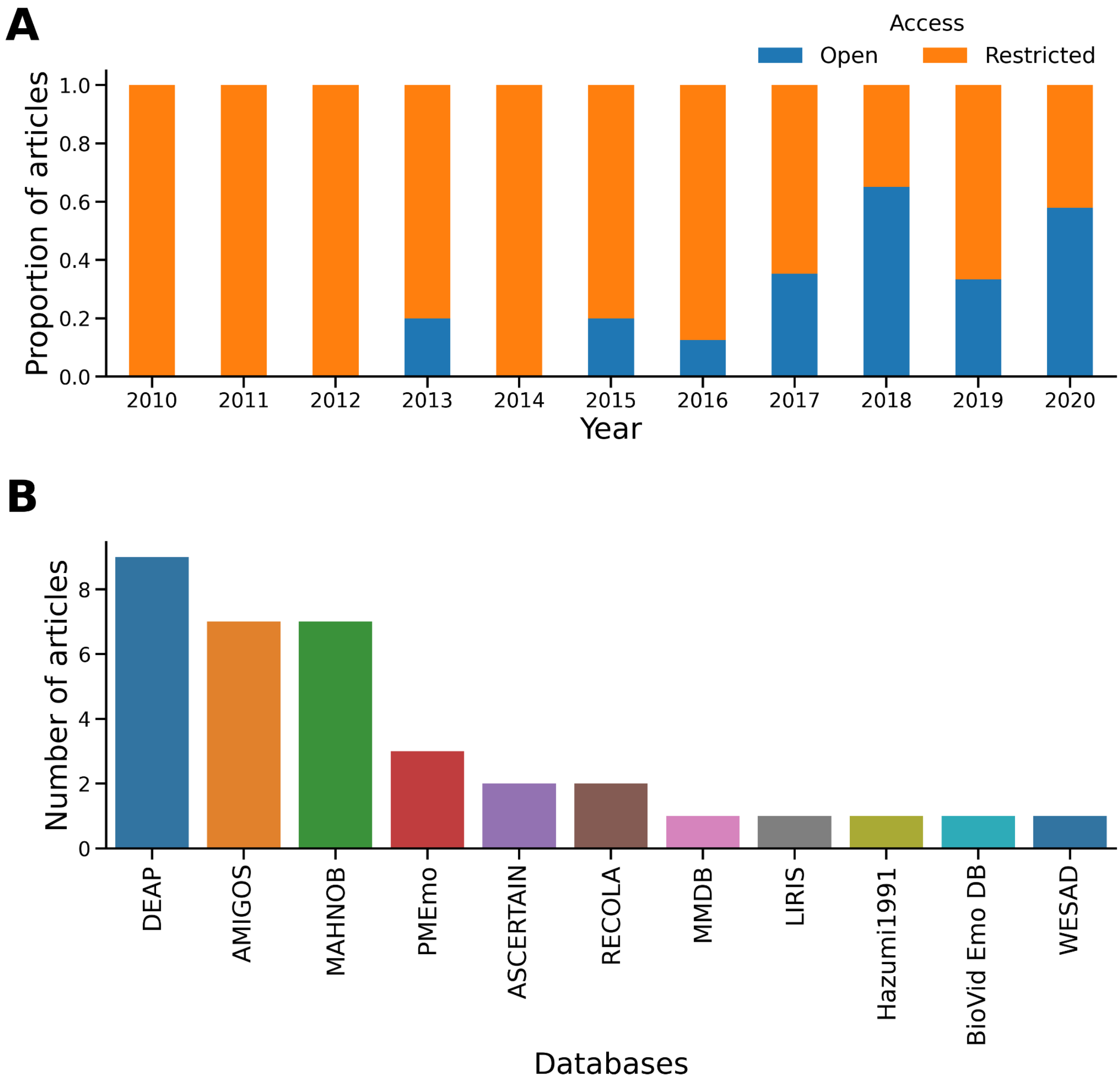

Figure 4. Trends and predominance in database usage for emotion recognition research with EDA. A. Chronological changes in the proportion of articles based on their accessibility. B. Frequency of use of different databases in the field of emotion recognition research with EDA.

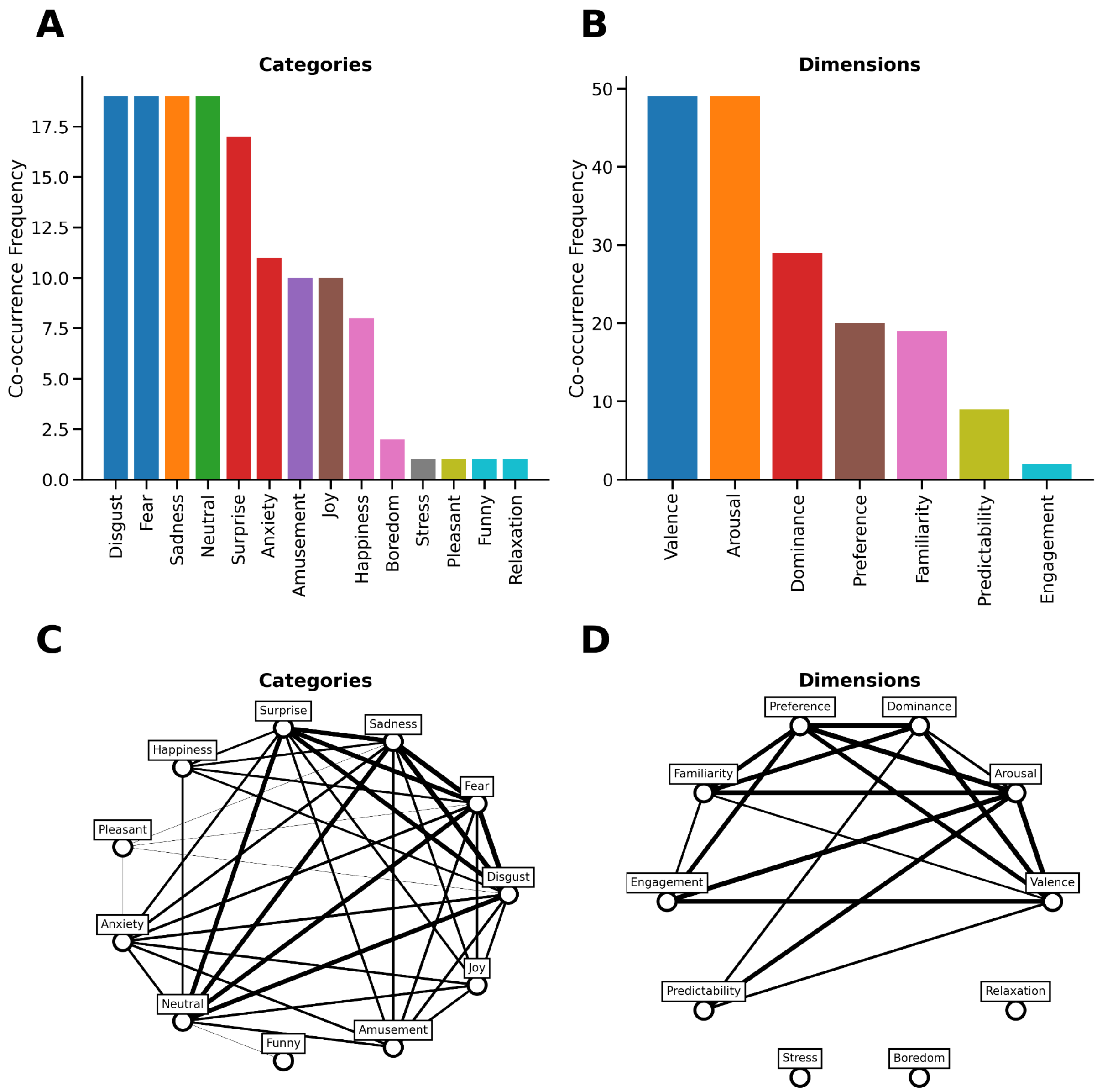

Figure 5. Comprehensive Representation of Emotional Categories and Dimensions in EDA-based Emotion Recognition Models. **A.** Count of emotion categories under examination in EDA-based emotion recognition models. The total count is model-based rather than article-based, which may result in a sum exceeding the total number of articles (99). **B.** Count of emotional dimensions investigated across the models in the field of emotion recognition with EDA, calculated based on the total number of models rather than individual articles. **C.** Visualization of the interconnectedness among emotion categories in emotion recognition models leveraging EDA. **D**. Representation of the connections among emotional dimensions within EDA-based emotion recognition models. *Note.* The thickness of each connecting line reflects the rate of their combined assessment.

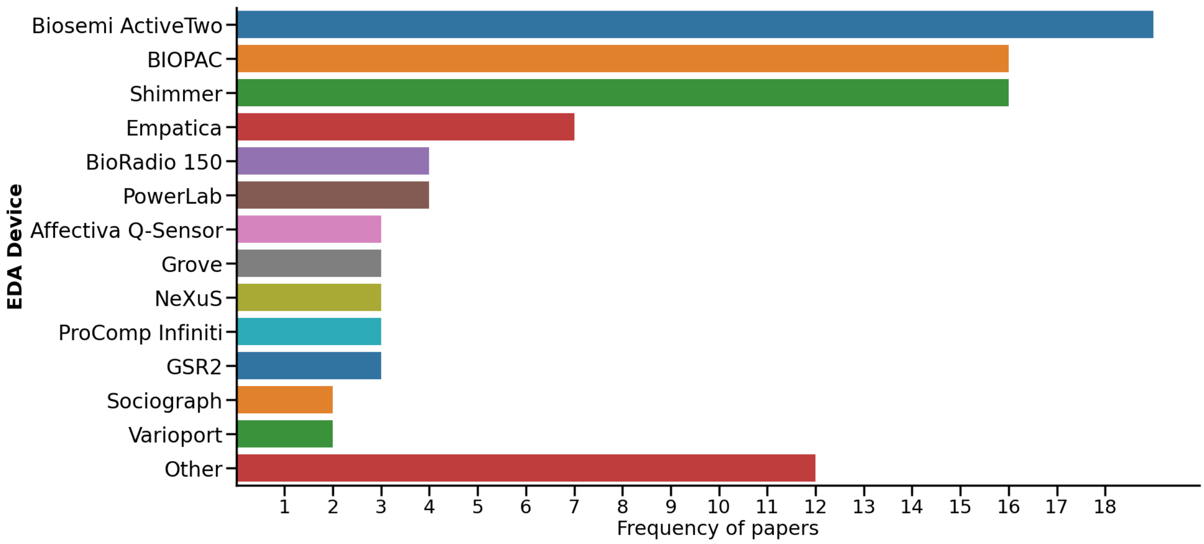

Figure 6. Most frequently used EDA devices in emotion recognition research with EDA. *Note.* A category labeled as 'Other'' accumulates all devices that were individually used only once.

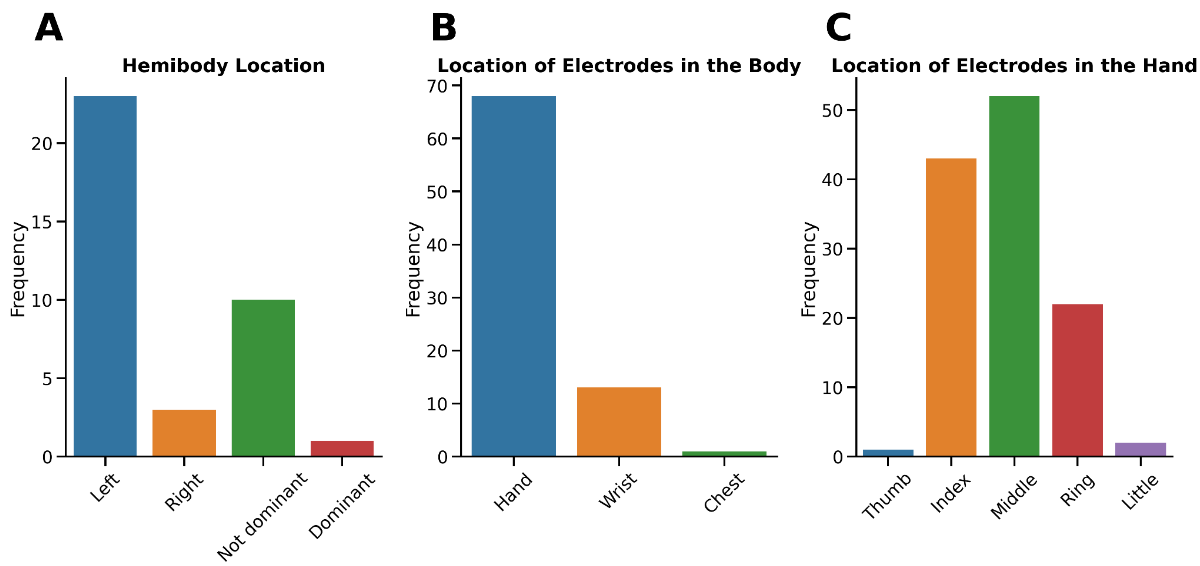

Figure 7. Three levels of analysis for electrode placement in emotion recognition studies using EDA. The hemibody level represents lateralization, the body level indicates the specific body part where electrodes are placed, and the fingers level details the specific finger chosen for electrode placement when the hand is the

selected body part.

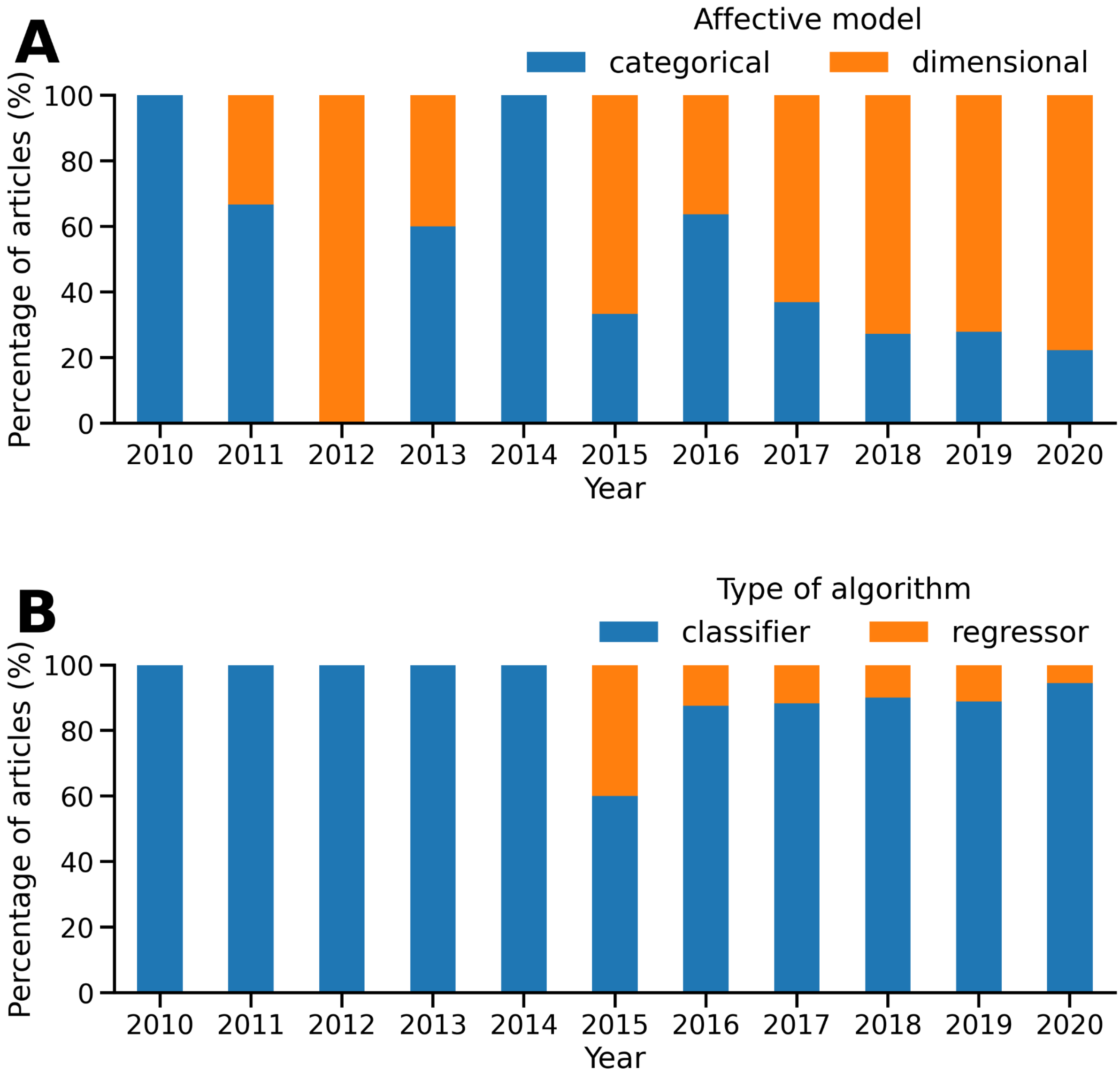

Figure 8. Chronological evolution of emotion model types and algorithm usage in emotion recognition research with EDA over a decade (2010-2020). A. Evolution of usage rates of emotion models, i.e. categorical and dimensional models. B. Evolution of algorithm type usage, i.e. classifier and regressor models.

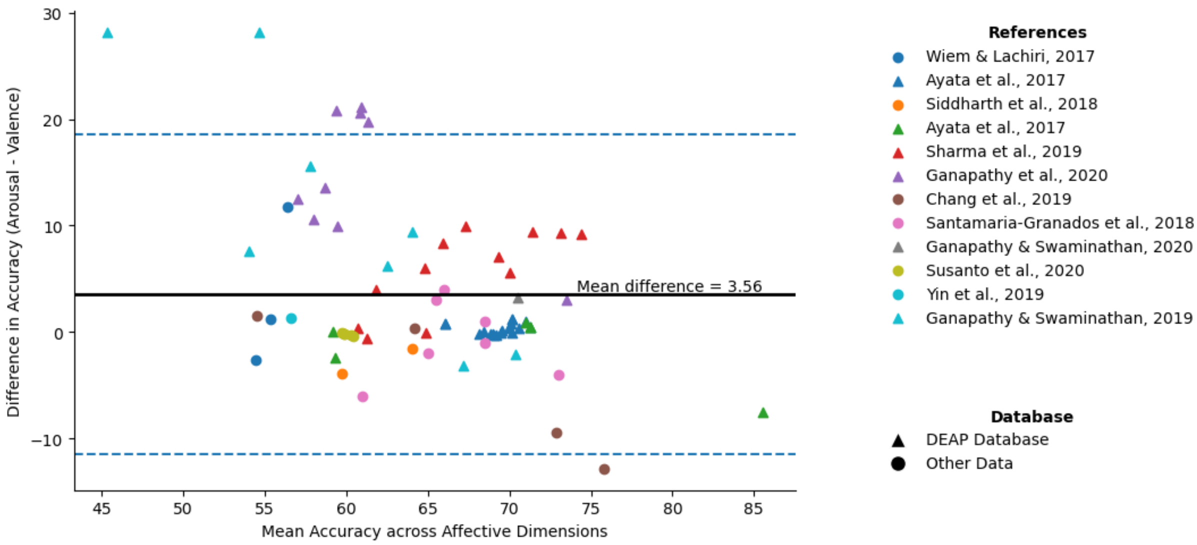

Figure 9. Comparative Analysis of Mean Accuracy and Difference in Accuracy for Arousal and Valence Models Across Studies. The scatterplot displays the relationship between the mean accuracy across affective dimensions (arousal and valence) and the difference in accuracy between these dimensions. Each point represents a unique model from a specific study, with triangles indicating models based on the DEAP database and circles representing models from other databases. A solid black horizontal line indicates the mean difference in accuracy, while dashed horizontal lines represent the 95% confidence interval for this mean difference. The plot also includes legends for study references and database types.