Ph.D. Program in Communication Markets and Society
XXXVIII Cycle

# Affect Detection in Consumer Neuroscience: a Multimodal Machine Learning Model based on bioelectrical and biometric signals

CANDIDATE
**Marco Bilucaglia**

REGISTRATION NUMBER
**1039972**

TUTOR
**Prof. Vincenzo Russo**

CO-TUTOR
**Prof. Luca T. Mainardi**

COORDINATOR
**Prof. Stefania Romenti**

Academic Year 2025/2026

# Dedications

To Anna, without further words, because no words could ever contain what you have been, are, and will always be to me. To Viola, my faithful and patient companion through the silence, the late nights, and the long days. To Matilde, who arrived at the very end of this journey and gave it a whole new meaning—and, quietly, also a new name. All of you have been the true reason I was able to keep going, through every difficulty, doubt, and hard moment along the way.

# Acknowlegments

To my long-standing colleagues Alessandro, Fiamma, and Mara, and to Riccardo, who was there from the very beginning and will always remain one of us, for the moments we shared beyond the laboratory, which allowed me to truly call them “friends”.

To Vincenzo and Margherita, who mentored me both scientifically and humanly throughout these years of research. To Enrico, for supporting my time with professional trust, and for making it possible for part of this work to take shape.

To my research companions Chiara, Cristina, Giuseppina, Federica, Sebastiano, Simone, and Valeria, with whom I shared experiments, ideas, difficulties, and pieces of this path.

To Prof. Tressoldi, who trusted me enough to let me take my first steps with EEG signals when I still had much to learn, and who, in doing so, opened the door to my academic journey.

# Contents

# List of Figures

# List of Tables

# List of Acronyms

**2F** Two-Factor theory

**ACC** Anterior Cingulate Cortex

**AD** Affect Detection

**ANN** Artificial Neural Network

**ANS** Autonomic Nervous System

**ANOVA** Analysis of Variance

**ASR** Artefact Subspace Reconstruction

**AUC** Area Under the Curve

**BF** Bayes Factor

**CAN** Central Autonomic Network

**CASH** Combined Algorithm Selection and Hyperparameter optimisation

**CI** Confidence Interval

**CNS** Central Nervous System

**CNN** Convolutional Neural Network

**DL** Deep Learning

**DT** Decision Tree

**DWT** Discrete Wavelet Transform

**ECG** Electrocardiogram

**EEG** Electroencephalogram

**EMG** Electromyogram

**EOG** Electrooculogram

**ERM** Empirical Risk Minimisation

**ESB** EEG Synchronisation Box

**ET** Eye Tracking

**fMRI** Functional Magnetic Resonance Imaging

**fNIRS** Functional Near-Infrared Spectroscopy

**GD** Gradient Descent

**GS** Gold Standard

**HF** High Frequency

**HRV** Heart Rate Variability

**HVHA** High Valence–High Arousal

**HVLA** High Valence–Low Arousal

**I DARE** IULM Dataset of Affective Responses

**IAF** Individual Alpha Frequency

**IAPS** International Affective Picture System

**IC** Independent Component

**ICA** Independent Component Analysis

**iEMG** Integral value of the EMG signal

**iid** Independent and Identically Distributed

**kCV** k-fold Cross-Validation

**kNN** k-Nearest Neighbours

**LDA** Linear Discriminant Analysis

**LF** Low Frequency

**LF/HF** Low-Frequency/High-Frequency ratio

**LOMO** Leave-One-Modality-Out

**LOO** Leave-One-Out

**LOSO** Leave-One-Subject-Out

**LPSO** Leave-P-Subjects-Out

**LVHA** Low Valence–High Arousal

**LVLA** Low Valence–Low Arousal

**MAD** Mean Absolute Deviation

**MAP** Maximum A Posteriori

**MAV** Mean Absolute Value

**MFCC** Mel-Frequency Cepstral Coefficient

**MATILDE** Multimodal Biosignal-based Modality-resilient Subject-independent Model for Affect Detection

**ML** Machine Learning

**MRMR** Minimum Redundancy Maximum Relevance

**MVPA** Multivariate Pattern Analysis

**NA** Negative Affect

**NAc** Nucleus Accumbens

**NB** Naive Bayes

**NCA** Neighbourhood Component Analysis

**NST** Nucleus of the Solitary Tract

**OASIS** Open Affective Standardized Image Set

**OFC** Orbitofrontal Cortex

**OVO** One-vs-One

**OVR** One-vs-Rest

**PA** Positive Affect

**PAG** Periaqueductal Gray

**PANA** Positive Affect–Negative Affect

**PCA** Principal Component Analysis

**PDF** Probability Density Function

**PFC** Prefrontal Cortex

**PNS** Peripheral Nervous System

**PPG** Photoplethysmogram

**PSD** Power Spectral Density

**PW** Parzen Window

**QDA** Quadratic Discriminant Analysis

**RAM** Random Access Memory

**RBF** Radial Basis Function

**REST** Reference Electrode Standardization Technique

**RF** Random Forest

**RMS** Root Mean Square

**RMSSD** Root Mean Square of Successive Differences

**SAM** Self-Assessment Manikin

**SC** Skin Conductance

**SCL** Skin Conductance Level

**SCR** Skin Conductance Response

**SCT** Sequential Check Theory

**SDNN** Standard Deviation of Normal-to-Normal intervals

**SGD** Stochastic Gradient Descent

**SMA** Spectral Magnitude Area

**SRM** Structural Risk Minimisation

**STFT** Short-Time Fourier Transform

**SVM** Support Vector Machine

**TF** Time–Frequency

**tML** Traditional Machine Learning

**TTL** Transistor–Transistor Logic

**VC** Vapnik–Chervonenkis

**vmPFC** Ventromedial Prefrontal Cortex

**WS** Within-Subject agreement

**XGBoost** Extreme Gradient Boosting

# Abstract

Emotions are integral to human cognition and behaviour, shaping attention, decision-making, and social interaction. In marketing, they act as antecedents, outcomes, mediators, and moderators of consumption-related processes, influencing how products, brands, and experiences are perceived and evaluated. Traditional emotion-measurement approaches, such as self-reports and behavioural observations, are affected by well-known biases, and physiological signals have been proposed as a more reliable, objective, subject-independent, and sensitive alternative.

Consumer Neuroscience commonly attempts to estimate marketing-related cognitive and emotional states from physiological signals through correlational evidence and reverse-inference reasoning. However, this approach is problematic when physiological measures lack specificity with respect to the psychological states. Affect Detection (AD), based on machine-learning models trained on multimodal physiological and biometric features, represents a promising alternative, but remains under-represented in Consumer Neuroscience.

The present thesis contributes to this area through the development, testing, and validation of a multimodal subject-independent AD model.

For this purpose, it first introduces I-DARE (IULM Dataset of Affective REsponses), a standardised multimodal dataset of emotional responses to static visual stimuli, combining self-report evaluations with electroencephalographic, skin-conductance, photoplethysmographic, electromyographic, and eye-tracking signals. I DARE includes 63 participants exposed to 32 static images across five recorded modalities, yielding a dataset dimension of 10080 and an associated minimum detectable effect size of $f = 0.095$, thereby surpassing existing datasets in terms of dimensionality and statistical sensitivity. Self-report evaluations showed fair agreement both within subjects ($\kappa_{WS} = 0.406$) and between subjects and reference labels ($\kappa_{GS} = 0.409$).

Based on I DARE, MATILDE (Multimodal biosignAl-based modaliTy-resilient subject-Independent modeL for affect DEtection) is proposed as a subject-independent multimodal AD

model for classifying affective responses in terms of valence, arousal, and valence–arousal quadrants. Its development combines multi-domain feature extraction, and automatic feature reduction, model selection and hyperparameter optimisation through Combined Algorithm Selection and Hyperparameter optimisation (CASH). Additionally, it provides robust error estimation by means of cross-validation and concentration inequalities. The best-performing models, based on Random Forests, achieved accuracies of 76.2% for quadrant classification, 83.4% for valence, and 85.3% for arousal. These results fall within the upper range of AD models based on traditional machine-learning classifiers.

The robustness of MATILDE is then assessed through dataset-level ablation studies. Both unimodal and Leave-One-Modality-Out configurations are considered using two complementary strategies: fixed-pipeline ablation, in which the classifier, hyperparameters, and feature-selection procedure of the full model are preserved, and re-CASH ablation, in which each reduced-modality configuration is treated as an independent model-selection and hyperparameter-optimisation problem. Across both strategies, most reduced-modality configurations showed no detectable degradation relative to the full multimodal models, whereas eye tracking alone was consistently the least informative modality. Re-CASH analyses further showed that preserving performance under modality reduction may require more complex model structures, especially for arousal classification. These results suggest that the proposed model can scale down to reduced-modality settings, including low-cost and wearable scenarios.

# Chapter 1

# Introduction

Emotions pervade human life: recent experience-sampling evidence suggests that they are experienced during at least 90% of waking time (Trampe et al., 2015).

The inquiry into emotions spans from ancient narratives to contemporary science. Classical texts such as the *Epic of Gilgamesh*, *Genesis*, the *Iliad*, and the *Mahabharata* testify to the centrality of affect in human affairs (Shields & Zawadzki, 2012). Philosophical traditions elaborated competing accounts – from the Stoics and Epicureans, to Plato and Aristotle, to Descartes and Hume – regarding the relation between reason and passion, and the constituents of emotional episodes (Dabrowski, 2016). A scientific reconceptualisation emerged with Darwin's *The Expression of the Emotions in Man and Animals* (Darwin, 1872), anticipating ideas about discrete emotions, their behavioural expressions, and evolutionary continuity (Ekman, 2009). In psychology, early models framed emotions as bodily or neural responses to eliciting events (Bard, 1929; Cannon, 1927; James, 1884; Lange, 1885), while later appraisal approaches emphasised the role of multiple cognitive evaluations (Arnold, 1960; Lazarus, 1966; Schachter & Singer, 1962). Early neuroscientific works (MacLean, 1949, 1970; Papez, 1937) revealed distributed, hierarchical processes spanning cortical and subcortical systems. Modern network-based theories have also underscored dynamic feedback loops with the autonomic nervous system (Kreibig, 2010; Panksepp, 1982, 1992).

Beyond individual experience, emotions function as social signals that modulate observers' cognition and behaviour (van Kleef & Côté, 2022) and systematically bias judgment and choice (Lerner et al., 2015). In marketing, they operate as antecedents, outcomes, mediators, and moderators of consumption-related processes (Bagozzi et al., 1999), and have been extensively investigated over several decades. Based on 453 articles published between 1981 and 2023 in 20 leading marketing journals, Aeron and Rahman (2024) applied bibliographic coupling and content analysis to organise the fragmented literature on emotions in marketing into coherent research streams, grouped around theory, characteristics, and context. According to their synthesis model (see Figure 1.1), emotions arise from marketing-related sources, such as appeals, offerings, and contexts, as well as from individual appraisals, physiological changes, subjective experiences, and social configurations. Whether integral or incidental, emotions influence consumer outcomes through cognitive and affective mediators and contextual moderators, leading to evaluative, regulative, and expressive behaviours, with potential implications for happiness and well-being (Aeron & Rahman, 2024).

Emotions are multifaceted phenomena that encompass subjective experiences, behavioural expressions, and physiological responses (Mauss & Robinson, 2009). Assessing them is therefore inherently challenging, as each component poses distinct methodological limitations. Behavioural observations tend to prioritise within-person experimental effects over between-person consistency (Dang et al., 2020), while self-reports are vulnerable to measurement error and social desirability biases (B. C. K. Choi & Pak, 2004). Physiological measures, in turn, are frequently presented as a more reliable alternative, being relatively objective, subject-independent, and sensitive (Baig & Kavakli, 2019).

Psychophysiology formalises the relationship between psychological states $\psi$ and corresponding physiological signals $\phi$ (Cacioppo & Tassinary, 1990). Applied to marketing, this programme underlies Consumer Neuroscience (Nemorin, 2016; Ramsøy, 2019; Yoon et al., 2006). A common approach in this field relies on reverse inference, whereby $\psi$ are inferred from $\phi$ on the basis of correlations reported in prior observational studies (Plassmann et al., 2015). Because $\phi$ often lacks specificity with respect to $\psi$, this approach can be fallacious (Poldrak, 2006).

Machine Learning (ML) provides a data-driven framework for estimating the probabilistic mapping from $\phi$ to $\psi$. When applied to emotion recognition, ML defines the field of Affective Computing and, more specifically, Affect Detection (AD) (Calvo & D'Mello, 2010; Picard, 1995, 1997). Despite demonstrable advantages over the reverse inference fallacy (Nathan & Del Pinal, 2017; Poldrack, 2011), AD remains underutilised in Consumer Neuroscience (Bilucaglia et al., 2025).

This work contributes to Consumer Neuroscience through the development, testing, and validation of a multimodal AD model based on bioelectrical and biometric signals. This objective is pursued through three steps. First, I DARE (IULM Dataset of Affective REsponses) is introduced

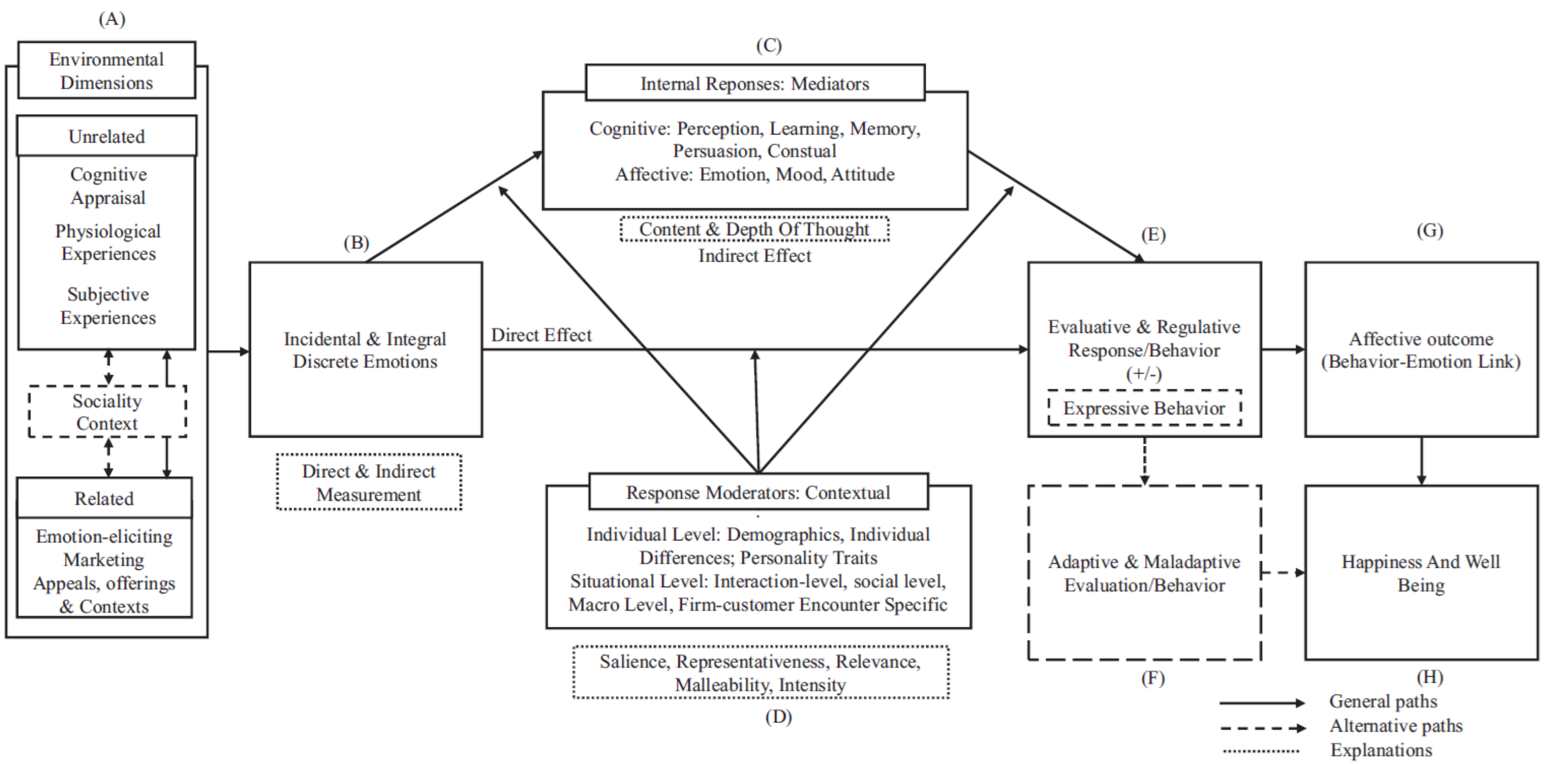


Figure 1.1: Integrative model of emotions in marketing. The model considers emotions as arising from marketing-related and individual determinants and influencing consumer outcomes through mediating and moderating mechanisms, with implications for behaviour, happiness, and well-being – from Aeron and Rahman (2024)

as a standardised multimodal dataset of emotional responses to static visual stimuli, combining self-report evaluations with bioelectrical and biometric recordings. Second, I DARE is used to develop and evaluate MATILDE (Multimodal biosignAl-based modaliTy-resilient subject-Independent modeL for affect DEtection), a subject-independent AD model for classifying affective responses in terms of valence, arousal, and combined valence–arousal configurations. Third, the robustness of MATILDE is assessed through modality-wise ablation studies, showing its resilience under reduced-modality configurations.

The thesis is organised into two main parts. The first part reviews the theoretical and methodological foundations, starting from psychological and physiological theories of emotion (Chapter 2). It then presents Consumer Neuroscience, the problem of reverse inference in psychophysiology, and AD (Chapter 3). Finally, it introduces the ML concepts required to understand the proposed modelling approach in depth, including statistical learning, model selection, and error estimation (Chapter 4). The second part presents the experimental and methodological contribution. It first describes the design, collection, and data consolidation of I DARE (Chapter 5). It then details feature extraction, model training, testing, and evaluation within MATILDE (Chapter 6). Finally, it assesses the robustness of MATILDE through ablation studies, considering both unimodal and Leave-One-Modality-Out configurations under fixed-pipeline and re-optimised settings (Chapter 7). Conclusions are presented in Chapter 8.

# Chapter 2

# Psychological and Physiological Perspectives on Emotion

This chapter examines emotions through psychological and physiological perspectives, tracing their evolution from classical to modern theories. It reviews the transition from early psychological models, such as the James–Lange and Cannon–Bard theories, to cognitive and appraisal-based frameworks, and then extends the discussion to physiological and neural models, from the Papez circuit to contemporary network and embodied approaches.

## 2.1 Defining emotions

According to a survey administered to 35 leading scholars from behavioural, cognitive, and computational neuroscience, including artificial intelligence and robotics, as well as clinical, cognitive, developmental, and social psychology, emotions were most commonly defined as a set of "neural circuits, response systems, feeling states and processes" that "motivate and organize cognition and action" (Izard, 2010).

Cognitive science frames emotions as "mental experiences" that exhibit a four-axis structure typical of sensations: time (i.e., limited duration after the eliciting stimulus), quality (including sensory inputs and imagination/memory processes), intensity (ranging from barely noticeable to highly arousing), and hedonic value (pleasure/displeasure). In particular, emotions are considered to be *any* mental experience characterised by both high intensity and pronounced hedonic tone, whether pleasant or unpleasant (Cabanac, 2002).

From an evolutionary perspective, emotions emerged "as *superordinate* mechanisms responsible for coordinating suites of other information-processing programs, including those of attention, perception, memory, categorization, learning and energy allocation, as well as the more typically considered elements of physiology and manifest behaviour" (Al-Shawaf et al., 2015).

Philosophical Ethology frames emotions as *reactive* components – or "modal willingness" – that position individuals within external reality, alongside *motivations* – or "modal desires" – that represent their *proactive* counterparts. These are shaped by "the multi-layer of phylogenetic history, then through the individual ontogenetic path" (Marchesini, 2021). They act as an interface between environmental inputs and behavioural outputs, enabling adaptive responses (Keltner & Gross, 1999).

## 2.2 Psychological perspectives

### 2.2.1 Classical theories

The first systematic theory of emotions stems from the independent work of William James (1884) and Carl G. Lange (1885), later integrated into a common framework by Dewey (1895). Similar arguments were independently proposed by the Italian anthropologist Giuseppe Sergi (1894). The James–Lange theory proposes that emotions arise from the perception of bodily changes automatically triggered by emotionally relevant stimuli, thereby inverting the traditional view of emotional order. Whereas folk theory assumed that emotions precede physiological responses, James and Lange argued that bodily reactions occur first, and emotional experience is the conscious awareness of these changes (see Figure 2.1).

Cannon (1927) and Bard (1929) challenged the James–Lange perspective, highlighting several empirical inconsistencies. They observed that emotions persist even when visceral–cerebral connections are severed, while direct manipulation of visceral activity fails to generate emotion. Furthermore, they noted that visceral responses are relatively slow, undifferentiated, and occur in both emotional and non-emotional states, undermining their role as unique markers. From these findings, the Cannon–Bard theory concluded that emotional experience and bodily reactions are parallel and independent – rather than causally linked – processes (see Figure 2.2).

A later development came with the work of Schachter and Singer (1962), who introduced a cognitive dimension to emotion formation. In their experiments, participants were given artificially induced physiological arousal (e.g., increased heart rate, trembling, flushing) and then

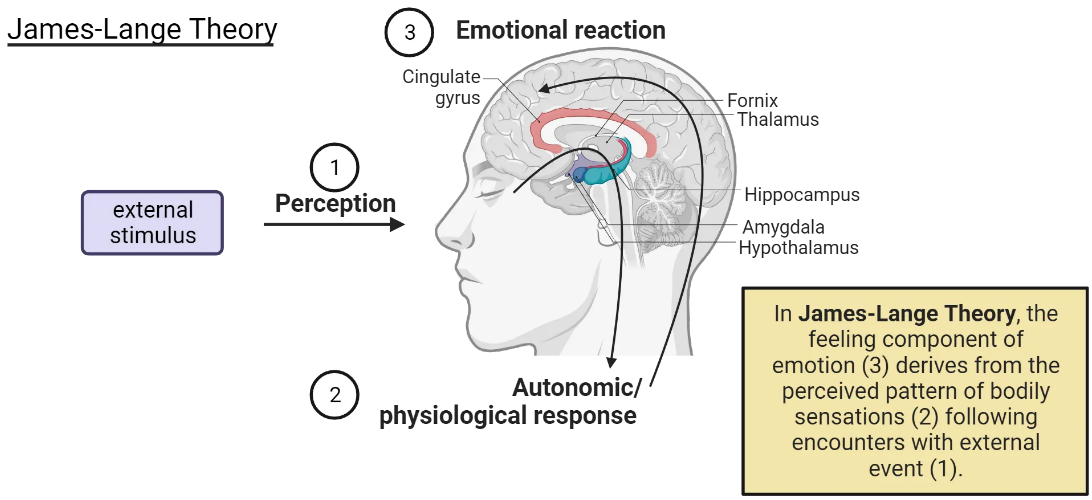


Figure 2.1: James–Lange theory of emotions – from Kirby et al. (2024)

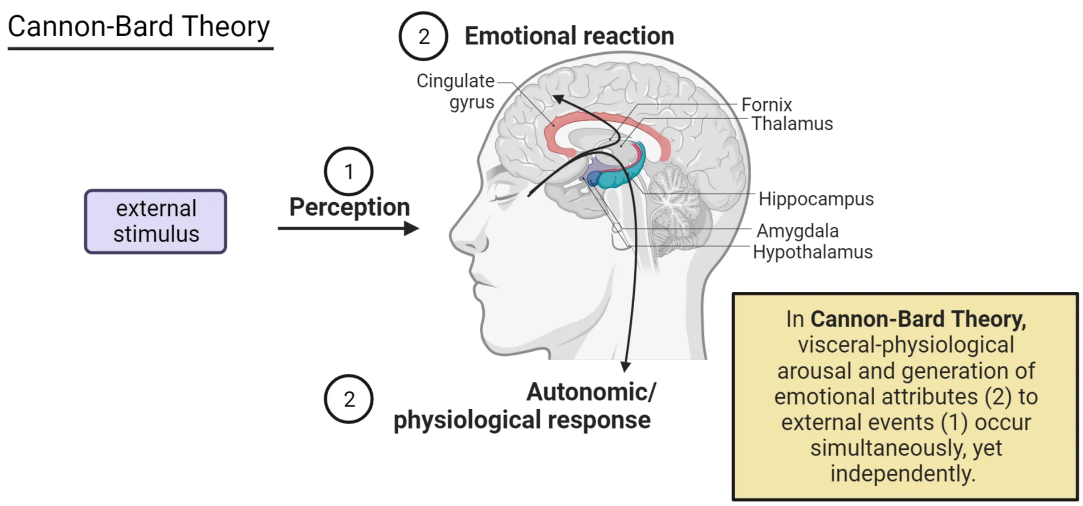


Figure 2.2: Cannon–Bard theory of emotions – from Kirby et al. (2024)

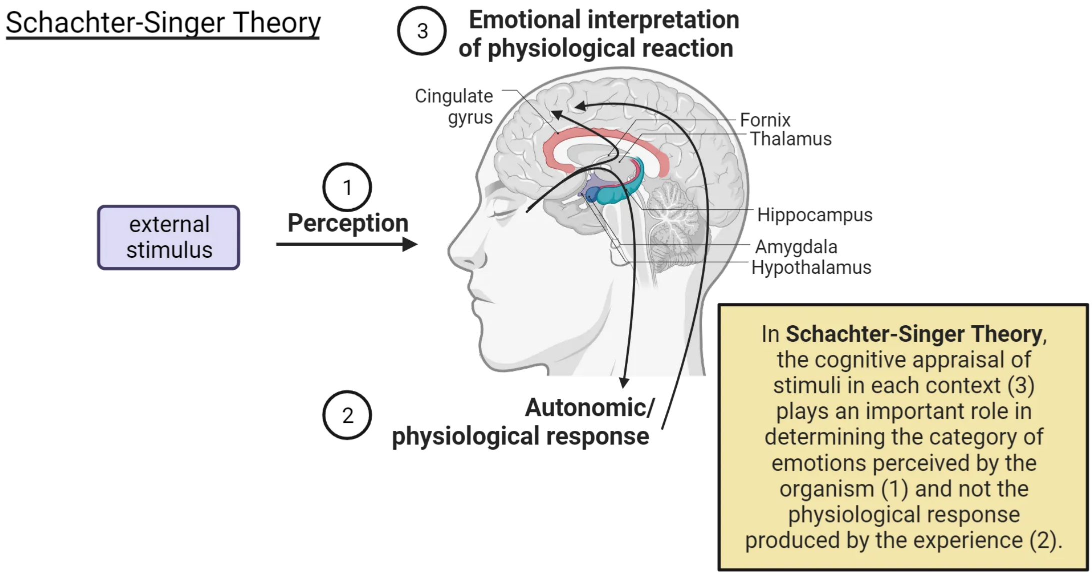


Figure 2.3: Schachter–Singer's two-factor theory of emotions – from Kirby et al. (2024)

exposed to a confederate displaying either euphoric or angry behaviours. Crucially, participants unaware of the arousal's true source adopted the emotional tone suggested by the context. Schachter and Singer argued that emotion results from two components: a state of physiological arousal and the cognitive label assigned to it. Known as the two-factor (2F) theory, this model highlights the interaction between bodily changes and cognitive interpretation, with cognition directing the emotional outcome (see Figure 2.3).

### 2.2.2 Modern view

Appraisal theories are considered the leading emotional perspective in psychology (Scherer & Moors, 2019), affective neuroscience (Mauss & Robinson, 2009), and artificial intelligence (Gratch & Marsella, 2015). Despite emerging in the 1960s, they represent a modern systematisation of ideas by early philosophers such as Aristotle, Hume, Spinoza, and Sartre. This family of theories proposes that emotions arise from *appraisal*, the evaluation of the significance of a stimulus for one's well-being in terms of goals, needs, and values (Moors et al., 2013). Like the 2F theory, appraisal theories assume cognition as an antecedent of emotion, but they emphasise that appraisal is often automatic and, crucially, occurs at the very onset of the emotional process, directly after stimulus perception and before physiological responses (Moors, 2009).

Magda Arnold (1960) is considered the pioneer of appraisal theories. She argued that physiological changes alone cannot generate emotions, stressing instead the role of spontaneous, perceptual-like appraisal in producing immediate feelings of attraction or aversion and corresponding tendencies of approach or withdrawal. Accordingly, the emotion process begins with a stimulus or situation (the *object*)[1] that is perceived and appraised. If the appraisal elicits an approach or avoidance tendency, it generates an emotion that may subsequently give rise to action tendencies, typically manifested through various bodily changes. The action tendencies can be perceived and evaluated through a *secondary appraisal*[2], eventually leading the emotion to grow in intensity and the physical effects to become cumulatively higher.

Richard Lazarus (1966) expanded this view by framing appraisal as a multi-stage process. *Primary appraisal* evaluates potential harm, while *secondary appraisal* assesses coping options. He reinterpreted the consequences of Arnoldian appraisal (action tendencies and physiological responses) as coping processes: affective experiences, motor actions, and physiological reactions are determined by the specific kind of appraisal involved. He reframed Arnold's secondary appraisal as appraised coping and renamed it *reappraisal*.

Lazarus's framework was further expanded with additional appraisal stages (or levels), such as goal relevance, certainty, agency (i.e., attribution of causality), and novelty. In addition to their number, in modern theories appraisal levels also differ in cardinality: finite and categorical, continuous and dimensional, or theoretically infinite but subjectively evaluated in discrete values (Moors et al., 2013). As an example, Scherer's (2001) Sequential Check Theory (SCT) assumes four appraisal steps: relevance (i.e., *does it directly affect me or my social reference group?*), implication (*what are the consequences of this event, and how do these affect my well-being and my immediate or long-term goals?*), coping potential (*how well can I cope with or adjust to these consequences?*), and normative significance (*what is the significance of this event with respect to my self-concept and to social norms and values?*). The order is fixed, but the behavioural and subjective outcomes of each step can influence subsequent evaluations, which in turn modulate the unfolding evaluation process (see Figure 2.4).

[1]The object can be physically present in the environment or even mentally generated (e.g., memory recall, anticipation of a future event, or imagination).

[2]Not to be conflated with Lazarus's secondary appraisal.

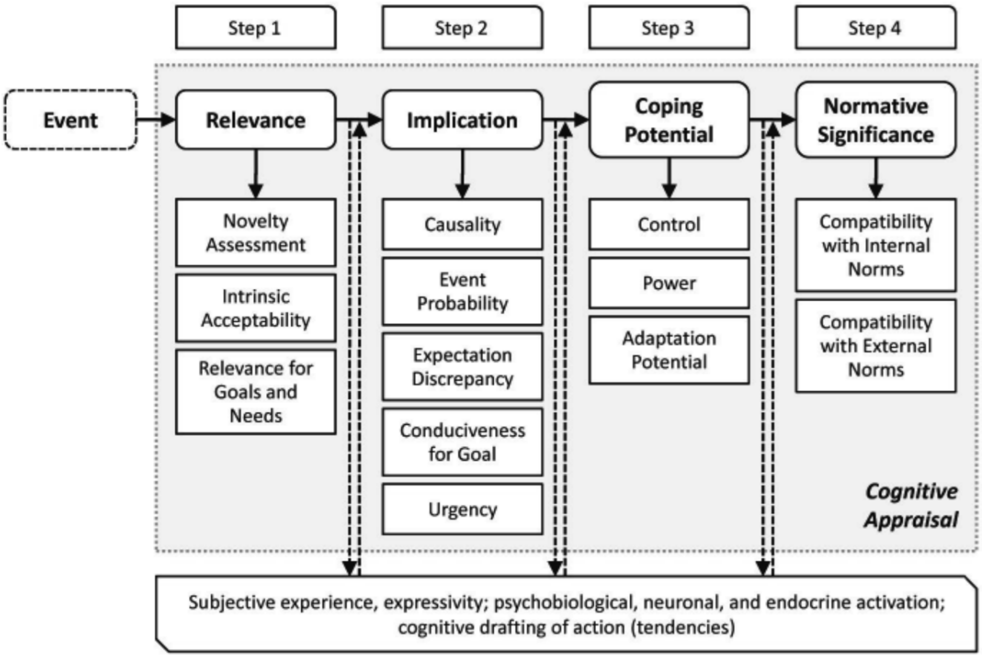


Figure 2.4: Scherer's Sequential Check Theory (SCT) – from Traue et al. (2013)

### Emotion categorisation

Modern theories of emotion converge on the view that both elicitation and differentiation rely on multilevel appraisal, which gives rise to emotional responses that vary in action tendencies (e.g., approach, avoidance), physiological changes (e.g., heart rate, muscle tension), and motor expressions (e.g., facial, vocal, gestural). These responses are then integrated into a non-verbal feeling state, which is subsequently categorised using either *discrete* terms or along *dimensional* axes (Scherer & Moors, 2019).

The discrete view emphasises a small set of basic emotions, each with its characteristic pattern, whereas dimensional approaches describe affective experience as continuously varying along multiple dimensions. This contrast in framing, compounded by differences in terminology – "discrete" versus "continuous", or "categories" versus "dimensions" – can create the impression of a strict opposition. In reality, the two approaches often operate at different descriptive levels and can be seen as complementary: discrete emotions typically occupy specific regions of the dimensional space (Barrett & Westlin, 2021).

**Discrete categorisation** The discrete categorisation of emotions has deep roots in Darwin's comparative treatise, which argued that many emotional expressions are biologically grounded and observable across species and cultures (Darwin, 1872).

Within an explicitly evolutionary framework, Robert Plutchik proposed eight *primary* emotions arranged in bipolar pairs (joy–sadness, trust–disgust, fear–anger, surprise–anticipation), with graded intensities and systematic blends forming *secondary* emotions (Plutchik, 1958, 1962, 1980).

In parallel, Silvan Tomkins proposed a repertoire of innate affective states: positive (interest–excitement, enjoyment–joy), neutral (surprise–startle), or negative (fear–terror, anger–rage, distress–anguish, shame–humiliation, disgust, dissmell). These affects organise motivation and expression and can be amplified or inhibited by control mechanisms (Tomkins, 1962, 1963).

Empirical support for cross-cultural recognition of basic categories was provided by Paul Ekman, who identified specific facial configurations for anger, fear, sadness, happiness, disgust, and surprise (sometimes including contempt). Recognition accuracies varied, but distinct patterns were observed across cultures (Ekman & Friesen, 1971; Ekman et al., 1969).

Izard's Differential Emotions Theory proposed a partly overlapping inventory (interest, joy, surprise, sadness, anger, disgust, contempt, fear, shame, and guilt), emphasising distinctive facial patterns and developmental trajectories (Izard, 1971, 1972).

Panksepp considered emotions as the result of basic subcortical *primary systems* that, when aroused, generate unconditioned action patterns and associated feelings. He initially identified SEEKING (also called "expectancy"), RAGE, FEAR, and PANIC (Panksepp, 1982) systems, later adding LUST, CARE, and PLAY[3] (Panksepp, 1992). These systems are considered *primary* to distinguish them from secondary (learned/conditioned) and tertiary (cognitive-based) processes. Primary systems have specific neuroanatomical and neurochemical substrates, and the behaviours they foster are characterised by either a positive (SEEKING, LUST, CARE, and PLAY) or negative (RAGE, FEAR, and PANIC) valence (Panksepp, 2011).

**Dimensional categorisation** Although first proposed in the 1870s by Wilhelm Maximilian Wundt (Wassmann, 2008), the dimensional categorisation of emotions was pioneered by the experimental work of Russell and Mehrabian (1974, 1977), who analysed the differential seman-

[3]Panksepp capitalises the names of the primary systems to differentiate them from the corresponding affective states, which are left uncapitalised.

tics of affect-related adjectives. They identified a three-factor structure that explained most of the variance in the data: *pleasure–displeasure* (also known as positive–negative valence), *arousal–nonarousal*, and *dominance–submissiveness*. Subsequent studies showed that, rather than spanning the entire three-dimensional space, emotions are primarily distributed along a circular subspace in the valence–arousal plane, termed *circumplex* (Russell, 1980).

Russell and Barrett (1999) adopted the circumplex as the coding scheme for the *core affect*, defined as "the most elementary consciously accessible affective feelings (and their neurophysiological counterparts) that need not be directed at anything" (Russell & Barrett, 1999, p. 806). Core affect represents one of the necessary elements of emotional experience, alongside stimulus-directed processes such as action tendencies, behaviours, appraisal, attention, conscious experience, and neural or bodily changes.

Russell's circumplex was further extended by Robert Thayer, who proposed arousal as the result of two partially independent activation systems: *Energetic Arousal* (vigour–fatigue) and *Tense Arousal* (calm–tension), yielding a two-axis activation–deactivation space for momentary mood (R. E. Thayer, 1989). Energetic arousal tracks circadian and sleep–metabolic state, whereas tense arousal indexes stress and anxiety. This distinction explains common combinations such as "nervous energy" versus "sleepy but keyed up", and highlights the need to separate fatigue or sleepiness from classic autonomic arousal in analysis.

An alternative dimensional model proposed by Watson, Clark, and Tellegen (1988) defined emotions along *Positive Affect* (PA; enthusiastic, active, alert) and *Negative Affect* (NA; distressed, upset). The Positive Affect–Negative Affect (PANA) plane, which accounts for mixed-state emotions (simultaneously high PA and high NA), relates to Russell's circumplex by a simple rotation: valence (pleasant–unpleasant) and arousal (calming–exciting) correspond to the diagonals.

The *vector model of emotion*, proposed by Bradley et al. (1992), conceptualises the emotional space as spanned by two axes pointing in non-opposite directions, forming a "boomerang-like" shape. The model assumes arousal is always present, while valence determines the direction of the emotion. Positive valence shifts the state along the upper vector, negative valence along the lower vector. High-arousal states are differentiated by valence, while low-arousal states cluster near the meeting point of the vectors.

## 2.3 Physiological perspectives

### 2.3.1 Classical theories

The James–Lange theory traces the whole emotion-generation process back to a chain of sensory perceptions. The emotion-evoking stimulus is sensed by receptors (R in Figure 2.5 – left) and routed (path 1) to appropriate cortical projection areas (C). Then, efferent "reflex currents" (path 2) generate a bodily change (either visceral, V, or muscular, SK M) that is sensed and projected (paths 3–4) to appropriate cortical areas, similarly to the original emotion-evoking stimulus. Finally, when these bodily sensations "combine with it in consciousness", the stimulus is transformed "from an object-simply-apprehended into an object-emotionally-felt" (James, 1884, p. 203).

The Cannon–Bard theory proposed a network of distinct brain regions involved in emotion-generation processes, with special emphasis on the thalamus; hence, it has also been known as the *Thalamic Theory* (Cannon, 1927, 1931). Accordingly, the input of an emotional stimulus goes (path 1 in Figure 2.5 – right) from the receptor (R) to the thalamus (T). There, it may directly excite hypothalamic mechanisms and cause a definite pattern to (path 2) visceral (V) and motor (SK M) regions. Alternatively, it may be regrouped and go to the cortex (C, through path 1'), generating a cortico-thalamic inhibition (path 3). Additionally, a thalamo-cortical path (path 4) may also occur (Bard, 1929, pp. 305–306).

Papez (1937) proposed emotions as generated from a network (now known as the *Papez circuit*) of cortical and subcortical structures. Cortical areas project (cp path in Figure 2.6) to the hippocampal formation (h) and then, by the fornix (f), to the mammillary body (m). Through the mammillothalamic tract (mt), the mammillary body projects to the anterior nuclei of the thalamus (a) and thence by medial thalamocortical radiation to the cortex of the cingulate gyrus (cg) (Papez, 1937, p. 728). The latter, in turn, projects to cortical areas through the internal capsule, closing the loop (Papez, 1937, p. 729).

MacLean (MacLean, 1949) expanded the Papez circuit into a broader system, initially termed "visceral brain" and later renamed "limbic system" (MacLean, 1952). Specifically, the hippocampal gyrus functions as an "affectoceptor cortex", the input gate for interoceptive and exteroceptive signals. The dentate gyrus serves as an integration and correlation centre. Finally, the hippocampus (drawn as a seahorse in Figure 2.7) and the amygdala act as an "affectomotor cortex", projecting to the hypothalamus (HYP), epithalamus, caudate, and septal area. These regions, in turn, project to the brainstem, producing physiological changes. The hypothalamus

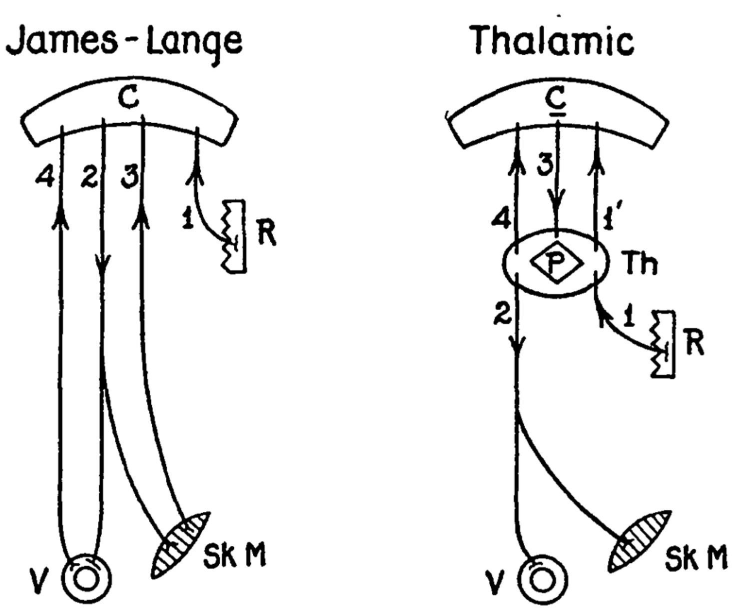


Figure 2.5: Brain areas involved in emotion generation according to James–Lange (left) and Cannon–Bard (right) theories – from Cannon (1931)

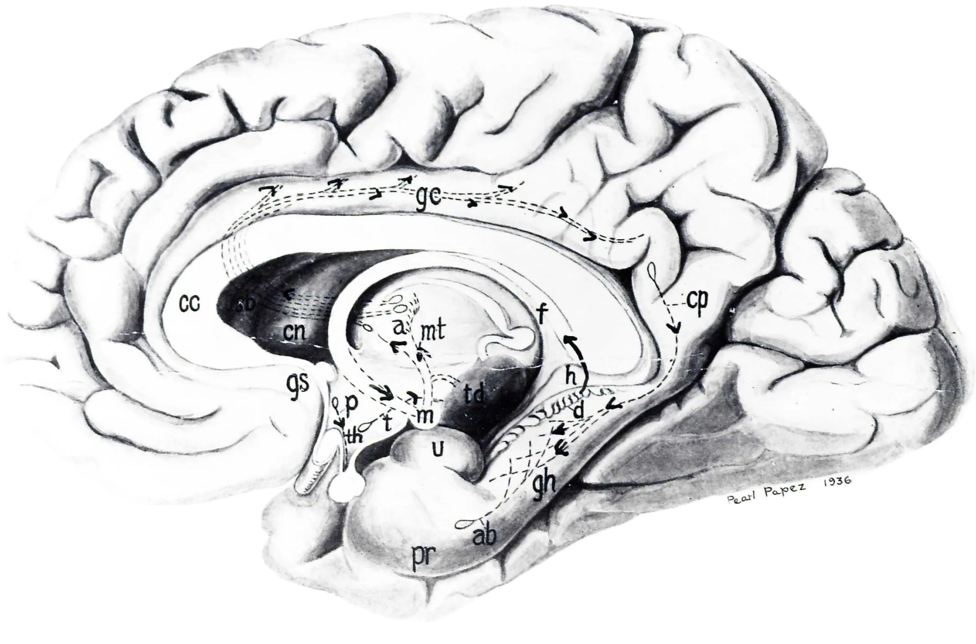


Figure 2.6: Medial view of the right cerebral hemisphere showing the Papez circuit – from Papez (1937).

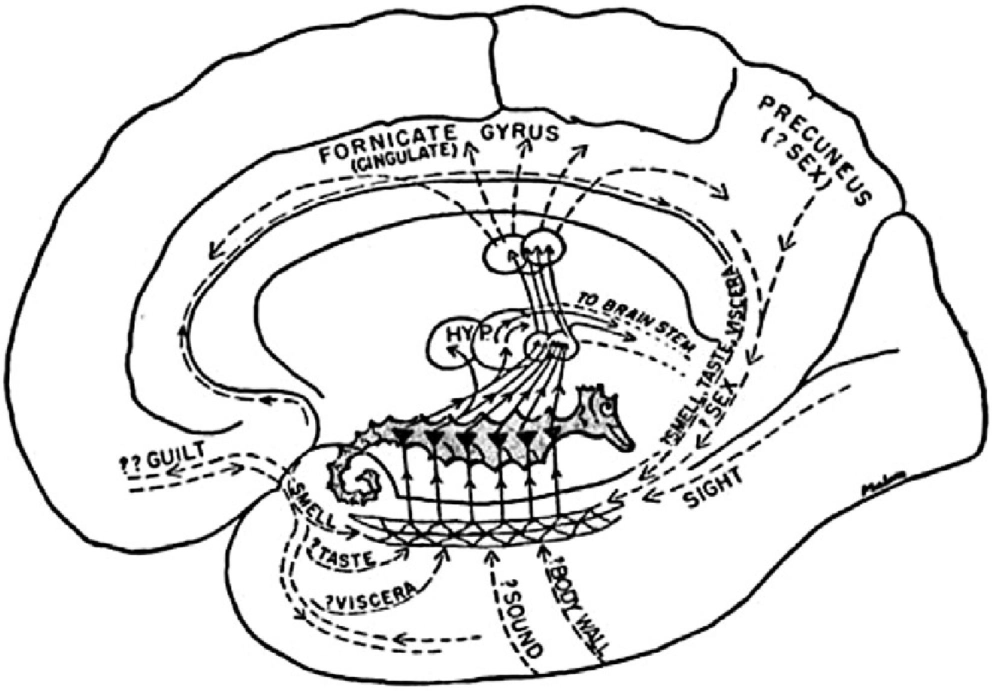


Figure 2.7: Anatomy and connectivity of the visceral brain – from MacLean (1949)

also projects to the anterior thalamus, which in turn projects to the cingulate gyrus (FORNICATE GYRUS), where the affective experience may enter conscious awareness (MacLean, 1949, p. 347).

MacLean also commented on the "animalistic and primitive" nature of the limbic system, as well as its tendency to elude "the grasp of the intellect" (MacLean, 1949, p. 348), suggesting a dichotomy between intellectual and emotional or non-verbal behaviour. This idea was further developed in his later "triune brain" hypothesis (MacLean, 1970). Accordingly, the human brain was described as comprising three evolutionarily layered systems: the "reptilian brain", associated with instinctual and primitive emotional responses; the "old mammalian brain" (the limbic system), which builds upon reptilian functions and enables social emotions; and the "neocortex", which integrates emotional experience with cognition and exerts top–down regulatory control.

### 2.3.2 Modern view

Contemporary accounts generally converge on the view that emotion generation and processing rely on distributed networks of subcortical, cortical, and paralimbic structures engaged across various affective states, reflecting a shared neural architecture that is anatomically widespread and loosely clustered (Malezieux et al., 2023).

This architecture has been interpreted as a hierarchical pipeline, involving both feed-forward and feedback loops across different levels. Subcortical structures handle primitive emotional features (the so-called *primary processes*), limbic circuits support emotional learning and stimulus integration (*secondary processes*), and cortical networks underlie higher-level cognitive processes (*tertiary processes*), with both the cortex and subcortical systems contributing (Panksepp, 2011).

Multilevel appraisal (Moors et al., 2013) can effectively model the brain's involvement in emotion generation and processing. Fast (within 100–350 ms of first perceiving an event) appraisal of novelty and concern relevance is supported by the amygdala; potential sources of reward by the nucleus accumbens (NAc) – part of the ventral striatum – and pain, threat, and harm by the periaqueductal grey (PAG). These structures receive interoceptive and exteroceptive inputs from the thalamus and project to both the hypothalamus and the cortex (Keltner et al., 2018). The amygdala and NAc are also involved in emotional learning and prediction of future affective states (Kirkland & Cunningham, 2011). Body awareness and subjective feelings of visceral changes are sustained by the anterior insular cortex, while higher processes such as emotion conceptualisation and regulation, as well as social cognition, are supported by the prefrontal cortex (PFC) (Keltner et al., 2018). Additionally, the anterior cingulate cortex (ACC) is considered a key substrate of conscious emotional experience (Kemp et al., 2015).

Rolls (2023) proposed a three-tiered organisation of the brain for emotion in primates, including humans (see Figure 2.8). In Tier 1, the brain constructs representations of sensory inputs involving primary and associative sensory areas. In Tier 2, reward value and emotional significance are assigned through the orbitofrontal cortex (OFC) and amygdala, with top–down attentional and cognitive modulation. In Tier 3, actions are learned in the supracallosal (dorsal) ACC to obtain rewards signalled by the OFC and amygdala via the pregenual ACC and ventromedial PFC (vmPFC). Within this tier, the vmPFC also enables decisions between stimuli of different reward values, the OFC connects with reasoning and language systems to integrate and communicate affective values, reinforcement learning supports the development of stimulus–response habits, and autonomic responses can be generated to emotion-provoking stimuli. Additionally, auditory inputs can directly reach the amygdala, and several thalamic nuclei convey somatosensory and taste information to their respective primary cortices.

Emotion generation and processing actively depend on bodily changes that act as more than a mere downstream expression of brain activity. Predictive coding (Barrett & Simmons, 2015)

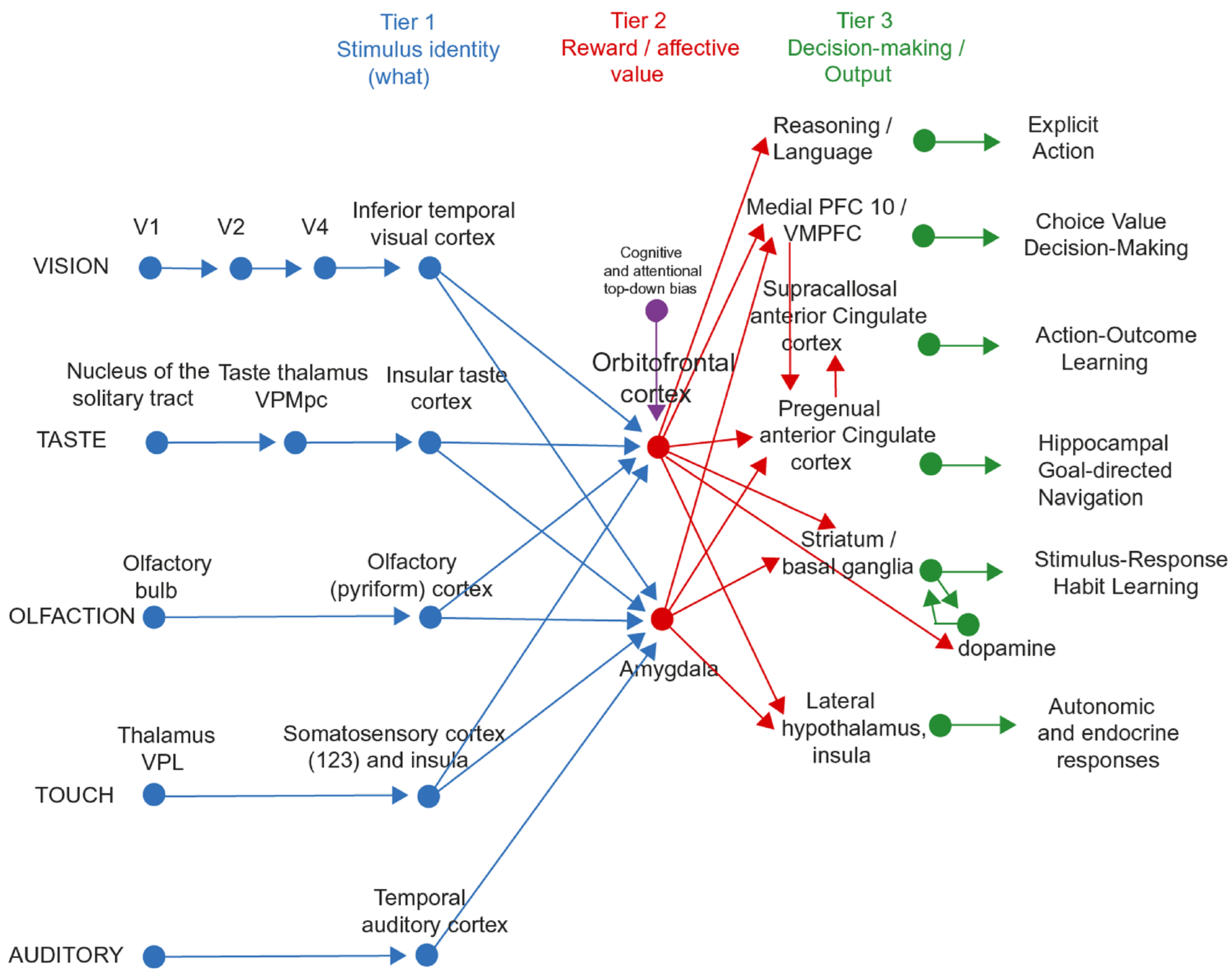


Figure 2.8: Three-tiered organisation of the brain for emotion – from Rolls (2023)

and interoceptive inference (Seth, 2013) models propose that the brain continuously anticipates and updates bodily states, forming a closed-loop *embodied* system of emotional perception and regulation.

Contrary to early views, the autonomic nervous system (ANS) is not an isolated and purely involuntary control system but is interconnected with central neural processes at multiple levels. While segmental reflexes are regulated by the spinal cord, suprasegmental integration within the brainstem is essential for respiration, blood pressure, swallowing, and pupillary control. More complex regulation arises from hypothalamic systems that influence brainstem autonomic subsystems. In turn, hypothalamic activity is modulated by cortical regions – including the insula, anterior cingulate, ventromedial prefrontal cortex, and the central nucleus of the amygdala – which process environmental inputs (Kreibig, 2010).

ANS changes also act as input for the central nervous system (CNS) through interoceptive mechanisms. Signals from bodily organs reach the brainstem nuclei via the vagus nerve and the spinal splanchnic nerve. They are then projected to other structures, including the thalamus, amygdala, hippocampus, insula, cingulate cortex, and sensory cortex (Camara et al., 2025).

Several embodied theories of emotion have been proposed, with the most influential being the Neurovisceral Integration Model, the Polyvagal Theory, the Somatic Marker Hypothesis, and the Homeostatic Model for Awareness (Kemp et al., 2015).

The Neurovisceral Integration Model (J. F. Thayer & Lane, 2000) proposes a Central Autonomic Network (CAN) comprising the PFC, ACC, insula, amygdala, and brainstem, which controls visceral responses to stimuli. The primary output of the CAN is heart rate variability (HRV), mediated by parasympathetic nervous system innervation (vagal inhibition) of the heart. Increased HRV is associated with positive emotions, while decreased HRV is linked to depression and anxiety.

Polyvagal Theory (Porges, 1995) further emphasises the role of vagal afferent feedback from the viscera and internal milieu to the nucleus of the solitary tract (NST) in modulating HRV. It also distinguishes between the myelinated and unmyelinated tracts of the vagus nerve (hence, "polyvagal"), assigning them different roles. The myelinated vagus underpins changes in HRV and approach-related behaviours (e.g., social engagement), while the phylogenetically older unmyelinated vagus, in combination with the sympathetic nervous system, is involved in dangerous or life-threatening events.

The Homeostatic Model of Awareness (Craig, 2002) relates approach behaviours, parasympathetic activity, and affiliative emotions to the left anterior insula and ACC, while it relates withdrawal behaviours, sympathetic activity, and heightened arousal to the right anterior insula and ACC. The suggested underlying brain network is the lamina I spinothalamocortical system, which projects sympathetic and somatic inputs to the spinal cord, brainstem, and, finally, to the insula and ACC, creating a thalamocortical representation of the state of the body.

Finally, the Somatic Marker Hypothesis (Damasio, 1996) highlights the role of the ventromedial PFC in translating the sensory properties associated with external stimuli into "somatic markers", which reflect their biological relevance and guide decision-making. The PFC indexes changes in heart rate, blood pressure, and glandular secretion, thereby contributing to both decision-making and affective experience.

# Chapter 3

# From Consumer Neuroscience to Affect Detection

This chapter describes the approaches of Consumer Neuroscience, Psychophysiology, and Affective Computing to the measurement of human emotions. It traces the development of Consumer Neuroscience as an experimental discipline that incorporates neurophysiological and biometric data to address the limitations of self-report methods. It then outlines Psychophysiology as the theoretical basis linking psychological states with physiological processes, introducing probabilistic and inferential models for interpreting brain–behaviour relationships. Finally, it examines Affect Detection within Affective Computing, where machine-learning models classify emotional states from multimodal biometric data.

## 3.1 Consumer Neuroscience

The term *Consumer Neuroscience* first appeared in a 2006 scientific paper, where it referred to the attempt "to investigate consumer behaviour and decision-making at the neural level (...) to complement more traditional methods of inquiry" (Yoon et al., 2006). A few years earlier, Ale Smidts had coined the term *Neuromarketing* to identify a novel discipline aimed at "better understand[ing] customers and their response to marketing stimuli by directly measuring processes in the brain and including them in theory building and stimulus development" (Smidts, 2002). Although the two terms are often used interchangeably, they were originally intended to differ in scope: Consumer Neuroscience denoted a primarily theoretical and academic field, whereas

Neuromarketing referred to a more practical and commercially oriented application (Ramsøy, 2019). Accordingly, Consumer Neuroscience has also been described as the scientific foundation of Neuromarketing, understood as the application of neuroscientific findings to managerial practice (Nemorin, 2016).

Their emergence was driven by the low reliability of "traditional" marketing techniques such as questionnaires, interviews, and focus groups. As these methods rely on self-reports, they depend on indirect and rationally reprocessed measures of individual experience and consequently exhibit limited predictive value (Hakim & Levy, 2018). The uncertainty associated with such measures predictably affects the success of commercial campaigns: between 40% and 80% of new products launched on the market fail, causing economic losses amounting to billions of dollars (Castellion & Markham, 2013).

Although Consumer Neuroscience is a relatively young discipline, it has gained increasing popularity within the scientific community. As of 2022, 862 articles had been published in international journals across diverse fields such as psychology (28.45%), business (24.75%), communication (24.08%), and neuroscience (15%), as well as management, information technology, and behavioural sciences (7.72%). Scientific production has increased at an annual growth rate of 7.5%, involving 2,354 contributors – an average of 3.83 co-authors per paper, with 24.97% of studies featuring international collaborations (Casado-Aranda et al., 2023).

Consumer Neuroscience is an experimental discipline. It collects neurophysiological and biometric signals (decoded into "neurometrics") from participants exposed to marketing stimuli or tasks to assess underlying cognitive and emotional processes (Cherubino et al., 2019). The most frequently recorded signals include the electroencephalogram (EEG; 56.14%), eye tracking (ET; 10.53%), skin conductance (SC), and cardiac activity (8.77%), as well as the electromyogram (EMG; 1.75%) (Rawnaque et al., 2020). Facial expressions are also commonly recorded (Mancini et al., 2023). Typical stimuli include video advertisements (Russo et al., 2022, 2023; Zito et al., 2021), radio commercials (Russo et al., 2020), product packaging (Russo et al., 2021), static visual creatives (Ciceri et al., 2020), and websites (Bengoechea et al., 2023). Common tasks involve purchasing processes (Cherubino et al., 2017), in-store experiences (Balconi et al., 2021), and food choices (Stasi et al., 2018). Memory, attention, and approach–withdrawal motivation are the main psychological processes assessed (Cherubino et al., 2019), though emotions, above all, have attracted the greatest scientific interest (Casado-Aranda et al., 2023).

## 3.2 Psychophysiology

The interplay between psychological and physiological phenomena is traditionally investigated within Psychophysiology, defined as "the scientific study of social, psychological and behavioural phenomena as related to and revealed through physiological principles and events in functional organisms" (Cacioppo et al., 2014, p. 6). Historically (Stern, 1964), Psychophysiology differs from the closely related Biological Psychology in the direction of the relationship: in Psychophysiology, physiological variables ($\phi \in \Phi$) are outputs of psychological manipulations ($\psi \in \Psi$), while in Biological Psychology the reverse holds. However, this distinction has been criticised as overly restrictive, excluding studies where physiological events serve as blocking variables and behaviour is treated as the dependent measure, as well as studies comparing physiological responses across known groups (Cacioppo et al., 2014, p. 6). For simplicity, we refer to the Psychophysiological approach as *direct* and to Biological Psychology as *inverse* (Bach et al., 2018). The following discussion focuses on the direct approach.

The $\Psi \sim \Phi$ relationship can be described along two orthogonal dimensions: "specificity" and "generality" (see Figure 3.1). Specificity (S) refers to the mapping type, which can be "one-to-one" ($\Psi \to \Phi$) or "one-to-many" (low S; $\Psi \rightrightarrows \Phi$)[1], while generality (G) refers to context dependence, ranging from "situation-" or "person-specific" (low G) to "cross-situational" or "pan-cultural" (high G). Combinations of generality and specificity define four relationship types: "outcome" (low S, low G), "concomitant" (low S, high G), "invariant" (high S, high G), and "marker" (high S, low G). Outcomes are typically the first relationships identified in psychophysiological studies; further experimentation can reclassify an outcome as a marker (once specificity is demonstrated), a concomitant (once generality is shown), or an invariant (once both are established) (Cacioppo & Tassinary, 1990).

Rather than seeking a deterministic relationship $\Phi = f(\Psi)$, a more rigorous probabilistic approach is preferred (Cacioppo et al., 2014, p. 11), consistent with the modern epistemological stance of "dynamic indeterminism" (Neyman, 1960). Relationships are thus treated as *inferences*, whose reliability is expressed by a probability $\mathcal{P}$.

Considering a contextual variable $\gamma \in \Gamma$, the psychophysiological outcome can be represented as the conditional probability $\mathcal{P}(\phi|\psi \cap \gamma)$, read as "the likelihood of observing a physiological

[1] "Many-to-one" and "many-to-many" relations can be reformulated as "one-to-one" and "one-to-many" by restricting the domain to a subset $\Psi' \subset \Psi$.

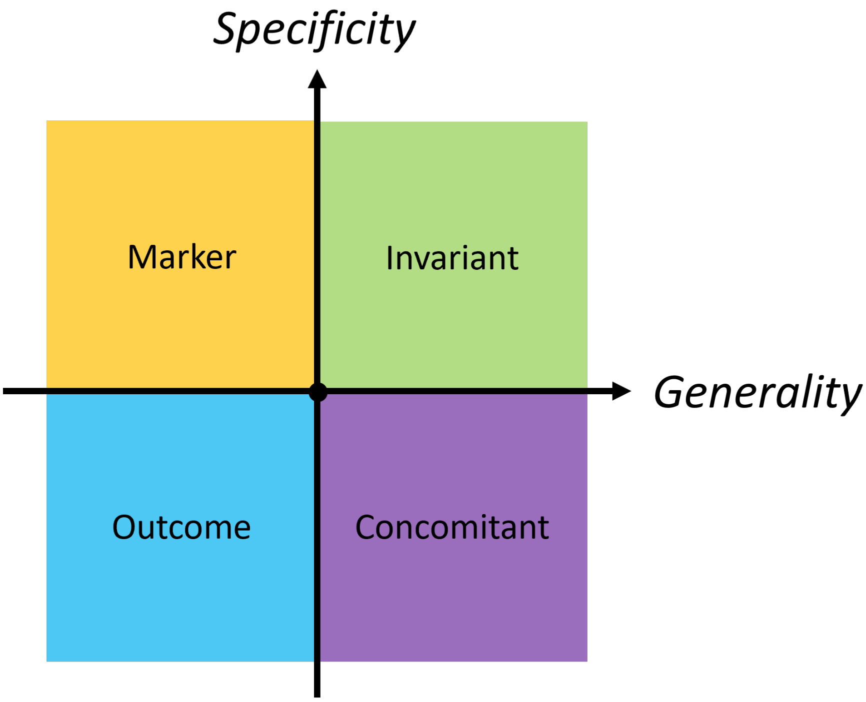


Figure 3.1: Psychophysiological relationships in terms of specificity and generality.

state $\phi$ given the psychological state $\psi$ and the context $\gamma$". For concomitants and invariants (high G), $\phi$ and $\gamma$ can be considered conditionally independent given $\psi$ (Dawid, 1979), and thus $\mathcal{P}(\phi|\psi \cap \gamma)$ simplifies to $\mathcal{P}(\phi|\psi)$. For markers (high S, low G), $\phi$ and $\psi$ can be considered conditionally independent given $\gamma$, yielding $\mathcal{P}(\phi|\psi \cap \gamma) = \mathcal{P}(\phi|\gamma)$. For invariants, the relationship is deterministic, $\mathcal{P}(\phi|\psi) = 1$ (Cacioppo & Tassinary, 1990).

To simplify, we embed generality and specificity within a single variable, $\psi \leftarrow \psi \cap \gamma$. Computations explicitly including $\psi \cap \gamma$ can be found in Hutzler (2014).

The *reverse inference* $\mathcal{P}(\psi|\phi)$ can be derived from the *direct inference* $\mathcal{P}(\phi|\psi)$ using Bayes' theorem and the law of total probability (Riley et al., 2006, p. 1132):

$$\mathcal{P}(\psi|\phi) = \frac{\mathcal{P}(\phi|\psi)\mathcal{P}(\psi)}{\mathcal{P}(\phi)} \equiv \frac{\mathcal{P}(\phi|\psi)\mathcal{P}(\psi)}{\mathcal{P}(\phi|\psi)\mathcal{P}(\psi) + \mathcal{P}(\phi|\bar{\psi})\mathcal{P}(\bar{\psi})}, \tag{3.1}$$

The so-called "reverse inference issue" arises when, regardless of $\mathcal{P}(\phi)$ and $\mathcal{P}(\psi)$, the approximation $\mathcal{P}(\phi|\psi) \approx \mathcal{P}(\psi|\phi)$ is erroneously assumed (Poldrak, 2006). Its logical form is as follows (Machery, 2014, p. 252):

When psychological process $\psi$ is recruited by a task $\mathcal{T}_0$, the physiological pattern $\phi$

is likely to be observed. In another task $\mathcal{T}_1$, the physiological pattern $\phi$ is found. Hence, psychological process $\psi$ was recruited by task $\mathcal{T}_1$.

Assuming uniform priors $\mathcal{P}(\psi) = \mathcal{P}(\bar{\psi}) = 0.5$, the only case where $\mathcal{P}(\psi|\phi) \approx 1$ is when $\mathcal{P}(\phi|\psi) \gg \mathcal{P}(\phi|\bar{\psi})$, meaning that the physiological state $\phi$ is *highly selective* for the psychological process $\psi$.

Otherwise, the likelihood of the reverse inference must be quantified by explicitly considering both posteriors, $\mathcal{P}(\phi|\psi)$ and $\mathcal{P}(\phi|\bar{\psi})$, which can be estimated through meta-analyses or large-scale synthesis databases (Poldrak, 2006). To simplify interpretation, the Bayes factor (BF) can be computed (Hutzler, 2014):

$$BF = \frac{\mathcal{P}(\psi|\phi)}{1-\mathcal{P}(\psi|\phi)}\frac{1-\mathcal{P}(\psi)}{\mathcal{P}(\psi)} \tag{3.2}$$

A positive Bayes factor supports the reverse inference, with thresholds for "moderate", "strong", and "very strong" evidence at 3, 10, and 30, respectively (Schmalz et al., 2023).

Alternatively, multiple measures can be collected as "successive pieces of evidence" $\{\phi_k\}_{k>1}$ to increase overall reliability (Delplanque & Sander, 2021). This approach, known as the "duck test"[2], is widely used in medical diagnosis, where the likelihood of a disease $\psi$ is supported by a collection of symptoms $\{\phi_k\}_{k>1}$ (Nalwanga & Henning, 2016). Using multiple measures to describe a single phenomenon is also central to the *within-method* strategy of *data triangulation* (Bekhet & Zauszniewski, 2012). In Neuromarketing, this corresponds to the "Type-III Inter-Technique Correlational Neurometric Approach" described by Bigne et al. (2025).

Under the assumptions of conditional independence among $\phi_{i\neq j}$ given $\psi$ and mutual independence of $\phi_{i\neq j}$, the reverse inference $\mathcal{P}\left(\psi|\bigcap_k \phi_k\right)$ simplifies to (Charniak, 1983):

$$\mathcal{P}\left(\psi|\bigcap_k \phi_k\right) = \mathcal{P}(\psi)\prod_k \frac{\mathcal{P}(\phi_k|\psi)}{\mathcal{P}(\phi_k)} \tag{3.3}$$

Assuming $\mathcal{P}(\psi) > 0$, the overall reverse inference increases monotonically, approaching 1 when $\mathcal{P}(\phi_k|\psi) \geq \mathcal{P}(\phi_k)$ – that is, when $\psi$ predicts $\phi_k$, even minimally.

Another method for estimating reverse inference involves Multivariate (or Multivoxel) Pattern Analysis (MVPA) (Nathan & Del Pinal, 2017; Poldrack, 2011), which employs Machine

[2]Colloquially summarised as: "If it looks like a duck, swims like a duck and quacks like a duck, then it probably is a duck."

Learning (ML) models to predict $\psi$ from $\phi$ or to discriminate $\psi$ from $\bar{\psi}$ based on $\phi$ (Peelen & Downing, 2023). Free from theoretical assumptions, ML methods can reveal latent data structures, providing new theoretical insights and hypotheses about observed phenomena (Verzelli et al., 2024).

Reverse inference remains a central issue in neuroimaging (Sinnott-Armstrong & Simmons, 2021) and, similarly, in Consumer Neuroscience (Bigne et al., 2025; Hakim & Levy, 2018), where few of the aforementioned solutions have yet been fully implemented. Posterior estimation has not been systematically applied to common Neuromarketing measures, while the use of multiple measures and ML remains limited in scope and dissemination. A recent review found that, on average, only $2.08 \pm 0.86$ (range: 1–4) different measures are used in Consumer Neuroscience experiments, with traditional techniques (60.7%), EEG (49.2%), ET (26.2%), and SC (21.3%) being most prevalent (Gupta et al., 2025). Moreover, only 35% of studies to date have adopted ML or Deep Learning (DL) methods, with adoption rates ranging from 27% to 50% in EEG-focused research (Bilucaglia et al., 2025).

## 3.3 Affect Detection

Affective Computing emerged in the late 1990s as a subfield of Computer Science "that relates to, arises from, or influences emotions" (Picard, 1995, p. 1), with the goal of designing "affective systems" (Picard, 1997, p. 86) capable of "recognising emotions, expressing emotions, having emotions and possessing emotional intelligence" (Picard, 1997, p. 50).

The emotion-recognition task, commonly referred to as *Affect Detection* (Calvo & D'Mello, 2010), has become increasingly pervasive, with applications extending far beyond basic research. In healthcare, it supports the diagnosis, monitoring, and treatment of neurodevelopmental, neurological, and psychiatric disorders (Guo et al., 2024; Pepa et al., 2023; Yoonesi et al., 2025). In education, it is used to enhance student engagement and learning effectiveness in both real and virtual settings, as well as to inform adaptive tutoring systems (Lampropoulos et al., 2024; Llurba & Palau, 2024; Mejbri et al., 2022). In transportation, human–robot (Spezialetti et al., 2020), human–computer (Z. Ding et al., 2024), and brain–machine (H. Chen et al., 2025) interactions, it has been proposed to improve safety, comfort, and overall user experience.

Since its inception, Affect Detection (AD) has been treated as a Machine Learning (ML) problem (Picard, 1995), solvable through models designed to infer *affective states* from measurable biometric variables grouped into *modalities* (Picard, 1997). Specifically, the model $f$ is trained to predict a numerical representation of the affective state $s \in \mathcal{S}$, termed *affect expression*, from a feature vector $\mathbf{m} \in \mathbb{R}^M$ whose elements $\{m_i\}_{i=1}^M$ are scalars extracted from the modalities. The model is expressed as $s = f(\mathbf{m})$.

Depending on the chosen emotion model, the affect expression $s$ may belong to a discrete set $\mathcal{S} \subseteq \mathbb{Z}$ or a continuous set $\mathcal{S} \subseteq \mathbb{R}$, defining a *classification* or *regression* task, respectively. In practice, classification is far more common – even when continuous models are used – by discretising dimensions into arbitrary categories (e.g., low, medium, high). Meta-analyses show that although 63% of studies reference continuous emotion models, 92% of affect detection systems are classifiers (D'Mello & Kory, 2015). This trend is corroborated by other reviews, reporting classifier usage rates of 90%–93% in voice, physiological, and hybrid systems (Imani & Montazer, 2019) and 98% in wearable applications (Schmidt et al., 2019). A survey of more than 380 studies published between 2000 and 2020 found no mention of regression models (Y. Wang et al., 2022).

Common biometric features used in $\mathbf{m}$ include facial expressions, voice, gaze movements, and physiological signals from the Peripheral Nervous System (PNS) or the Central Nervous System (CNS) (Calvo & D'Mello, 2010). Signals from the PNS include the electromyogram (EMG), skin conductance (SC), electrocardiogram (ECG), photoplethysmogram (PPG), and electrooculogram (EOG). Signals from the CNS include electroencephalogram (EEG) and, more recently, functional magnetic resonance imaging (fMRI) and functional near-infrared spectroscopy (fNIRS) (Bota et al., 2019; Shu et al., 2018).

The features are typically *multimodal* – that is, derived from at least two modalities – rather than *unimodal*. Their relative prevalence varies from 85% (D'Mello & Kory, 2015) to 93% (Schmidt et al., 2019). Among these, voice and facial expressions are the most common, with reported individual percentages ranging from 82% and 77% (D'Mello & Kory, 2015) to 21% and 54% (Y. Wang et al., 2022), respectively. Physiological signals are used in approximately 16% of affect detection systems overall (Y. Wang et al., 2022), with at least 10% incorporating PNS signals and 5% incorporating CNS signals (D'Mello & Kory, 2015). Wearable applications present a distinct scenario: all employ PNS signals (100%), while only 15% include CNS signals, 2% include facial expressions, and none incorporate voice data (Schmidt et al., 2019).

Scalar features extracted from each modality can be integrated – a process commonly referred to as *fusion* – at different stages of the training process. In the *feature-level* or *early fusion* approach (Verma & Tiwary, 2014), modality-dependent features $m_i^k$ are concatenated into a single feature vector $\mathbf{m} = (m_1^1, m_2^1, \ldots, m_M^N)^T$, regardless of modality $k \leq N$. A single classifier $f$ is then trained to predict the affective state as $s = f(\mathbf{m})$. In the *decision-level* or *late* *fusion* approach (Verma & Tiwary, 2014), scalar features are first grouped into modality-specific feature vectors $\mathbf{m}_k = (m_1^k, m_2^k, \ldots, m_M^k)^T$, which are used to train modality-dependent classifiers $g_k$ such that $o_k = g_k(\mathbf{m}_k)$. The final affect detection is performed by a classifier $f$, which takes as input the concatenated outputs of these classifiers $\mathbf{o} = (o_1, o_2, \ldots, o_N)^T$, yielding $s = f(\mathbf{o})$. The *model-level* or *intermediate fusion* approach takes into account interdependencies (e.g., correlations) among modality-dependent features in both feature representation and classification stages (D'Mello & Kory, 2015). Specifically, model components $\boldsymbol{c}_k$, fed with modality-specific feature vectors $\boldsymbol{m}_k$, serve to create modality-specific intermediate (or "hidden") representations $\boldsymbol{h}_k = \boldsymbol{c}_k(\boldsymbol{m}_k)$. Then, a fusion model $\boldsymbol{e}$ projects them into a common *embedding space* as $\boldsymbol{y} = \boldsymbol{e}\left(\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_M\right)^T$, which serves as input to the final classifier as $s = f\left(\boldsymbol{y}\right)$ (Guarrasi et al., 2025). Finally, the *hybrid fusion* approach combines elements of both feature- and decision-level methods. In this case, the classifier $f$ is fed with a combined vector $\mathbf{n}$ consisting of both concatenated feature vectors $\mathbf{m}$ and classifier outputs $\mathbf{o}$, expressed as $\mathbf{n} = (\mathbf{m}^T, \mathbf{o}^T)^T$ (D'Mello & Kory, 2015). The recent advent of Deep Learning (DL) architectures, where feature extraction, data integration, and decision-making are jointly learned within the model, has blurred the boundaries of this conventional taxonomy (F. Zhao et al., 2024). Figure 3.2 summarises the main multimodal fusion strategies.

Reported usage rates for feature-, decision-, model-, and hybrid-level fusions vary, respectively, between 39%, 36%, 20%, and 5% (D'Mello & Kory, 2015) and 44%, 28%, 24%, and 4% (Y. Wang et al., 2022).

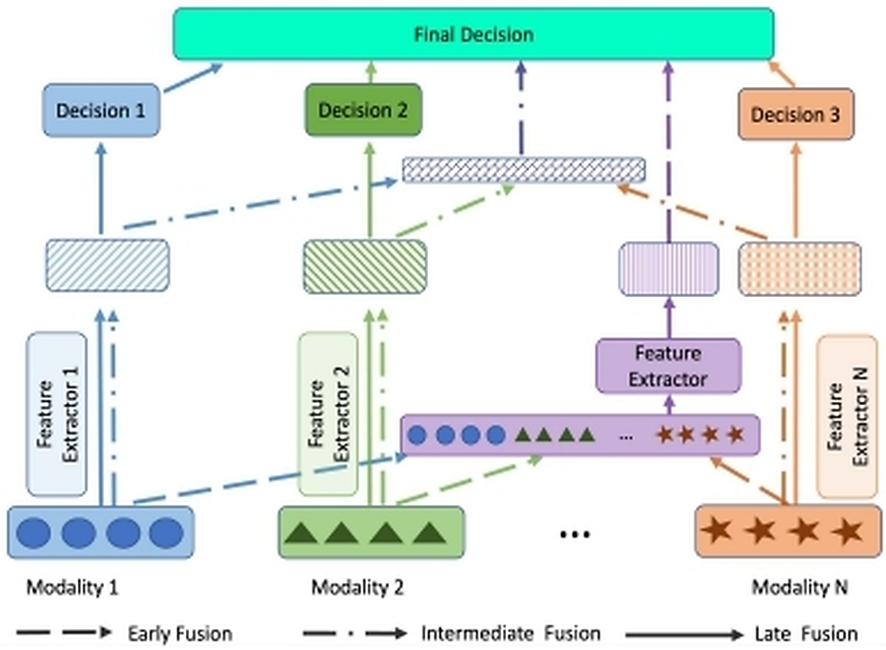


Figure 3.2: Schematic representation of early (or feature-level), intermediate (or model-level), and late (or decision-level) fusion approaches – from F. Zhao et al. (2024).

# Chapter 4

# Machine Learning Foundations

This chapter offers a concise overview of supervised Machine Learning, with specific emphasis on classification. It first introduces Machine Learning as a form of inductive inference and presents the statistical learning framework, including Bayes optimality, Empirical Risk Minimisation (ERM), and Structural Risk Minimisation (SRM). It then surveys the main classifier families, including informative, discriminative, non-metric, and ensemble classifiers. It presents in detail Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), $k$-nearest neighbours (kNN), Naive Bayes (NB), Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Tree (DT), and ensemble methods. Finally, it addresses the problem of model selection and evaluation, focusing on resampling methods and theoretical bounds for error estimation.

## 4.1 Defining Machine Learning

The term *Machine Learning* (ML) dates back to the late 1950s in relation to programs designed to play games (tic-tac-toe and chess, respectively) without relying on a fully predefined set of rules (Martens, 1959; Samuel, 1959). These early studies anticipated several key elements of modern ML, including feedback-driven adaptation, parametrised decision rules, and performance evaluation under uncertainty.

In modern terms, ML can be defined as "a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty" (Murphy, 2012, p. XXVII). The word "pattern" refers to "regularities in data" (Bishop, 2006, p. 1), which justifies the synonym *Pattern Recognition*, common in engineering contexts (Bishop, 2006, p. vii).

Machine Learning can be viewed as an application of logical *inference* – the process of deriving conclusions from assumed truths. Specifically, it belongs to *inductive* inference, where the facts are *observations* (i.e., specific cases) and the conclusions are *general rules*. In inductive inference, the observations offer probabilistic support for the conclusions – raising their likelihood of truth – rather than guaranteeing them, as in the deductive counterpart (Boghossian, 2012).

Machine Learning is commonly subdivided into *unsupervised* and *supervised* branches. Unsupervised learning is *exploratory* (or *descriptive*), seeking "interesting" structures in data to group (*clustering*) or transform (*dimensionality reduction*) patterns without external information. Supervised learning, by contrast, requires patterns (numerically coded as *feature vectors*) paired with target variables (or *labels*) and seeks a mapping rule between them. Its main tasks include *regression* (continuous targets) and *classification* (discrete targets) (Murphy, 2012, pp. 1–16).

In the following, we focus on supervised ML and, specifically, on classification.

## 4.2 Statistical Learning Framework

The Statistical Learning framework (Vapnik, 1999) provides a general formulation of the classification problem. Let $\boldsymbol{x} \in \mathcal{X}$ denote a feature vector and $y \in \mathcal{Y}$ its class label, both generated by a *fixed* joint probability distribution $p(\boldsymbol{x}, y) = p(y|\boldsymbol{x})p(\boldsymbol{x})$. Independent and identically distributed observations $\mathcal{D} = \{(\boldsymbol{x}_k, y_k)\}_{k=1}^{N} \in \mathcal{Z}^N \doteq (\mathcal{X} \times \mathcal{Y})^N$, drawn from $p(\boldsymbol{x}, y)$, form the *training set*.

Despite being fixed, both $p(\boldsymbol{x}, y)$ and the factors $p(\boldsymbol{x}|y)$ and $p(\boldsymbol{x})$ are almost always *unknown*. The feature–label mapping $p(y|\boldsymbol{x})$ is generally *unknown* and is approximated by means of a *classifier* $f \in \mathcal{F},\ f : \mathcal{X} \to \mathcal{Y}$. Often, the function class $\mathcal{F}$ consists of a set of functions $f$ indexed by a set of *parameters* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, namely $\mathcal{F} \equiv \{f_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$.

Classifier performance is evaluated through a *loss function* $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, so that $L(y, f(\boldsymbol{x}))$ measures the penalty incurred by predicting $f(\boldsymbol{x})$ when the true class is $y$.

The *expected risk* of a generic classifier $f$ is defined as:

$$R(f) = \mathbb{E}_{\boldsymbol{x},y}\left\{L\left(y, f(\boldsymbol{x})\right)\right\} \tag{4.1}$$

Let $\mathfrak{F} = \{f | f : \mathcal{X} \to \mathcal{Y}\}$ be the set of all functions from $\mathcal{X}$ to $\mathcal{Y}$. The *Bayes classifier*, $f_B$, and the corresponding *Bayes risk*, $R_B$, are defined, respectively, as:

$$
\begin{aligned}
f_B &= \arg\inf_{f \in \mathfrak{F}} R(f) \\
R_B &= R(f_B)
\end{aligned}
\tag{4.2}
$$

The *approximated Bayes classifier* over $\mathcal{F} \subset \mathfrak{F}$, $f_{\mathcal{F}}$, and the corresponding risk, $R_{\mathcal{F}}$, are defined, respectively, as:

$$
\begin{aligned}
f_{\mathcal{F}} &= \arg\inf_{f \in \mathcal{F}} R\,(f) \\
R_{\mathcal{F}} &= R(f_{\mathcal{F}}),
\end{aligned}
\tag{4.3}
$$

Both $f_B$ and $f_{\mathcal{F}}$ represent *optimal* classifiers. However, their computation requires minimising the expected risk $R(\cdot)$, which is practically impossible because it is based on the expectation $\mathbb{E}_{\boldsymbol{x},y}\{\cdot\}$ and therefore requires knowledge of $p(\boldsymbol{x}, y)$. Empirical Risk Minimisation (ERM) addresses this problem by means of the *induction principle*: minimising the risk over the training set $\mathcal{D}$ can lead to a classifier with low expected risk.

### Empirical Risk Minimisation

Given a training set $\mathcal{D} = \{(\boldsymbol{x}_k, y_k)\}_{k=1}^{N}$, the *empirical risk* of a generic classifier $f$, $R_{emp}$, and its infimum, $f'$, are defined as:

$$
\begin{aligned}
R_{emp}(f) =& \frac{1}{N}\sum_{i=1}^{N} L\,(y_i, f(\boldsymbol{x}_i)) \\
f' =& \arg\inf_{f \in \mathcal{F}} R_{emp}(f)
\end{aligned}
\tag{4.4}
$$

Because $f'$ depends on $\mathcal{D}$ through $R_{emp}$ and its minimisation, it can be considered as generated by a *learning algorithm* $\mathcal{A}_{\boldsymbol{\gamma}} : \mathcal{Z}^N \to \mathcal{F}$, so that $f' = \mathcal{A}_{\boldsymbol{\gamma}}(\mathcal{D})$, where $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$ are the *hyperparameters*. Given a fixed $\mathcal{D}$, different algorithm–hyperparameter configurations generate different function classes (Oneto, 2020, pp. 17–20).

The Empirical Risk Minimisation (ERM) induction principle (Vapnik, 1991; Vapnik, 1999) approaches $R_{\mathcal{F}}$ through $f'$ by assuming that the expected risk of $f'$ is close to the infimum,

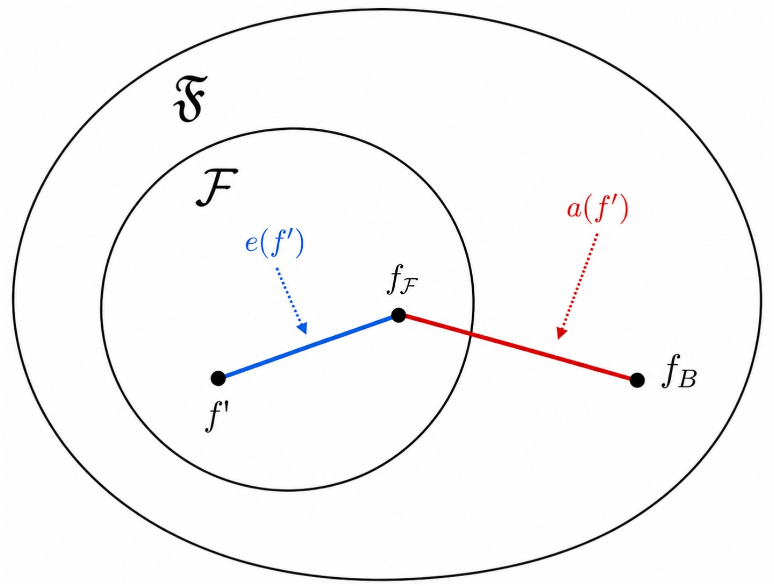


Figure 4.1: Estimation $e(f')$ and approximation $a(f')$ errors of the classifier $f' \in \mathcal{F}$ with respect to the approximated Bayes classifier in $\mathcal{F}$, $f_{\mathcal{F}}$, and the Bayes classifier, $f_B$.

namely:

$$R(f') \approx R(f_{\mathcal{F}}) \equiv \inf_{f \in \mathcal{F}} R(f) \tag{4.5}$$

The distance (in risk terms) between $f'$ and the Bayes classifier $f_B$ is measured by the so-called *excess risk* $R(f') - R_B$. The following decomposition, also involving the approximated Bayes classifier $f_{\mathcal{F}}$, can be given (Brown & Ali, 2024):

$$R(f') - R_B = \underbrace{[R(f') - R_{\mathcal{F}}]}_{e(f')} + \underbrace{[R_{\mathcal{F}} - R_B]}_{a(f')} \tag{4.6}$$

The term $e(f')$, the *estimation error*, reflects the uncertainty introduced by the random sampling that generates $\mathcal{D}$ and, in turn, $f'$. The term $a(f')$, the *approximation error*, arises from the limited size of $\mathcal{F}$ relative to $\mathfrak{F}$. There is a natural *trade-off* between $a(f')$ and $e(f')$: as we increase $\mathcal{F}$, $a(f')$ will likely decrease (dropping potentially to zero when $f_B \in \mathcal{F}$), while $e(f')$ will increase (as it becomes harder to find $f'$ in such a larger space) (Von Luxburg & Schölkopf, 2011, pp. 662–663). This trade-off is related to, but should not be confused with, the classical bias–variance decomposition, which concerns the expected behaviour of learning algorithms over repeated samples (Brown & Ali, 2024).

Equation 4.5 can be quantitatively expressed in terms of the *capacity* (or *complexity*) of the function class $\mathcal{F}$. For binary classifiers, this can be quantified by the Vapnik–Chervonenkis (VC) dimension $h$, defined as the maximum number of points that can be correctly separated for every arbitrary class assignment (i.e., "shattered" by $\mathcal{F}$). Complexity can be generalised to multiclass settings through measures such as the Natarajan dimension (Natarajan, 1989).

Given a function class $\mathcal{F}$ with VC dimension $h$, with probability $1-\epsilon$, and for all $f \in \mathcal{F}$, the following bound holds:

$$\begin{aligned} R(f) &\leq R_{emp}(f) + C_\epsilon\left(h, N\right), \\ C_\epsilon\left(h, N\right) &= \sqrt{\frac{h \ln\left(2N/h\right) + 1 - \ln\left(\epsilon\right)}{N}} \end{aligned} \tag{4.7}$$

where $C_\epsilon(\cdot,\cdot)$ is the *confidence term* and $N$ is the size of the training set.

When $N \gg h$, $C_\epsilon\left(h, N\right) \approx 0$, and Equation 4.7 implies $R(f) \approx R_{emp}(f)$ with high probability. By contrast, when $h \gg N$, one may have $R_{emp}(f) \approx 0$ while $R(f) \neq 0$, since $C_\epsilon\left(h, N\right) \neq 0$. In the latter case, ERM is ineffective, and either $R(f)$ or $R_{emp}(f) + C_\epsilon\left(h, N\right)$ must be directly minimised. This motivates the Structural Risk Minimisation (SRM) framework described next.

### Structural Risk Minimisation

The Structural Risk Minimisation (SRM) framework (Guyon et al., 1991; Vapnik, 2000) is based on simultaneously minimising the empirical risk and the confidence term. Because the latter depends on the VC dimension, SRM requires searching over different function classes, not just over different parameter values.

SRM introduces a nested structure $\{\mathcal{F}_i\}_{i=1}^N$ with associated VC dimensions $\{h_i\}_{i=1}^N$, such that $\mathcal{F}_i \subset \mathcal{F}_{i+1}$ and $h_i < h_{i+1}$. For each $\mathcal{F}_i$, it finds the classifier $f_i'$ that is optimal in the ERM sense. It then finds the optimal classifier $f^*$ that minimises the sum of the minimum empirical risk and the confidence term. In other words, SRM selects the classifier $f^*$ that minimises the upper bound of $R$ given by $R_{emp}$ plus $C_\epsilon$:

$$f^* = \arg \inf_{\substack{i \leq N \\ f \in \mathcal{F}_i}} \left\{R_{emp}\left(f\right) + C_\epsilon\left(h_i, N\right)\right\} \tag{4.8}$$

This involves a trade-off: for fixed $N$, as $i$ increases, $h_i$ increases and $R_{emp}(f_i')$ decreases, while $C_\epsilon\left(h_i, N\right)$ increases. As $C_\epsilon\left(h_i, N\right)$ is not always easily estimable, an alternative upper bound based on risk over an unseen "test set" is commonly used, and their minima have been shown to be close.

There are various ways to construct effective families $\{\mathcal{F}_i\}_{i=1}^N$. For some classifiers, such as linear models (Burges, 1998) and neural networks (Bartlett et al., 2019), the VC dimension relates to the number of parameters. One may therefore pre-process data by selecting $N < M$ features and training on them, or apply *pruning* after training by setting $N - M$ weights to zero. Additionally, in *regularisation* (Cherkassky & Ma, 2009), the empirical risk in ERM is replaced by:

$$R_{reg}(f, \lambda) = R_{emp}(f) + \lambda\Omega(f) \tag{4.9}$$

where $\lambda > 0$ is a regularisation parameter and $\Omega(\cdot)$ is a measure of the complexity of $f$, usually a norm $\|\cdot\|$. Finally, SRM can also be implicitly approached during training algorithms such as stochastic gradient descent (SGD), where, as iterations progress, functions of increasing complexity are learned (Nakkiran et al., 2019).

## 4.3 Classifier Families

### 4.3.1 Informative Classifiers

Under the hypotheses of a finite set of classes $\mathcal{Y} = \{y_k\}_k$ and an action space coinciding with $\mathcal{Y}$, with respect to 0–1 loss, the Bayes classifier coincides with the following pointwise maximum-a-posteriori (MAP) classifier $f_{MAP}$ (Bassett & Deride, 2019):

$$f_{MAP}(\boldsymbol{x}) = \arg\max_{y\in\mathcal{Y}} \{p(y|\boldsymbol{x})\} \equiv \arg\max_{y\in\mathcal{Y}} \{p(\boldsymbol{x}|y)p(y)\} \tag{4.10}$$

This justifies the so-called *informative* (Rubinstein, Hastie, et al., 1997) or *generative* (Drummond, 2006; Ng & Jordan, 2001) classifiers, which estimate either the posterior probability $p(y|\boldsymbol{x})$ directly or the class-conditional density $p(\boldsymbol{x}|y)$ together with the prior $p(y)$ (Rubinstein, Hastie, et al., 1997). Once estimated, the probabilities are inserted into Equation 4.10, so that these classifiers are also referred to as *plug-in* classifiers (Holmström & Koistinen, 2010; Jain et al., 2000).

#### Parametric Informative Classifiers

*Parametric* informative classifiers assume a specific form for $p(\boldsymbol{x}|y_k)$ and estimate its parameters from the training data (Bishop, 2006, p. 68). In Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), the class-conditional densities $p(\boldsymbol{x}|y_k)$ are assumed to be multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ (Tharwat, 2016).

Equation 4.10 can be equivalently solved by replacing the class-conditional probabilities $p(\boldsymbol{x}|y_k)$ with the following quadratic functions[1], obtaining QDA:

$$g(\boldsymbol{x}, y_k) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| + \ln p(y_k). \tag{4.11}$$

Assuming, for all $y_k \in \mathcal{Y}$, $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$, Equation 4.11 simplifies to the following linear function, corresponding to LDA:

$$g(\boldsymbol{x}, y_k) = \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln p(y_k). \tag{4.12}$$

The mean vectors and covariance matrices are usually estimated following the Maximum Likelihood approach, while prior probabilities $p(y_k)$ are usually replaced by the relative frequencies $n_k/N$ (Tharwat, 2016). Maximum Likelihood estimation is equivalent to ERM, considering, as a loss function, the negative log-likelihood of $\mathcal{N}_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x})$, as (Vapnik, 2000, p. 24):

$$\arg\min_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k} \frac{1}{N_k}\sum_{i=1}^{N_k} -\ln\left(\mathcal{N}_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}\left(\boldsymbol{x_i}^k\right)\right), \tag{4.13}$$

where $\boldsymbol{x}_i^k$ is the $i$-th point (over $N_k$) in $\mathcal{D}$ belonging to class $y_k$.

When covariance estimates are singular or ill-conditioned, as may happen in high-dimensional or small-sample settings, the inverse covariance matrix can be replaced by a pseudo-inverse $\boldsymbol{\Sigma}^\dagger$ or by a regularised estimate $\boldsymbol{\Sigma}_\lambda = \boldsymbol{\Sigma} + \lambda\boldsymbol{I},\ \lambda > 0$, yielding pseudo-inverse or regularised versions of LDA and QDA, respectively (Raudys & Duin, 1998).

### Non-parametric Informative Classifiers

*Non-parametric* informative classifiers avoid imposing a fixed form on the class-conditional probabilities and directly estimate either $p(\boldsymbol{x}|y_k)$ or $p(y_k|\boldsymbol{x})$. When needed, the probabilities $p(y_k)$ are simply estimated by the relative frequencies.

[1] $g(\boldsymbol{x}, y_k)$ can be obtained, under Gaussian assumptions, by applying the natural logarithm to $p(\boldsymbol{x}|y_k)$ and ignoring the class-independent terms, as shown by Tharwat (2016).

The *Parzen-window* (PW) classifier estimates $p(\boldsymbol{x}|y_k)$ as (Jain & Ramaswami, 1988):

$$p(\boldsymbol{x}|y_k) \approx \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{1}{h^M} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i^k}{h}\right), \tag{4.14}$$

where $K(\cdot)$ is a kernel function, $h > 0$ is the smoothing parameter, $M$ is the feature dimensionality, and $\{\boldsymbol{x}_i^k\}_{i=1}^{N_k}$ are the training samples of class $y_k$.

The $k$-nearest-neighbour (kNN) classifier instead estimates the posterior probability of class $y_k$ in the neighbourhood of $\boldsymbol{x}$ as (Jain & Ramaswami, 1988):

$$p(y_k|\boldsymbol{x}) \approx \frac{1}{K} \sum_{i=1}^{K} I(y_{k_i}, y_k), \tag{4.15}$$

where $y_{k_i}$ are the labels of the $K$ nearest points from $\boldsymbol{x}$ and $I(\cdot, \cdot)$ is the indicator function. The neighbourhood is defined based on a distance metric. Common choices include, among others, Euclidean, Manhattan, and Chebyshev distances (Abu Alfeilat et al., 2019).

Finally, the Naive Bayes (NB) classifier assumes conditional independence among features and factorises the joint class-conditional probabilities as:

$$p(\boldsymbol{x}|y_k) \approx \prod_{i=1}^{M} p(x_i|y_k) \tag{4.16}$$

Rather than being a novel classifier, NB represents a simplified approach to composing marginal probabilities $p(x_i|y_k)$, which still have to be estimated using either parametric (e.g., Gaussian assumption) or non-parametric (e.g., Parzen-window) approaches (Hand & Yu, 2001).

### 4.3.2 Discriminative Classifiers

Discriminative classifiers forgo the MAP approach in pursuit of "directly" learning a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ (Drummond, 2006; Ng & Jordan, 2001; Rubinstein, Hastie, et al., 1997) that, according to ERM, can still approach the Bayes risk.

The discriminative framework expresses $f$ through discriminant functions $g : \mathcal{X} \rightarrow \mathbb{R}$ representing the evidence, score, or compatibility of $\boldsymbol{x}$ with the classes. More specifically, a set of classes $\{y_k\}_{k=1}^{K>2}$ is associated with a set of $K$ discriminant functions $\{g_k\}_{k=1}^{K}$ so that (Duda et al., 2000, p. 29):

$$\begin{aligned} &f(\boldsymbol{x}) = y_k \\ &k = \arg\max_{i \le K} \{g_i(\boldsymbol{x})\} \end{aligned} \tag{4.17}$$

In the binary case $\mathcal{Y} = \{y_1, y_2\}$, only one discriminant function is usually needed, namely $g(\boldsymbol{x}) \doteq g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$. The resulting classifier, also called a *dichotomiser*, simply performs a check on the value or sign of $g(\boldsymbol{x})$.

**Support Vector Machine**

The *Linear Support Vector Machine* (SVM) (Burges, 1998; Cortes & Vapnik, 1995) is an affine discriminant function $g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}+b$ which provides binary classifications as $f(\boldsymbol{x}) = \text{sign}\,\{g(\boldsymbol{x})\}$. It applies to linearly separable datasets $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, namely where $g(\boldsymbol{x}_i) \geq 1, \forall i \leq N$. Geometrically, it maximises the *margin* $2/\|\boldsymbol{w}\|$ based only on a limited number of points $\{\boldsymbol{x}_k\}_{k<N} \subset \mathcal{D}$, called *support vectors*, for which the relationship $y_k g(\boldsymbol{x}_k) = 1$ holds. It solves the following optimisation problem:

$$\begin{aligned} &\arg\min_{\boldsymbol{w},b} \left\{ \frac{1}{2}\|\boldsymbol{w}\|^2 \right\} \\ &\text{s.t.} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 \end{aligned} \tag{4.18}$$

With non-separable datasets, Linear SVM cannot be applied. The so-called *soft-margin* SVM (Burges, 1998; Cortes & Vapnik, 1995) considers non-negative slack variables $\xi_i$ and a *penalty* coefficient $C$ and rewrites the margin maximisation as [2]:

$$\begin{aligned} &\arg\min_{\boldsymbol{w},b} \left\{ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^N \xi_i \right\} \\ &\text{s.t.} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \end{aligned} \tag{4.19}$$

By introducing the hinge loss as $L_{Hinge}\,(g(\boldsymbol{x}_i), y_i) = \max\{1 - y_i g(\boldsymbol{x}_i), 0\}$, the SVM optimisation problem can be reframed as SRM (Karaçalı et al., 2004) and, specifically, as a regularised

[2] Obviously, with $\xi_i = 0,\ \forall i \leq N$, the soft-margin SVM is equivalent to the Linear SVM.

ERM (Cherkassky & Ma, 2009) with the following form (Evgeniou et al., 2000):

$$g^* = \arg\min_g \left\{ \underbrace{\frac{1}{N}\sum_{i=1}^{N} L_{Hinge}\left(g(\boldsymbol{x}_i), y_i\right)}_{R_{emp}(g)} + \underbrace{\frac{1}{2C}}_{\lambda} \underbrace{\|g\|^2}_{\Omega(g)} \right\} \tag{4.20}$$

Although an SVM is inherently linear, it can achieve non-linear behaviour if applied to transformed vectors $\phi(\boldsymbol{x}_i)$, where $\phi : \mathcal{X} \rightarrow \mathcal{X}'$ is a non-linear mapping to a higher-dimensional space $\mathcal{X}'$. This is the basis of the *kernel* SVM and kernel methods in general (Muller et al., 2001).

In both training[3] and prediction phases, SVMs rely solely on scalar products between vectors $\boldsymbol{x}_i^T \boldsymbol{x}_j$, which in the target space take the form $\phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$. Direct computation can be infeasible in very high dimensions. However, appropriate mappings can be implicitly defined by replacing the dot product in $\mathcal{X}'$ with a kernel function $\kappa : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ such that $\phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Thanks to $\kappa$, scalar products in $\mathcal{X}'$ can be computed implicitly, without explicitly using or even knowing $\phi$: this is the *kernel trick* (Muller et al., 2001). Typical kernels include the *Gaussian* (or radial basis function, RBF) $\kappa(\boldsymbol{x}, \boldsymbol{y}) = e^{-\|\boldsymbol{x}-\boldsymbol{y}\|^2}$ and the *polynomial* $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \left(1 + \boldsymbol{x}^T \boldsymbol{y}\right)^{p>1}$, as well as the dummy *linear* $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \left(\boldsymbol{x}^T \boldsymbol{y}\right)$ associated with the identity mapping $\phi(\boldsymbol{x}) = \boldsymbol{x}$.

SVMs are natively binary and, to achieve multiclass ($K > 1$) behaviour, several SVMs must be combined. In the one-vs-rest (OVR) approach, $K$ binary classifiers $\{f_k\}_{k=1}^K$ are trained such that $f_i$ discriminates class $y_i$ versus $\bar{y}_i$, the aggregate of the remaining classes $\{y_j\}_{j \neq i}$. In the one-vs-one (OVO) approach, $K(K-1)/2$ classifiers discriminate each pair $y_i$ versus $y_j$, with $i \neq j$ (Platt et al., 1999). Finally, in the Error-Correcting Output Code (ECOC) framework, $M$ classifiers represent a binary coding of the classes; the simplest is base-2 integer coding using $2^{K-1} - 1$ classifiers, though additional methods (e.g., BCD) can be applied (Dietterich & Bakiri, 1995).

### Artificial Neural Networks

Artificial Neural Networks (ANNs), also known as multilayer perceptrons, are discriminative classifiers composed of artificial neurons (or perceptrons) arranged in layers (Higham & Higham, 2019). Each neuron applies a non-linear activation function $\sigma(\cdot)$ to an affine transformation of

[3] This is particularly evident in the Wolfe dual problem; see Burges (1998, p. 135).

its inputs as $f(\boldsymbol{x}) = \sigma\left(\boldsymbol{w}^T\boldsymbol{x} + b\right)$, where $\boldsymbol{w}$ is the weight vector, $b$ is the bias term, and $\sigma$ is the activation function. By stacking neurons into *fully connected layers*, a *feed-forward* ANN defines a composition of non-linear transformations. For the $i$-th layer,

$$\boldsymbol{h}^{(i)} = \boldsymbol{\sigma}^{(i)}\left(\boldsymbol{W}^{(i)}\boldsymbol{h}^{(i-1)} + \boldsymbol{b}^{(i)}\right) \tag{4.21}$$

with $\boldsymbol{h}^{(0)} = \boldsymbol{x}$. The matrices $\boldsymbol{W}^{(i)}$ and vectors $\boldsymbol{b}^{(i)}$ contain the trainable parameters, while the number of layers, the number of neurons per layer, and the activation functions are considered hyperparameters.

The first *input* layer usually matches the dimensionality of the features, while the last *output* layer, which provides class-probability estimates, is chosen according to the target structure (Bishop, 2006, pp. 232–236). Binary classification usually employs a single output neuron with sigmoid activation:

$$\sigma(h) = \frac{1}{1 + \exp(-h)} \tag{4.22}$$

Multiclass ($K > 1$) classification commonly uses an output layer of $K$ neurons with the softmax output, namely:

$$\sigma_i(\boldsymbol{h}) = \frac{\exp(h_i)}{\sum_{j=1}^{K}\exp(h_j)} \quad i \le K \tag{4.23}$$

ANN training is based on ERM, with cross-entropy, negative likelihood, and mean-square error as popular loss functions (Terven et al., 2025). Minimisation is achieved through the following gradient descent (GD) iteration (Deisenroth et al., 2020, pp. 227–229):

$$\boldsymbol{p}^{(k+1)} = \boldsymbol{p}^{(k)} - \eta \nabla R_{emp}\left(\boldsymbol{p}^{(k)}\right), \tag{4.24}$$

where $\boldsymbol{p}$ is the vector collecting the trainable parameters and $\eta > 0$ is the learning rate. Instead of computing the sum over all training points, mini-batch GD and stochastic GD approximate the gradient using, respectively, subsets of the training set or single observations (Tian et al., 2023).

The gradients are efficiently computed by applying the chain rule backwards through the network according to the *backpropagation* algorithm (Higham & Higham, 2019).

### 4.3.3 Non-metric Classifiers

Non-metric classifiers do not require an explicit distance measure between observations. A prominent example is the Decision Tree (DT), which classifies samples through a sequence of conditional rules organised from a root node to terminal leaves, where class decisions are assigned (Duda et al., 2000, pp. 394–396).

Training a DT usually consists of two phases: growing and pruning (Hernández et al., 2021). During *growing*, the training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ is recursively partitioned according to a splitting criterion. At each node, candidate splits are evaluated through a quality measure, and the best split is selected. The procedure is repeated until a stopping condition is reached. The final node is then labelled according to a predefined rule, usually the majority class. *Pruning* subsequently reduces the complexity of the tree by removing branches that do not improve generalisation.

Splits may be univariate, when they involve a single feature, or multivariate, when they combine multiple features (Brodley & Utgoff, 1995). They may also be binary or $N$-ary, producing two or more ($N > 2$) child nodes, respectively (Hernández et al., 2021).

Quality measures, usually interpreted in terms of *impurity*, are based on the class proportions $\{\mathcal{P}(y_i)\}_{i=1}^{K}$. The most common impurity measures include the Gini index $I_G$ and deviance (or entropy) $I_D$, defined as (Aluja-Banet & Nafria, 1998):

$$
\begin{aligned}
I_G(\mathcal{D}) &= \sum_{j \neq k} \mathcal{P}(y_j)\mathcal{P}(y_k) \\
I_D(\mathcal{D}) &= -\sum_{j=1}^{K} \mathcal{P}(y_j) \log\left(\mathcal{P}(y_j)\right)
\end{aligned}
\tag{4.25}
$$

A split is then evaluated by the impurity reduction from the parent to the resulting child nodes. Given a split $s(\cdot)$ that partitions $\mathcal{D}$ into subsets $\{\mathcal{D}_m\}_{m=1}^{M}$, this reduction can be written as (Duda et al., 2000, p. 401):

$$
\Delta_I(s) = I(\mathcal{D}) - \sum_{m=1}^{M} \frac{N_m}{N} I(\mathcal{D}_m), \tag{4.26}
$$

where $I(\cdot)$ is a quality index and $N_m$ is the number of observations in partition $\mathcal{D}_m$. Alternative criteria, such as the gain ratio or the twoing index, normalise or reformulate this impurity reduction to account for the structure of the split.

Common stopping criteria include a maximum depth, a maximum number of nodes or splits, a minimum number of samples per leaf, or a minimum impurity improvement. When growing stops, the final node is labelled by the most frequent class.

Unlike previous classifiers, optimising a DT is NP-hard, and the standard top-down greedy algorithm is only locally optimal (Bertsimas & Dunn, 2017). However, an ERM-like bound (with probability $1 - \delta$) can be given as (Mansour & McAllester, 2000):

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{\ln(2)|T|\ln(1/\delta)}{2N}}, \tag{4.27}$$

where $|T|$ is the encoded bit length of the DT, which plays the role of a complexity measure.

### 4.3.4 Ensemble Classifiers

Ensemble classifiers combine multiple base learners to obtain a final prediction. Let $\{f_k\}_{k=1}^{M}$ be a set of base classifiers. An ensemble classifier can be written as $\phi(\boldsymbol{x}) = \gamma\left(f_1(\boldsymbol{x}), \ldots, f_M(\boldsymbol{x})\right)$, where $\gamma$ is a proper aggregating function (Sagi & Rokach, 2018).

The intuition behind ensemble classifiers is related to the Condorcet jury principle. If $M$ *independent* classifiers each have probability $p$ of making a correct binary decision, the probability of a correct majority vote is given by (Lam & Suen, 1997):

$$P(M) = \sum_{i=k}^{M} \binom{M}{i} p^i (1-p)^{M-i}, \tag{4.28}$$

where $k = \lfloor M/2 \rfloor + 1$, with $\lfloor \cdot \rfloor$ being the floor function. If $p > 0.5$ and the independence assumption holds, then $P(M) \to 1$ as $M \to \infty$.

Ensemble classifiers require diversity (low correlation) among weak learners, which is typically achieved by training on differently sampled inputs (input manipulation), altering learning settings or hyperparameters (algorithm manipulation), training on different feature- or instance-wise partitions (partitioning), or transforming outputs via different coding schemes (output manipulation). Hybridising these techniques is also common (Sagi & Rokach, 2018).

Aggregation functions are typically *weighting* methods and *meta-learning* (Rokach, 2009). Weighting implements a linear combination $\phi(\boldsymbol{x}) = \sum_k w_k f_k(\boldsymbol{x})$, $w_k \in \mathbb{R}$, while meta-learners train a classifier on the weak predictions $\boldsymbol{\xi} = (f_1(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x}))^T$ and output $\phi(\boldsymbol{\xi})$.

Major ensemble families include bagging, boosting, stacking, and random subspace methods, alongside their combinations and variants (K. Sharma et al., 2023). In *bagging*, homogeneous classifiers $\{f_k\}_k$ are trained in parallel on bootstrapped versions of the original dataset, namely

$\mathcal{D}_k = \mathcal{B}_k(\mathcal{D})$, and their predictions are then combined, typically by majority vote or weighting. In *boosting*, homogeneous classifiers are trained sequentially on the dataset $\mathcal{D} = \{(\boldsymbol{x}_k, y_k)\}_{k=1}^N$, assigning at iteration $k$ a cost $w_i^k L(f_k(\boldsymbol{x}_i), y_i)$ to each observation. The weights are updated according to the performance of the current classifier, for instance as $w_i^{k+1} = w_i^k + \delta$, with $\delta > 1$ when $\boldsymbol{x}_i$ is misclassified by $f_k$ and $\delta < 1$ otherwise, starting from $w_i^1 = 1/N$. The final prediction is obtained through a weighted aggregation of the individual classifiers. In *stacking*, heterogeneous classifiers $\{f_k\}_k$ are trained in parallel, either on the original dataset $\mathcal{D}$ or on resampled datasets $\mathcal{D}_k = \mathcal{B}_k(\mathcal{D})$; their predictions are then collected into a vector $\boldsymbol{\xi} = (f_1(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x}))^T$ and used as inputs to a meta-learner. Finally, in *random subspace* methods, homogeneous classifiers are trained in parallel on feature-wise partitions $\mathcal{D}_k = \mathcal{S}_k(\mathcal{D})$. Given $\boldsymbol{x} = (x_1, \ldots, x_M)^T \in \mathcal{D}$, each subspace representation $\boldsymbol{x}^k \in \mathcal{D}_k$ contains only $N < M$ selected features, $\boldsymbol{x}^k = (x_1^k, \ldots, x_N^k)^T$, with $x_i^k \in \{x_j\}_{j=1}^M$; the resulting classifiers are then aggregated by voting or weighting.

A popular ensemble method is the Random Forest (RF), originally proposed as a random subspace method of oblique DTs (Ho, 1995), later extended to bagging of DTs (Breiman, 2001).

Most ensemble classifiers are either Bayesian or asymptotically tend to achieve Bayes expected risk (Le & Clarke, 2018).

## 4.4 Error Estimation and Model Selection

Estimating the expected risk $R(\cdot)$ of a classifier $f$ is challenging. Even for informative classifiers – which are Bayes-optimal and for which an explicit $R(f)$ exists – the finite-sample setting biases both parameter and density estimates. In discriminative classifiers, where confidence intervals tighten with increasing sample size $N$, the upper bound on $R(f)$ remains uncertain because the VC dimension $h$ is not always computable. Similar considerations apply to non-metric classifiers, replacing $h$ with the encoded bit length $|T|$.

Given these limitations, $R(f)$ is empirically estimated on a separate set (Guyon et al., 1991). The simplest approach, *holdout*, splits $\mathcal{D}$ into a learning set $\mathcal{L}$ and a validation set $\mathcal{V}$ to train and evaluate $f$. In *resubstitution*, $\mathcal{V}$ coincides with $\mathcal{L}$. Holdout is pessimistically biased and depends on the particular split; resubstitution is optimistically biased, especially when $N$ is small[4] (Jain et al., 2000).

[4]Resubstitution assumes $R(f) \approx R_{\text{emp}}(f)$, ignoring the confidence interval in the ERM bound.

Resampling methods – Monte Carlo, bootstrap, and cross-validation – are based on multiple splits $\left\{\mathcal{L}^{(i)}, \mathcal{V}^{(i)}\right\}_{i=1}^{N}$ to compute an averaged error estimate. In Monte Carlo, $\mathcal{L}^{(i)}$ and $\mathcal{V}^{(i)}$ are random partitions of $\mathcal{D}$, typically 75% and 25%, respectively. In bootstrap, $\mathcal{L}^{(i)}$ is bootstrapped from $\mathcal{D}$ and $\mathcal{V}^{(i)} = \mathcal{D} \setminus \mathcal{L}^{(i)}$. $k$-fold cross-validation divides $\mathcal{D}$ into $k$ folds (commonly 5 or 10), iteratively using one as $\mathcal{V}^{(i)}$ and the remaining $k-1$ as $\mathcal{L}^{(i)}$. Variants include leave-one-out (LOO), where each instance serves once as $\mathcal{V}^{(i)}$, and repeated $(n \times k)$-fold cross-validation (Nakatsu, 2021).

The *no-free-lunch* theorem (Wolpert, 1996) states that it is impossible to identify a learning algorithm $A(\cdot)$ inducing a classifier $f$ that universally outperforms all others. Model selection addresses no-free-lunch by searching, among candidates, the classifier $f^*$ (or, more precisely, the inducing algorithm $A^*$) minimising expected error. Let $\mathcal{A}$ be the set of candidate learning algorithms (possibly differing by hyperparameters). Let $\mathcal{D} = \{(\boldsymbol{x}_k, y_k)\}_{k=1}^{N}$ be the training set with $(\boldsymbol{x}_k, y_k) \sim p(\boldsymbol{x}, y)$. Given a loss function $L(\cdot, \cdot)$, model selection aims to find (J. Ding et al., 2018):

$$A^* = \arg\min_{A \in \mathcal{A}} \left\{ \underbrace{\mathbb{E}_{\boldsymbol{x},y} \left\{ L(y, \underbrace{A(\mathcal{D})}_{f}(\boldsymbol{x})) \right\}}_{R(f)} \right\} \tag{4.29}$$

As noted, model selection is tightly linked to error estimation: for fixed $\mathcal{A}$, the sharper the estimate of the generalisation error $R(f)$, the higher the probability of selecting the best model. State-of-the-art model selection and error estimation methods belong to (or combine) six main approaches and provide estimates of the form $\mathcal{P}\{R(f) \geq \Delta\} \leq \delta(\Delta)$, where $\delta$ depends on the accuracy $\Delta$ and vice versa (Oneto, 2020).

In the following, we describe the resampling approach. Given a dataset $\mathcal{D} = \{(\boldsymbol{x}_k, y_k)\}_{k=1}^{N}$ and a set of $N_R$ splits, each with disjoint subsets[5] $\mathcal{D} = \mathcal{L}^{(i)} \cup \mathcal{V}^{(i)} \cup \mathcal{T}^{(i)}$, $\forall i = 1, \ldots, N_R$, let the split sizes be $\left|\mathcal{L}^{(i)}\right| = N_{\mathcal{L}}$, $\left|\mathcal{V}^{(i)}\right| = N_{\mathcal{V}}$, and $\left|\mathcal{T}^{(i)}\right| = N_{\mathcal{T}}$. Let $\Sigma$ be the search space for rules.[6] Let $R_{emp}\left(\mathcal{A}\left(\mathcal{S}_1\right), \mathcal{S}_2\right)$ be the empirical error over $\mathcal{S}_2$ of a model trained on $\mathcal{S}_1$:

$$R_{emp}\left(\mathcal{A}\left(\mathcal{S}_1\right), \mathcal{S}_2\right) \equiv \frac{1}{N_{\mathcal{S}_2}} \sum_{i=1}^{N_{\mathcal{S}_2}} L\left(y_i, f(\boldsymbol{x}_i)\right), \tag{4.30}$$

[5] $\mathcal{L}^{(i)} \cap \mathcal{V}^{(i)} = \emptyset,\ \mathcal{V}^{(i)} \cap \mathcal{T}^{(i)} = \emptyset,\ \mathcal{T}^{(i)} \cap \mathcal{L}^{(i)} = \emptyset,\ \forall i \leq N_R$

[6] For simplicity, hyperparameter tuning is considered within model selection.

where $f = \mathcal{A}(\mathcal{S}_1)$, $N_{\mathcal{S}_2} = |\mathcal{S}_2|$, and $L(\cdot)$ is a loss function. Model selection is performed as:

$$\mathcal{A}^* = \arg\min_{\mathcal{A}\in\Sigma} \left\{ \frac{1}{N_R} \sum_{i=1}^{N_R} R_{emp}\left(\mathcal{A}\left(\mathcal{L}^{(i)}\right), \mathcal{V}^{(i)}\right) \right\}, \tag{4.31}$$

while error estimation of the best algorithm $\mathcal{A}^*$ yielding the model $f(\mathcal{D})$, is given by:

$$\epsilon = \frac{1}{N_R} \sum_{i=1}^{N_R} R_{emp}\left(\mathcal{A}^*\left(\mathcal{L}^{(i)} \cup \mathcal{V}^{(i)}\right), \mathcal{T}^{(i)}\right) \tag{4.32}$$

Under 0–1 loss, with probability $(1-\delta)$, the expected error $R(f)$ can be bounded as:

$$R(f) \in \begin{bmatrix} Q_{\delta/2}\left\{\mathcal{B}\left(N_{\mathcal{T}}\epsilon, N_{\mathcal{T}} - N_{\mathcal{T}}\epsilon + 1\right)\right\} \\ Q_{1-\delta/2}\left\{\mathcal{B}\left(N_{\mathcal{T}}\epsilon + 1, N_{\mathcal{T}} - N_{\mathcal{T}}\epsilon + 1\right)\right\} \end{bmatrix}^T, \tag{4.33}$$

where $Q_n$ is the $n$-th quantile and $\mathcal{B}(\alpha, \beta)$ is the beta distribution with shape parameters $\alpha,\ \beta$.

# Chapter 5

# I DARE – Dataset Development[1]

This chapter presents I DARE (IULM Dataset of Affective REsponses), a novel multimodal physiological dataset. I DARE includes five modalities (EEG, SC, PPG, EMG, and ET) recorded from 63 participants exposed to 32 emotional images, selected from standard databases through a semi-automatic, bias-controlled procedure. Participants' arousal and valence ratings showed fair agreement both between subjects and with respect to the reference scores. I DARE surpasses existing datasets by relying on static stimuli and by achieving the largest dimension (10080) and the smallest minimum detectable effect size ($f = 0.095$).

## 5.1 Introduction

Affective datasets are collections of modalities annotated in terms of the corresponding affective states and are used to train Affect Detection models (D'Mello et al., 2017). The modalities can be either extracted from existing data (e.g., facial expressions from video clips) or newly recorded during experiments (e.g., EEG data from a sample of subjects). Furthermore, affective states can be induced by external stimuli or emerge naturally, as well as being posed or spontaneously emitted (D'Mello & Kory, 2015; S. Zhao et al., 2021).

In the literature, multiple affective datasets can be found, largely varying in terms of modalities, labels, and origin of the affective state. We took into account a representative sample of 58

[1]This chapter has been published as: Bilucaglia, M., Zito, M., Fici, A., Casiraghi, C., Rivetti, F., Bellati, M. and Russo, V. (2024). I DARE: IULM dataset of affective responses. *Frontiers in Human Neuroscience*, 18, 1347327. doi: 10.3389/fnhum.2024.1347327

Table 5.1: Qualitative comparison between I DARE and the eight most popular physiology-based datasets in terms of modalities and stimulus type.

| **Dataset** | **Modalities** | **Stimuli** |
|---|---|---|
| AMIGOS | EEG, ECG, SC | Video clips |
| ASCERTAIN | EEG, ECG, SC | Movie clips |
| BIO-VID-EMO DB | ECG, EMG, SC | Film clips |
| DEAP | EEG, SC, EMG, PPG, EOG, Resp | Music videos |
| DREAMER | EEG, ECG | Film clips |
| MAHNOB-HCI | EEG, ECG, SC, Resp, Temp, ET | Video clips |
| MPED | ECG, EEG, PPG, SC, Resp | Video clips |
| SEED | EEG | Movie videos |
| I DARE | EEG, EMG, SC, PPG, ET | Images |

datasets, gathered from two reviews (Ahmed et al., 2023; Siddiqui et al., 2022) and three recent research papers (J. Chen et al., 2022; Pant et al., 2023; Saganowski et al., 2022). Accordingly, the number of modalities varies from 1 to 7 (M=2.28, SD=1.24) and includes face (71.2%), voice (54.2%), EEG (23.7%), ECG (20.3%), PPG (17%), EMG (6.8%), and ET (6.8%). Overall, 81% of the datasets contain newly recorded modalities. The affective states, which are mostly labelled according to a discrete model (70.7%), are mainly induced (69%) and spontaneous (63%).

A recent survey (Ahmad & Khan, 2022) compared the stimuli, modalities, and sample size of the eight most popular datasets for the training of physiology-based AD models: AMIGOS (Miranda-Correa et al., 2018), ASCERTAIN (Subramanian et al., 2016), BIO-VID-EMO DB (L. Zhang et al., 2016), DEAP (Koelstra et al., 2011), DREAMER (Katsigiannis & Ramzan, 2017), MAHNOB-HCI (Soleymani et al., 2011), MPED (Song et al., 2019), and SEED (W.-L. Zheng et al., 2017). The stimuli, varying from 10 to 40 (M=23.5, SD=9.943), are all induced and include spontaneous responses. The number of modalities, varying from 1 to 6 (M=3.38, SD=1.60), includes EEG (87.5%), ECG (75%), SC (37.5%), respiration (37.5%), PPG (25%), EMG (25%), peripheral temperature – Temp (12.5%), and ET (12.5%). The sample size ranges from 15 to 86 (M=38, SD=23.43). The "dataset dimension", defined as the product between stimuli, modalities, and sample sizes, ranges from 225 to 7680 (M=3257.88, SD=2581.12). Finally, the stimuli are all dynamic, including videos (37.5%), films (25%), movie clips (12.5%), and music videos (12.5%). The features of the eight most popular datasets, plus the proposed one, are summarised in Tables 5.1 and 5.2.

For each dataset, we conducted statistical analyses using JASP (Love et al., 2019) to ex-

Table 5.2: Quantitative comparison between I DARE and the eight most popular physiology-based datasets in terms of number of modalities, stimuli, and subjects, as well as dimension and minimum detectable effect size (Cohen's f).

| **Dataset** | **Modalities** | **Stimuli** | **Subjects** | **Dimension** | **Cohen's f** |
|---|---|---|---|---|---|
| AMIGOS | 3 | 16 | 40 | 1920 | 0.149 |
| ASCERTAIN | 3 | 36 | 58 | 6264 | 0.095 |
| BIO-VID-EMO DB | 3 | 15 | 86 | 3870 | 0.104 |
| DEAP | 6 | 40 | 32 | 7680 | 0.125 |
| DREAMER | 2 | 18 | 23 | 828 | 0.191 |
| MAHNOB-HCI | 5 | 20 | 27 | 2700 | 0.170 |
| MPED | 4 | 28 | 23 | 2576 | 0.166 |
| SEED | 1 | 15 | 15 | 225 | 0.253 |
| I DARE | 5 | 32 | 63 | **10080** | **0.095** |

plore relationships between the number of modalities, stimuli, and sample size. Additionally, we performed sensitivity analyses using G*Power (Faul et al., 2007) to estimate, from the sample size and the number of stimuli, the minimum detectable effect size. No significant correlations between the number of modalities and subjects (Spearman's correlation: $\rho = 0.270, p = 0.518$), modalities and stimuli (Spearman's correlation: $\rho = 0.700, p = 0.053$), or subjects and stimuli (Spearman's correlation: $\rho = 0.084, p = 0.843$) emerged, suggesting that the dataset-specific characteristics are not bounded by technological or experimental constraints. The sensitivity analysis, which was based on the repeated-measures Analysis of Variance (ANOVA) model (stimulus ID as within-subject factor, $\alpha = 0.05$, $1 - \beta = 0.95$, $r = 0.5$, $\epsilon = 1$), showed an average minimum detectable effect size of $f = 0.157 \pm 0.051$, ranging from 0.095 to 0.253, interpretable as "small" to "medium" (Cohen, 1992).

Among the limitations of existing datasets, we believe that the strictest one is the use of videos as a unique elicitation method, which may lead to AD models unsuitable for the evaluation of static or short-term stimuli (e.g., static creatives, product packaging, or landing pages), as well as for a dynamic classification (Bilucaglia et al., 2021) of multi-frame stimuli (e.g., video commercials). A less severe, but still present, limitation is the varying number of stimuli, which affects the minimum detectable effect size and, in turn, the reliability and generalisability of the results (Funder & Ozer, 2019). Finally, modalities such as ET, SC, PPG, and EMG are not as widely represented.

To fill these gaps, we created I DARE, the IULM Dataset of Affective REsponses, a multimodal (five modalities: EEG, SC, PPG, EMG, and ET) physiological affective dataset gathered from 63 participants and based on 32 images as eliciting stimuli, corresponding to a dimension of $5 \times 32 \times 63 = 10080$. A sensitivity analysis, performed using G*Power under the same previously described conditions, showed a minimum detectable effect size of $f = 0.095$, considered as "small". As shown in Table 5.2, I DARE exceeds existing datasets in three key aspects: it is based on static pictures, it has the highest dimension, and it allows the smallest minimum detectable effect size.

## 5.2 Materials and Methods

### 5.2.1 Stimulus Extraction

We extracted the emotional stimuli from the International Affective Picture System (IAPS) (Lang et al., 2008) and the Open Affective Standardized Image Set (OASIS) (Kurdi et al., 2017). Rather than relying on a single database, we combined the two to benefit from their complementary features. IAPS is the most widely adopted benchmark in affective research (Branco et al., 2023), but its stimuli were selected before the web era and rated by a relatively limited convenience sample. OASIS, in turn, provides more contemporary stimuli that have been evaluated through large-scale online assessments with a broader participant base.

IAPS consists of 1194 realistic images, rated for arousal, valence, and dominance through a nine-point Self-Assessment Manikin (SAM) (Bradley & Lang, 1994). The raters were college students enrolled in the experiment as part of a course requirement. Each picture was rated by approximately 100 participants (about 50% females). Valence and arousal scores showed strong internal consistency (split-half $r_S \geq .93$) and high test–retest reliability ($r_S \geq .97$). For every image, the dataset provides the mean $\left(m_{(\cdot)}\right)$ and standard deviation $\left(s_{(\cdot)}\right)$ across raters for valence ($v$), arousal ($a$), and dominance ($d$).

OASIS consists of 900 realistic images rated for arousal and valence through a seven-point Likert scale. The raters were 827 paid participants (about 51% females), recruited through a mass crowdsourcing website. Each participant rated either the valence or arousal of a random subset of 255 images, and each picture obtained 101 to 108 different ratings. Valence and arousal scores showed strong internal consistency (split-half $r_S \geq .93$). For every image, the dataset provides the mean $\left(m_{(\cdot)}\right)$ and standard deviation $\left(s_{(\cdot)}\right)$ across raters for the valence ($v$) and arousal ($a$) dimensions.

In order to reduce experimental biases, for each dataset we performed the stimulus selection through a seven-step semi-automatic procedure. Steps 1–5 (automatic) were based on the geometrical properties of the stimuli in the two-dimensional valence–arousal space, which were considered either as points $\boldsymbol{x} = (m_a, m_v)^T$, or Gaussian vectors $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = (m_v, m_a)^T$ and $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} s_v & 0 \\ 0 & s_a \end{smallmatrix}\right)$. Steps 6–7 (manual) were based on the ranking of the ratings provided by independent judges.

Figure 5.1 shows the distribution of the stimuli (selected ones are marked with circles) and the centroids (squares) across the four quadrants of the IAPS and OASIS datasets.

**Step 1.** We computed the central point $\boldsymbol{c}$ and the peak-to-peak range $p$ as, respectively:

$$\boldsymbol{c} \doteq (v_c, a_c)^T = \left(min\{R\} + \frac{max\{R\} - min\{R\}}{2}, min\{R\} + \frac{max\{R\} - min\{R\}}{2}\right)^T \tag{5.1}$$

$$p = max\{R\} - min\{R\} \tag{5.2}$$

**Step 2.** We extracted the subsets of points belonging to the four affective quadrants HVHA (High Valence High Arousal), HVLA (High Valence Low Arousal), LVHA (Low Valence High Arousal), and LVLA (Low Valence Low Arousal), defined as:

$$\text{HVHA} = \{\boldsymbol{x} \mid v > v_c + 0.1 \cdot p, a > a_c + 0.1 \cdot p\} \tag{5.3}$$

$$\text{HVLA} = \{\boldsymbol{x} \mid v > v_c + 0.1 \cdot p, a < a_c - 0.1 \cdot p\} \tag{5.4}$$

$$\text{LVHA} = \{\boldsymbol{x} \mid v < v_c - 0.1 \cdot p, a > a_c - 0.1 \cdot p\} \tag{5.5}$$

$$\text{LVLA} = \{\boldsymbol{x} \mid v < v_c - 0.1 \cdot p, a < a_c - 0.1 \cdot p\} \tag{5.6}$$

The above-defined equations excluded the set of near-neutral points defined as:

$$\{\boldsymbol{x} \mid v_c - 0.1 \cdot p \leq v \leq v_c + 0.1 \cdot p, a_c - 0.1 \cdot p \leq a \leq a_c + 0.1 \cdot p\} \tag{5.7}$$

**Step 3.** For each quadrant $l = \{\text{HVHA, HVLA, LVHA, LVLA}\}$, we computed Gaussian centroid vectors $\boldsymbol{\pi}_l = \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$, where the mean vector and the covariance matrix are defined, respectively, as:

$$\begin{aligned} \boldsymbol{\mu}_l &= \frac{1}{\#\mathcal{S}_l} \sum_{\boldsymbol{x} \in \mathcal{S}_l} \boldsymbol{\mu}_{\boldsymbol{x}} \\ \boldsymbol{\Sigma}_l &= \frac{1}{\#\mathcal{S}_l} \sum_{\boldsymbol{x} \in \mathcal{S}_l} \boldsymbol{\Sigma}_{\boldsymbol{x}} \end{aligned} \tag{5.8}$$

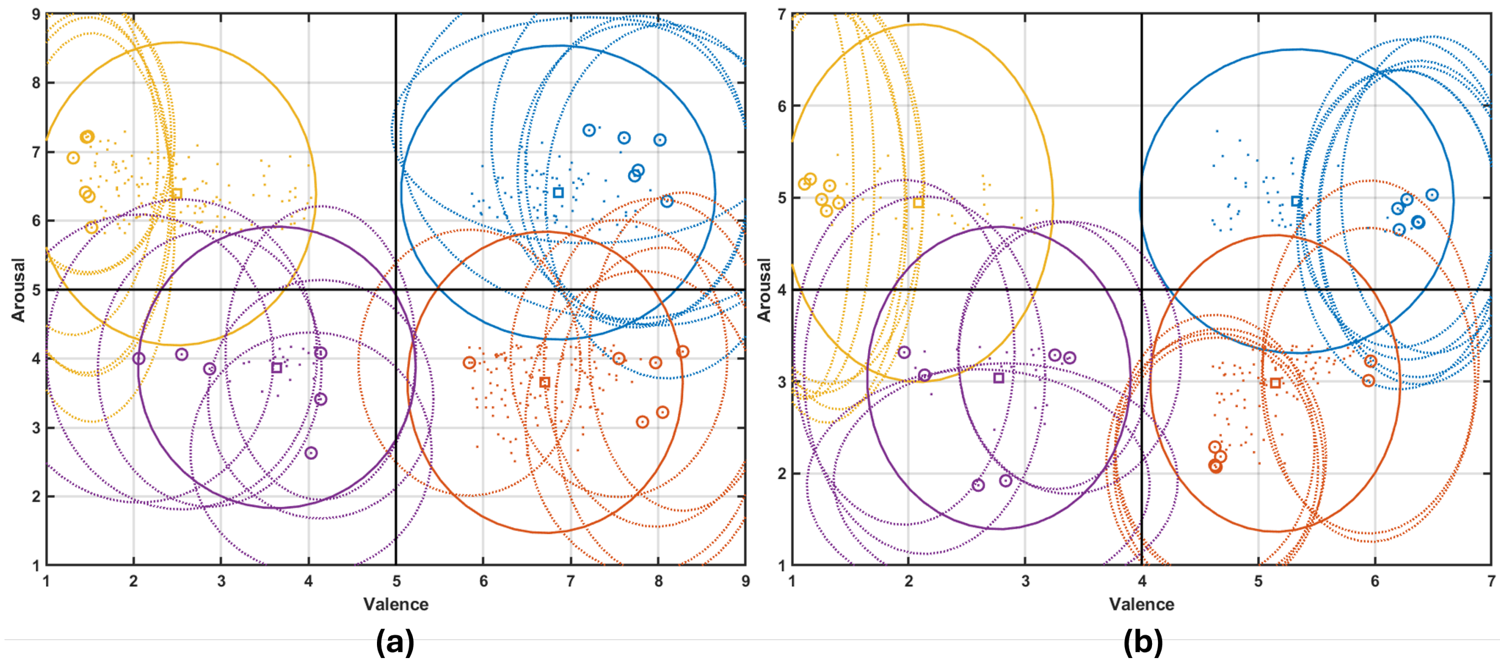


Figure 5.1: Distribution of the stimuli (dots) and centroids $\boldsymbol{\mu}_l$ (squares) for the IAPS (panel **a**) and OASIS (**b**) datasets. The six selected stimuli per quadrant are marked as circles. For both the centroids and the selected stimuli, the ellipsoids are drawn with solid and dashed lines, respectively.

where $\boldsymbol{\mu_x}$ and $\boldsymbol{\Sigma_x}$ are, respectively, the mean vector and covariance matrix of $\boldsymbol{x}$, $\mathcal{S}_l$ is the set of $\boldsymbol{x}$ belonging to the quadrant $l$, and $\#\mathcal{S}_l$ is the cardinality of $\mathcal{S}_l$.

**Step 4.** For each vector $\boldsymbol{x} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathcal{S}_l$, we computed its distribution similarity with the centroid vector $\boldsymbol{\pi}_l$ by means of the Bhattacharyya distance $d(\boldsymbol{x}, \boldsymbol{\pi}_l)$, defined as (E. Choi & Lee, 2003):

$$d(\boldsymbol{x}, \boldsymbol{\pi}_l) = \frac{1}{8}(\boldsymbol{\mu} - \boldsymbol{\mu}_l)^T \left(\boldsymbol{\Sigma}'\right)^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_l) + \frac{1}{2} \ln \left( \frac{\det\left(\boldsymbol{\Sigma}'\right)}{\sqrt{\det(\boldsymbol{\Sigma}) \det(\boldsymbol{\Sigma}_l)}} \right) \tag{5.9}$$

where $\boldsymbol{\Sigma}' = (\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_l)/2$, $\ln(\cdot)$ is the natural logarithm, and $\det(\cdot)$ is the determinant.

**Step 5.** From each quadrant $l$, we extracted the six stimuli with the highest Bhattacharyya distances.

**Step 6.** Nine independent raters selected, from each set of six stimuli, a subset of four.

**Step 7.** We selected the four stimuli that obtained the highest percentage agreement (McHugh, 2012) across the raters. In cases of parity, we considered the highest distance from the centroid.

### 5.2.2 Sample Population and Experimental Protocol

Sixty-three people (24 males) aged between 20 and 30 years (M=22.89, SD=2.38) were enrolled in the study. Statistical analyses performed using JASP showed that the sample was not significantly gender unbalanced in terms of proportions (binomial test: Male $= 38.1\%, p = 0.077$) or mean age (Mann–Whitney U test: $W = 3995, p = 0.325$).

Each subject sat on a chair placed in front of a 23.8" ($527 \times 296.5mm^2$) FlexScan EV2451 (Eizo, KK) monitor located in a $7 \times 3m^2$ room. At the beginning of the experiment, the subject performed a 60s-long eyes-closed baseline and a 120s-long neutral baseline (white fixation dot on a black background). This was preceded by the baseline instructions and followed by the experiment description. Then, the 32 emotional blocks were presented in random order. Each block consisted of a black screen (5s), used as a short inter-stimulus interval to reduce immediate carry-over from the previous trials, followed by the emotional stimulus (5s) and the valence/arousal evaluation, performed by means of a 9-point SAM. Figure 5.2 shows a schematic representation of the experimental protocol.

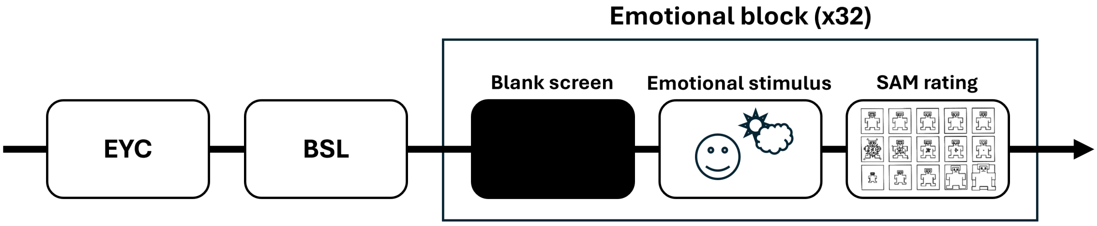


Figure 5.2: Schematic representation of the experimental protocol.

The quality of the signals was visually checked before starting the recording, and the contact impedance of the EEG electrodes was reduced below $10k\Omega$ (Sinha et al., 2016) by means of a scrub cream (Neurprep by Spes Medica, S.p.A.) and a conductive gel (Neurgel by Spes Medica, S.p.A.).

### 5.2.3 Instrumentation

We recorded the EEG signal using the NVX-52 device (Medical Computer Systems, Ltd.) at a sample frequency of $f_s = 1kHz$. Forty passive wet Ag / AgCl electrodes were placed at standard locations (Russo et al., 2023) of the 10-10 system (Nuwer, 2018), internally referenced and grounded to the left mastoid (M1) by means of 1 Ag/AgCl passive adhesive patch. We recorded the SC and PPG signals using, respectively, the GSRSens and FpSens (Medical Computer Systems, Ltd.) sensors, both connected to the auxiliary inputs of the NVX-52. We placed 2 passive dry Ag/AgCl electrodes of the GSRSens on the index and ring finger from the non-dominant hand, and the FpSens on the ring finger from the same hand. We recorded the EMG signal using two passive Ag/AgCl adhesive patches placed over the left *zygomaticus major* and *corrugator supercilii* muscles. The electrodes, connected to the bipolar inputs of the NVX-52, were referenced to M1. The recordings were controlled by the NeoRec software (Medical Computer Systems, Ltd.). We recorded the ET signal using the Tobii Spectrum (Tobii, LLC) device at a sample frequency of $f_s = 150Hz$. The collected measures included the horizontal and vertical coordinates of the right and left eyes ($x_R$, $y_R$, $x_L$, $y_L$, respectively), expressed in $mm$ with respect to the upper-left corner of the screen.

We used the iMotions software (iMotions, A/V) to deliver the stimuli, as well as to collect the eye-tracking data and the self-reports. At the beginning of the experiment, iMotions generated a Transistor–Transistor Logic (TTL) pulse that was fed into the digital inputs of the NVX-52 using the EEG Synchronisation Box (ESB) (Bilucaglia et al., 2020). This served to perform an offline synchronisation between the recorded data and the stimulus timestamps.

### 5.2.4 Data Processing

We aligned the recorded data with the TTL pulse and appended the stimuli according to the iMotions-generated timestamps. Then, we cleaned the data from noise and artefacts for subsequent feature extraction (see Section 6.2.1), as described below. The operations were performed within the Matlab (The Mathworks, Inc.) environment, and the processed data were saved as `.mat` files.

**EEG**

We processed the EEG data using the EEGLab toolbox (Delorme & Makeig, 2004).

First, we applied a re-reference to the linked earlobes, a resampling to $f_S = 512Hz$, and a band-pass filter ($0.1 - -40Hz$, $4^{th}$ order, Butterworth type). To avoid phase distortions, the filter was acausal and time-reversed (i.e., *zero-phase*) (de Cheveigné & Nelken, 2019). Then, we applied an adaptive filter (`cleanline` plug-in) to remove the sinusoidal noise at frequencies $f = \{50, 100\}\,Hz$. The method computes, in a moving-window fashion ($w$), the least-square estimate of the noise component in the form of $s_w(t) = \alpha\cos(2\pi t) + \beta\sin(2\pi f t)$ and subtracts it from the original signal (Bigdely-Shamlo et al., 2015).

We corrected non-stationary artefacts through Artefact Subspace Reconstruction (ASR) (Mullen et al., 2015), which performs a moving-window Principal Component Analysis (PCA) decomposition of the EEG signal, compares the variance of the projected components with a calibration-derived reference threshold controlled by the parameter $\kappa$, and reconstructs the signal after suppressing high-variance components. We set $\kappa = 20$, since values in the 20–30 range have been shown to provide an effective trade-off between removing artefacts and preserving brain-related EEG activity (Chang et al., 2020).

We corrected stereotypical artefacts, such as blinks, eye movements, and EMG noise, through Independent Component Analysis (ICA) (Hyvärinen & Oja, 2000), which decomposes the EEG data matrix into statistically independent component activations (or simply Independent Components, ICs) by means of an unmixing matrix $W$, estimated through iterative algorithms. The corresponding mixing matrix, used for back-projection to the original channel space, is denoted by $A = W^{-1}$. Artefactual ICs, identified from their temporal activity and scalp topographies, are removed before back-projection, yielding the cleaned EEG signal (Jung et al., 1997).

ICA was achieved through the `RunICA` algorithm (Akhtar et al., 2012). To speed up the computations, we applied it to a copy of the original data that had been band-pass filtered ($1 - 30Hz$, $4^{th}$ order zero-phase Butterworth filter) and down-sampled ($f_s = 100Hz$) (Luck, 2022). The obtained unmixing matrix was then saved and multiplied by the original data to obtain the ICs. We identified artefactual ICs through the neural-net classifier ICLabel (Pion-Tonachini et al., 2019) as those with probability of "not-brain" `P{!brain}>0.9`; we set them to zero and back-projected them to the original space.

Finally, we re-referenced the artefact-free EEG data to the approximately zero potential through Reference Electrode Standardization Technique (REST) (Dong et al., 2017).

### SC

We resampled the SC data at $f_s = 32\,\mathrm{Hz}$ and applied a band-pass filter ($4^{th}$-order, zero-phase, Butterworth type) between 0.001 and $0.35Hz$. Artefactual samples were corrected using fixed amplitude and rate-of-change thresholds (Kleckner et al., 2018). Specifically, a sample $x$ at time $t_i$ was marked as artefactual if:

$$x(t_i) < 0.05\mu S \texttt{ OR } x(t_i) > 60\mu S \texttt{ OR } \left|\frac{x(t_i) - x(t_{i-1})}{t_i - t_{i-1}}\right| > 8\,\mu S/s \tag{5.10}$$

The identified samples and all the neighbours within a centred $5s$-long window were then deleted and linearly interpolated based on the remaining valid points.

### PPG

We resampled the PPG signal at $f_s = 32\,\mathrm{Hz}$ and applied a band-pass filter (0.001–$0.35Hz$, $2^{nd}$-order zero-phase Butterworth filter). Then, to attenuate slow drifts and motion artefacts, we applied the following demodulation (Dall'Olio et al., 2020):

$$x'(t) = \frac{x(t)}{\sqrt{x^2(t) + \mathcal{H}\{s(t)\}^2}}, \tag{5.11}$$

where $\mathcal{H}\{\cdot\}$ is the Hilbert transform.

### EMG

Since the power spectrum of the EMG does not overlap with common interference signals, a simple cascade of a band-stop and a pass-band filter is considered adequate for preprocessing

(Boyer et al., 2023). Accordingly, we applied a notch filter ($50Hz,\ 100Hz,\ q = 40$, zero-phase Infinite Impulse Response filter) to remove the power-line noise and a band-pass filter ($10 - -400Hz$, $4^{th}$ order zero-phase Butterworth filter) to attenuate low-frequency motion artefacts (Merletti & Cerone, 2020).

**ET**

We corrected temporal jitters by re-aligning the timestamps with the equipment-designed sample frequency ($fs = 150Hz$) and linearly interpolating the missing values (Liu et al., 2018).

To reduce outliers and missing data, including blink-related gaps, we filtered the coordinates $x_R$, $y_R$, $x_L$, and $y_L$ using the MFLWMD algorithm (Gucciardi et al., 2022). Specifically, for each sample point, the median of the available non-missing samples within a centred window of length $L$ was computed. Consecutive missing-data segments shorter than a threshold $g$ were ignored during median filtering, whereas longer gaps were used to split the time series into separate segments. We chose the following parameters: $L = 140ms$ and $g = 280ms$.

Finally, from the filtered coordinates, we computed the horizontal and vertical gaze positions as, respectively, $g_X = (x_R + x_L)/2$ and $g_Y = (y_R + y_L)/2$ (Kar, 2020).

## 5.3 Results

### 5.3.1 Stimuli Selection

The selected stimuli obtained a percentage agreement ranging from 56.56% to 100% (M=79.20%, SD=10.50%). Statistical analyses performed using JASP did not show any significant main effect for the factors "Database" ($F(1, 24) = 0.143,\ p = 0.709$) and "Quadrant" ($F(3, 24) = 2.381,\ p = 0.095$), or for the "Database×Quadrant" ($F(3, 24) = 0.048,\ p = 0.986$) interaction. Additionally, the agreement did not significantly correlate with the mean ($m$) and standard deviations ($s$) of the overall ratings in either the valence ($\rho_m = 0.194,\ p = 0.288; \rho_s = -0.175,\ p = 0.339$) or the arousal ($\rho_m = -0.142,\ p = 0.439;\ \rho_s = 0.109, p = 0.553$) dimension.

### 5.3.2 SAM Evaluations

We coded the subjects' evaluations (SAM numerical scores) into labels corresponding to the four quadrants (HVLA, LVLA, HVHA, LVHA), considering the midpoint value 5 as the threshold for low/high discrimination. We measured the overall agreement within the subjects ($WS$) and between the subjects and the gold standard ($GS$) given by the selection process, by means of Fleiss' kappa (Fleiss, 1971). The analyses, performed using JASP, showed values of $\kappa_{WS} = 0.406$ and $\kappa_{GS} = 0.409$, both interpreted as a "fair agreement" (Landis & Koch, 1977).

## 5.4 Discussion

Existing physiological datasets rely on dynamic eliciting stimuli, making them unsuitable for training AD models for the evaluation of images or the dynamic frame-by-frame classification of videos. Furthermore, the available datasets are characterised by a widely varying number of subjects, stimuli, and modalities, negatively affecting their reliability and generalisability.

To fill these gaps, we created I DARE (the IULM Dataset of Affective REsponses), a multimodal physiological dataset that outperforms existing ones in terms of dimension (5 modalities $\times$ 32 stimuli $\times$ 63 subjects) and minimum detectable effect size ($f = 0.095$, small). We selected the stimuli (static pictures) from the two standard databases (IAPS and OASIS) through a semi-automatic procedure to reduce selection biases. We obtained "fair" agreements within subjects and between subjects and the original scores ($\kappa_{WS} = 0.406$ and $\kappa_{GS} = 0.409$, respectively).

I DARE may support several applications. In Consumer Neuroscience, it could be used for the emotional assessment of static stimuli, such as product packaging and landing pages, and for the dynamic frame-by-frame classification (Bilucaglia et al., 2021) of video commercials. More broadly, it could support the training and benchmarking of AD models in Affective Computing and Neuroergonomics, the evaluation of emotional selectivity and invariance of existing neurometrics, and the study of the physiological correlates of emotion in Psychology and Cognitive Neuroscience.

Several limitations of I DARE should be acknowledged. First, although the sample was not significantly gender unbalanced, it mainly consisted of young adults. This may limit the generalisability of the dataset and of models trained on it, since age can modulate physiological responses to emotional stimuli, including electrodermal, cardiovascular, pupillary, and facial-muscle reactivity (Gomez et al., 2016; Neiss et al., 2009). Future extensions should therefore include broader age ranges to increase both sample size and representativeness. Second, the experimental timing may not have ensured complete physiological recovery between trials. The 5 s black screen provided a short inter-stimulus interval, but slower autonomic responses, especially skin-conductance and cardiac dynamics, can outlast this interval and partially overlap across trials (Boucsein et al., 2012; Shaffer & Ginsberg, 2017). Future protocols should consider longer or adaptive recovery periods, particularly when autonomic features are central to the analysis. Finally, the use of standardised affective images increased experimental control but may limit ecological validity. Affective databases such as IAPS and OASIS provide useful normative ratings, but stimulus effectiveness can vary across historical, cultural, and demographic contexts (Betella & Verschure, 2016; Riegel et al., 2017). Moreover, static images only partially resemble Consumer Neuroscience settings, where emotional responses often emerge during exposure to advertisements, products, websites, social-media content, or purchase-related decisions. AD models trained on I DARE should therefore be further validated using more ecological and marketing-relevant stimuli.

# Chapter 6

# MATILDE – Model Development

This chapter presents MATILDE (Multimodal biosignAl-based modaliTy-resilient subject-Independent modeL for affect DEtection), a multimodal subject-independent affect detection (AD) model based on the standard I DARE dataset. Temporal, spectral, and time–frequency features are extracted from electroencephalographic (EEG), skin-conductance (SC), photoplethysmographic (PPG), electromyographic (EMG), and eye-tracking (ET) modalities. A Combined Algorithm Selection and Hyperparameter optimisation (CASH) problem is formulated over multiple classifiers and feature-selection methods and solved via Bayesian optimisation. Random Forest models achieved the best performance (76.2% for quadrants, 83.4% for valence, and 85.3% for arousal), all above the benchmark under the adopted conservative criterion. Feature-importance analyses reveal EEG domination and a prevalence of time-domain features across modalities.

## 6.1 Introduction

The development of an affect detection (AD) model typically involves four main steps (Bota et al., 2019; Shu et al., 2018): (1) feature construction, (2) feature selection, (3) model training, and (4) model testing. In machine learning (ML) terminology, steps (1) and (2) constitute the feature-engineering phase, whereas steps (3) and (4) belong to the learning phase (Theodoridis & Koutroumbas, 2009, pp. 4–7).

Both phases depend on the chosen fusion strategy (see Section 3.3), which, for most physiology-based AD models, involves either early or late fusion approaches (Ahmad & Khan, 2022). With large datasets, early fusion is often preferred, possibly combined with dimensionality-reduction techniques. It offers conceptual simplicity and enables the model to exploit interdependencies across modalities. Conversely, with limited data, late fusion may be more effective, as it leverages the complementary strengths of multiple classifiers, improving robustness and generalisation (Pereira et al., 2023).

In traditional ML (tML) for AD, the pipeline is modality-specific and relies on domain knowledge for feature extraction and selection (Bota et al., 2019). By contrast, deep-learning approaches integrate data representation in the training phase, reducing reliance on explicit, hand-crafted feature engineering (Chutia & Baruah, 2024).

Nevertheless, the view that DL has fully superseded hand-crafted feature engineering and careful model selection and hyperparameter tuning is overstated. Deep models typically require substantial labelled data to achieve competitive performance (Younis et al., 2024). With specific data formats, tML methods such as decision trees can even outperform deep neural networks (Grinsztajn et al., 2022). Moreover, DL models are generally less interpretable (X. Zhang et al., 2025) and can be brittle under distribution shift, whereas pipelines built on hand-crafted features have shown better out-of-distribution generalisation (Bento et al., 2022).

In the following, we present a review of state-of-the-art AD models based on tML methods, together with the associated design choices for the feature-engineering and learning phases. When possible, we discuss each step separately for the different modalities considered in this work (i.e., EEG, SC, PPG, EMG, and ET).

### 6.1.1 Feature Construction

Features are compact representations of signals that are obtained not only in the time domain but also through appropriate transformations in the frequency and time–frequency domains (Bota et al., 2019; Shu et al., 2018).

Among the 53 EEG studies included in the review by Al-Nafjan et al. (2017), 17 (32%) used time-domain features, 29 (55%) frequency-domain features, and 7 (13%) time–frequency features. This trend is confirmed, more recently, by Erat et al. (2024): out of 216 studies examined, 96 (44%) included time-domain features, 147 (68%) frequency-domain features, and 76 (35%) time–frequency features.

Among the 92 PPG studies examined by Sayed Ismail et al. (2024), 11 (12%) included only time-domain features, 1 (1%) only frequency-domain features, and 65 (71%) combinations of both. No time–frequency features were reported.

In the study by Shukla et al. (2021), a total of 25 SC works were examined. Among these, 22 (88%) used time-domain features, 6 (24%) frequency-domain features, and 3 (12%) time–frequency features, computed on the raw SC signal or on its decomposition into Skin-Conductance Level (SCL) and Skin-Conductance Response (SCR).

In her experimental work, Rutkowska et al. (2024) generically mentioned 31 EMG studies using only time-domain features. Considering the scarcity of works on EMG in affect detection, the literature survey was broadened to the more general use of ML methods for EMG signal classification, finding the review by Afroza et al. (2023), which focused on hand-movement identification. Out of the 33 studies, they found 22 (66%) using time-domain features, 4 (12%) frequency-domain features, and 2 (6%) time–frequency features.

The review by Skaramagkas et al. (2021) highlighted that, out of a total of 22 ET studies, only time-domain features were used, computed on fixations (duration: 7, 32%; quantity: 10, 45%; total time: 5, 23%; filtered amplitude: 2, 9%), saccades (amplitude: 2, 9%; velocity: 1, 5%; quantity: 1, 5%; fractions of micro-saccades: 3, 14%), blink frequency (1, 5%), and pupil diameter (2, 9%).

### 6.1.2 Feature Selection

Training a model on a dataset where $n$ is greater than, or even comparable to, the number of instances $N$ leads to several issues (e.g., sparsity, multicollinearity, inflation of statistical significance, and overfitting), which fall under the so-called "curse of dimensionality" (Altman & Krzywinski, 2018). Furthermore, the feature-extraction process may generate irrelevant, redundant, or noisy elements that negatively affect model performance (Chandrashekar & Sahin, 2014). Feature selection aims to reduce the dimensionality of features $\boldsymbol{x} \in \mathbb{R}^n$ in order to maximise their discriminative power while simultaneously minimising the computational costs associated with training and using the model. It is based on the selection, among the $n$ elements $\{x_k\}_{k=1}^n$, of the best subset $\{x_{s_i}\}_{i=1}^d$ , $s_i \in [1, n]$ (Bota et al., 2019; Shu et al., 2018).

The two most common feature-selection families in AD are unsupervised and supervised methods, with the latter further subdivided into filter and wrapper types (Bota et al., 2019; Khare et al., 2024; Shu et al., 2018). Filter methods are based on ranking techniques to score each element $x_k$ and remove those below a proper threshold. They are computationally efficient and help avoid overfitting, but may select redundant or suboptimal subsets and overlook inter-feature relationships or algorithm compatibility. Wrapper methods use the classifiers as black boxes and the corresponding performance as the objective function to evaluate the feature subsets $\{x_{s_i}\}_{i=1}^d$ , $s_i \in [1, n]$. They are usually associated with better performance, but are computationally expensive and prone to overfitting (Chandrashekar & Sahin, 2014).

### 6.1.3 Model Training

AD models employ various classifiers, including generative-parametric (e.g., LDA and QDA) and non-parametric (e.g., kNN and Naive Bayes) methods, as well as informative (e.g., SVM and ANN), decision-tree (DT), and ensemble methods.

Al-Nafjan et al. (2017) identified that, out of a total of 45 EEG studies, 24 (53%) used SVM as a classifier, 10 (22%) used kNN, 4 (9%) used Bayesian (B) or discriminant analysis (either LDA or QDA) classifiers, and 7 (16%) used ANN. This trend is partially confirmed by Erat et al. (2024): out of 216 studies examined, 87 (40%) were based on SVM, 35 (16%) on kNN, 39 (18%) on B or DA, and 78 (36%) on ANN. In addition, 9 (4%) studies emerged based on decision trees (DT) and 23 (11%) on ensemble classifiers. Finally, in the review by Khare et al. (2024), which included 49 studies, 15 (30%) adopted SVM, 4 (8%) kNN, 22 (44%) ANN, and 5 (10%) ensemble methods.

Sayed Ismail et al. (2024) highlighted that, out of a total of 92 SC studies, 61 (66%) used SVM as a classifier, 30 (33%) kNN, 27 (29%) DA or B, 47 (51%) ANN, and 48 (52%) DT. Khare et al. (2024) partially confirmed this trend, observing, in 23 studies, 7 (30%) based on SVM, 2 (9%) on kNN, 3 (13%) on DA or B, and 2 (9%) on ensemble methods.

The review by Khare et al. (2024) reported that, out of a total of 18 PPG studies, 5 (28%) were based on SVM, 1 (6%) on kNN, 1 (6%) on DA or B, and 2 (11%) on ensemble methods.

From the review by Afroza et al. (2023), which, as already mentioned, examined 33 studies on hand-movement recognition from EMG signals, the following results emerged: 15 (45%) studies were based on SVM, 7 (21%) on kNN, 5 (15%) on DA or B, and 19 (57%) on ANN.

Skaramagkas et al. (2021) highlighted that, out of a total of 22 ET studies, 15 (68%) used SVM, 3 (14%) kNN, 6 (27%) DA or B, 3 (14%) ANN, 3 (14%) DT, and 7 (32%) ensemble methods. This trend is partially confirmed by Khare et al. (Khare et al., 2024), who observed, in 6 studies, 3 (50%) based on SVM and 3 (50%) on ANN.

Model selection and hyperparameter tuning aim to find the best algorithm $\mathcal{A}^*$ and hyperparameters $\boldsymbol{\lambda}^*$ for a given dataset $\mathcal{D}$.

Although originally considered dependent but distinct tasks, they have recently been framed as a simultaneous Combined Algorithm Selection and Hyperparameter optimisation (CASH) problem (Feurer et al., 2015). Accordingly, given a set of learning algorithms $\mathcal{A}$ with associated ordered set of hyperparameters $\Lambda$, and given a dataset $\mathcal{D}$, the CASH problem aims to find a solution to:

$$\arg\min_{\substack{A\in\mathcal{A}\\ \boldsymbol{\lambda}\in\Lambda}} \left\{ R\left(A_{\boldsymbol{\lambda}}\left(\mathcal{D}\right)\right)\right\}, \tag{6.1}$$

The CASH problem rarely has a universal objective function to be optimised. Thus, it is often treated as a black-box problem to be tackled by methods including random and grid search, genetic algorithms, Monte Carlo techniques, reinforcement learning, and Bayesian optimisation. The latter, in particular, has been shown to be particularly promising because of its speed in large and conditional search spaces, as well as its ability to handle both numerical and categorical variables (Baratchi et al., 2024).

Bayesian optimisation is a sequential model-based search algorithm. It considers the objective function $f(\cdot)$ as unknown but evaluable at every query point $\boldsymbol{x} \in \mathcal{X}$. The evaluation produces a noise-corrupted output $y \in \mathbb{R}$ such that $f(\boldsymbol{x}) = \mathbb{E}\left\{y|f(\boldsymbol{x})\right\}$. At each iteration $i$, a location $\boldsymbol{x}_{i+1}$ is selected at which $f(\cdot)$ is queried and $y_{i+1}$ is observed. $\boldsymbol{x}_{i+1}$ is selected as the maximiser of an acquisition function $a_n : \mathcal{X} \to \mathbb{R}$ that evaluates the utility of candidate points for the next evaluation based on past points $\boldsymbol{x}_{k\leq i}$ and evaluations $y_{k\leq i}$ (Shahriari et al., 2016).

In AD, classifier training can typically follow two different strategies: the subject-dependent approach, which involves a classifier for each subject, and the subject-independent approach, which trains a single classifier on the entire dataset (Bota et al., 2019; Shu et al., 2018). According to D'Mello and Kory (2015), 62% of the AD models are subject-dependent, while only 38% are subject-independent. Subject-dependent and subject-independent approaches present complementary trade-offs: the former often attains higher within-subject accuracy at the cost of large per-subject data and limited generalisation, whereas the latter generalises to unseen subjects but usually with lower performance (Arevalillo-Herráez et al., 2019).

### 6.1.4 Model Testing

Model testing refers to the assessment of discrepancies between existing and required functionalities of a system. It includes, among other aspects[1], *correctness*, namely the probability of error when classifying future data (J. M. Zhang et al., 2022).

Among the six main approaches for model correctness testing[2], the so-called resampling (or

[1]According to J. M. Zhang et al. (2022), model assessment includes (i) correctness, (ii) relevance, (iii) robustness, (iv) security, (v) privacy, (vi) efficiency, (vii) fairness, and (viii) interpretability.
[2]According to Oneto (2020, p. 21), they include: (i) resampling methods, (ii) complexity-based methods, (iii) compression bound, (iv) algorithmic stability theory, (v) PAC-Bayes theory, and (vi) differential privacy theory.

out-of-sample) methods estimate generalisation by means of a set of data not previously "seen" during the training phase. More specifically, the training set $\mathcal{D}$ is resampled (with or without replacement) $N$ times to build three independent datasets, called *learning* $\left(\mathcal{D}_l^{(i)}\right)$, *validation* $\left(\mathcal{D}_v^{(i)}\right)$, and *testing* $\left(\mathcal{D}_t^{(i)}\right)$. As detailed in Section 4.4, model selection and hyperparameter tuning are performed by training the model on $\mathcal{D}_l^{(i)}$ and testing on $\mathcal{D}_v^{(i)}$, while error estimation is performed by training it on $\mathcal{D}_l^{(i)} \cup \mathcal{D}_v^{(i)}$ and testing on $\mathcal{D}_t^{(i)}$. The computed errors at each repetition $i \leq N$ are finally averaged.

Subject-dependent AD models typically follow the $k$-fold cross-validation (kCV) or leave-one-out (LOO) schemes (Shu et al., 2018). kCV randomly splits the training set into $k$ disjoint partitions and, for each $i \leq K$ repetition, considers the $k$-th partition as $\mathcal{D}_v^{(i)}$ and the remaining folds as $\mathcal{D}_l^{(i)}$. LOO is a special case of kCV where $k$ is equal to the number of elements in the training set (Airola et al., 2011).

In order to prevent subject-wise data leakage (Aljalal et al., 2024), subject-independent AD models typically follow the leave-$p$-subjects-out (LPSO) or leave-one-subject-out (LOSO) schemes (Ahmad & Khan, 2022). LPSO splits the training set by subject, rather than per feature as in kCV, and considers the set of $p$ subjects as $\mathcal{D}_t$ (Vázquez et al., 2021). LOSO is a special case of LPSO that considers one subject only as $\mathcal{D}_t$ (Fazli et al., 2009).

## 6.2 Materials and Methods

We designed, trained, and evaluated MATILDE (<u>M</u>ultimodal biosign<u>A</u>l-based modali<u>T</u>y-resilient subject-<u>I</u>ndependent mode<u>L</u> for affect <u>DE</u>tection), a multimodal, subject-independent AD model to predict valence, arousal, and affective quadrants from physiological signals. To this end, we developed a complete pipeline encompassing feature extraction, feature selection, model training, and evaluation within a unified optimisation framework.

I DARE (see Chapter 5) served as the affective dataset. Features were extracted from all modalities (EEG, SC, PPG, EMG, and ET), and the original gold-standard annotations were used as target variables, as they showed agreement with subjective ratings of $\kappa = 0.409$, interpreted as "fair" according to Landis and Koch (1977). In addition, quadrant ($Q$, 4 classes: HVLA, LVLA, HVHA, LVHA), valence ($V$, 2 classes: HV, LV), and arousal ($A$, 2 classes: HA, LA) labels were considered.

Feature-engineering and learning phases were performed in MATLAB (The MathWorks, Inc.) using the Signal Processing and Statistics & Machine Learning Toolboxes. All computations were carried out on a workstation equipped with an AMD Ryzen™ Threadripper PRO 5975WX CPU (32 cores, 64 threads, 3.6 GHz base clock) and 256 GB DDR4-3200 MHz ECC RAM. No GPU acceleration was used; however, Bayesian optimisation for model selection and hyperparameter tuning was executed in parallel (32 workers) using the Parallel Computing Toolbox.

### 6.2.1 Feature Extraction

**EEG**

The Individual Alpha Frequency (IAF) was computed as the centre of gravity of the Power Spectral Density (PSD) in the extended alpha band $\alpha' = [7.5,\ 12.5]$Hz as (Klimesch, 1999):

$$\mathrm{IAF} = \frac{\int_{\alpha'} f p'(f) df}{\int_{\alpha'} p'(f) df}, \tag{6.2}$$

where $p'(f)$ is the mean occipital PSD, averaged across the occipital channels PO3, POz, PO4, PO7, O1, Oz, and O2. The channel-wise PSDs were estimated according to Welch's method ($1s$-long Hamming window, 50% overlap), considering the eyes-closed (EYC) epoch only (Bilucaglia et al., 2019).

The IAF served to define the following subject-specific $\delta,\ \theta,\ \alpha,\ \beta$, and $\gamma$ EEG bands (Borghini et al., 2019):

$$\begin{aligned}
\delta =& [0,\ \mathrm{IAF} - 6]\,\mathrm{Hz} \\
\theta =& [\mathrm{IAF} - 6,\ \mathrm{IAF} - 2]\,\mathrm{Hz} \\
\alpha =& [\mathrm{IAF} - 2,\ \mathrm{IAF} + 2]\,\mathrm{Hz} \\
\beta =& [\mathrm{IAF} + 2,\ \mathrm{IAF} + 16]\,\mathrm{Hz} \\
\gamma =& [\mathrm{IAF} + 16,\ \mathrm{IAF} + 25]\,\mathrm{Hz}
\end{aligned} \tag{6.3}$$

Features belonging to the time, frequency, information, and time–frequency domains were extracted for each EEG channel $x^C(t)$.

Temporal features included the following mean $\bar{x}^C$ and median $\tilde{x}^C$ values, as well as Hjorth's

(1970) activity $\mathcal{A}^C$, mobility $\mathcal{M}^C$, and complexity $\mathcal{C}^C$ parameters:

$$
\begin{aligned}
\bar{x}^C &= \mathrm{Mean}_t \left\{ x^C(t) \right\}, \\
\tilde{x}^C &= \mathrm{Med}_t \left\{ x^C(t) \right\} \\
\mathcal{A}^C &= \mathrm{Var}_t \left\{ x^C(t) \right\} \\
\mathcal{M}^C &= \sqrt{\frac{\mathrm{Var}\left\{ \Delta^1 x^C(t) \right\}}{\mathrm{Var}\left\{ x^C(t) \right\}}} \\
\mathcal{C}^C &= \frac{\sqrt{\frac{\mathrm{Var}\{\Delta^2 x^C(t)\}}{\mathrm{Var}\{\Delta^1 x^C(t)\}}}}{\mathcal{M}^C},
\end{aligned}
\tag{6.4}
$$

where $\mathrm{Mean}_t\{\cdot\}$, $\mathrm{Med}_t\{\cdot\}$, and $\mathrm{Var}_t\{\cdot\}$ are, respectively, the sample mean, median, and variance operators, computed across the time dimension. $\Delta^k x^C(t),\ k = 1, 2$, is the $k$-th finite difference of $x^C(t)$.

As spectral features, the normalised spectral powers in the bands $B \in \{\delta, \theta, \alpha, \beta, \gamma\}$ were computed as (Bilucaglia et al., 2022):

$$
p_B^C = \frac{\int_B x^C(f)\, df}{\int x^C(f)\, df}, \tag{6.5}
$$

where $x^C(f)$ is the PSD of channel $C$, estimated using Welch's method with a 1 s Hamming window and 50% overlap.

Information-related features included measures of entropy and fractal dimension. Entropy consisted of Pincus' (1991) Approximate Entropy:

$$
\mathcal{E}^C \doteq \mathrm{ApE}\left\{ x^C(t) \right\}(m, \tau, r) \tag{6.6}
$$

with embedding dimension $m = 2$, lag $\tau = 1$, and radius $r = 0.2 \cdot \mathrm{Std}\left\{ x^C(t) \right\}$, where $\mathrm{Std}\{\cdot\}$ is the sample standard deviation operator, computed across the time dimension.

Fractal dimension was computed according to Katz's (1988) definition:

$$
\mathcal{F}^C = \frac{\log_{10}(n)}{\log_{10}(n) + \log_{10}(d/L)}, \tag{6.7}
$$

where $n$ is the number of steps in $x^C(t)$ (one less than the number of sample points $N$). $L$ and

$d$ are, respectively, the total length and the diameter of the curve, defined as:

$$\begin{aligned} L &= \sum_{i=1}^{N-1} \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|_2 \,, \\ d &= \max_{i \le N} \|\boldsymbol{x}_i - \boldsymbol{x}_1\|_2 \,. \end{aligned} \tag{6.8}$$

where $\|\cdot\|_2$ is the Euclidean norm and $\boldsymbol{x}_i$ is the $i$-th point in the time-amplitude space defined as $\boldsymbol{x_i} = \left(t_i, x^C(t_i)\right)^T$.

To compute the time–frequency features, a 3-level Discrete Wavelet Transform (DWT) was applied to each channel-wise EEG signal $x^C(t)$. Under the filter-bank interpretation, the DWT recursively decomposes the signal into approximation ($a_k^C[\ell]$) and detail ($d_k^C[\ell]$) coefficients through low-pass and high-pass filtering, followed by downsampling by a factor of 2, as (Mallat, 1989):

$$\begin{aligned} a_0^C[n] &\doteq x^C[n] \\ a_k^C[\ell] &= \sum_n h[n - 2\ell]\, a_{k-1}^C[n] \\ d_k^C[\ell] &= \sum_n g[n - 2\ell]\, a_{k-1}^C[n] \end{aligned}, \tag{6.9}$$

where $x^C[n] \doteq x^C\left(nT_s\right)$ is the sampled signal at sample interval $T_s$. $h[\cdot]$ and $g[\cdot]$ are the low-pass and high-pass filters associated with the selected mother wavelet. $2\ell$ accounts for the downsampling by a factor of 2 after filtering.

The 3-level DWT was applied through a Daubechies 4 wavelet (Daubechies, 1988). Then, from $a_3^C[n]$, the following mean ($\mu$), standard deviation ($\sigma$), and kurtosis ($\kappa$) values were computed as (Kumar G S et al., 2022):

$$\begin{aligned} \mu^C &= \frac{1}{N} \sum_{\ell=1}^{N} a_3^C[\ell], \\ \sigma^C &= \sqrt{\frac{1}{N-1} \sum_{\ell=1}^{N} \left(a_3^C[\ell] - \mu^C\right)^2}, \\ \kappa^C &= \frac{1}{N} \sum_{\ell=1}^{N} \left(\frac{a_3^C[\ell] - \mu^C}{\sigma^C}\right)^4 \end{aligned} \tag{6.10}$$

where $N$ is the length of $a_3[\ell]$.

The following Shannon entropy ($H$) was additionally computed (Kumar G S et al., 2022):

$$H^C = -\sum_n \mathcal{P}_n \log_2(\mathcal{P}_n), \tag{6.11}$$

where the probabilities $\mathcal{P}_n$ are estimated as relative wavelet energies (Rosso et al., 2001) by means of a $[\log_2(N) - 1]$-level Maximum Overlap DWT (Percival & Walden, 2000, pp. 159–205).

The EEG feature vector (of dimensionality 608) was obtained by concatenating the 5 temporal, 5 spectral, 2 information, and 4 time–frequency features from each of the 38 channels $C$, namely:

$$\boldsymbol{x}_{EEG} = \left(\left\{\bar{x}^C, \tilde{x}^C, \mathcal{A}^C, \mathcal{M}^C, \mathcal{C}^C, \left\{p_B^C\right\}_B, \mathcal{E}^C, \mathcal{F}^C, \mu^C, \sigma^C, \kappa^C, H^C\right\}_C\right)^T \in \mathbb{R}^{608}, \tag{6.12}$$

**PPG**

The PPG signal was first pre-processed for peak identification using the qPPG algorithm (Charlton et al., 2022). From the time points corresponding to the identified peaks $\{t_k\}_{k=1}^M$, the pulse-to-pulse interval tachogram $\{x_k \doteq t_{k+1} - t_k\}_{k=1}^n$ was computed, with $n = M - 1$. Temporal and spectral features were extracted. As temporal features, Mean Absolute Deviation (MAD), Root Mean Square of Successive Differences (RMSSD), and Standard Deviation of Normal-to-Normal intervals (SDNN) of the tachogram $\{x_k\}_{k=1}^n$ were computed as (Malik et al., 1996):

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{k=1}^{n} x_k, \\
MAD &= \frac{1}{n}\sum_{k=1}^{n} |x_k - \bar{x}|, \\
RMSSD &= \sqrt{\frac{1}{n-1}\sum_{k=1}^{n-1}(x_{k+1} - x_k)^2}, \\
SDNN &= \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}.
\end{aligned} \tag{6.13}$$

where $\bar{x}$ is the average pulse-to-pulse interval.

The PSD of the tachogram, $x(f)$, was computed by means of the Lomb–Scargle periodogram (Van Dongen et al., 1999), and the following powers in Low Frequency ($p_{LF}$) and High Frequency

($p_{HF}$), as well as the Low-Frequency-to-High-Frequency power ratio ($p_{LFHF}$), were computed as (Malik et al., 1996):

$$\begin{aligned} p_{LF} &= \int_{0.04}^{0.15} x(f)df \\ p_{HF} &= \int_{0.15}^{0.4} x(f)df \\ p_{LFHF} &= \frac{p_{LF}}{p_{HF}} \end{aligned} \tag{6.14}$$

The PPG feature vector $\boldsymbol{x}_{PPG}$ (of dimensionality 6) was obtained by concatenating the 3 temporal and 3 spectral features, namely:

$$\boldsymbol{x}_{PPG} = (MAD, RMSSD, SDNN, p_{LF}, p_{HF}, p_{LFHF})^T \in \mathbb{R}^6 \tag{6.15}$$

**SC**

The phasic SCR and tonic SCL components were deconvolved from the raw signal $SC(t) \approx SCL(t) + SCR(t)$ through the `cvxEDA` algorithm (Greco et al., 2016).

As time-domain features, the mean value of SCL ($\overline{SCL}$), mean amplitude of SCR peaks ($\overline{pSCR}$), and Area Under the Curve (AUC) of the SC signal were computed as (Greco et al., 2017):

$$\begin{aligned} \overline{SCL} &= \text{Mean}_t \{SCL(t)\} \\ \overline{pSCR} &= \frac{1}{N_P} \sum_{k=1}^{N_P} SCR(t_k) \\ AUC &= \int SC(t)dt \end{aligned} \tag{6.16}$$

where $\text{Mean}_t \{\cdot\}$ is the sample mean operator computed across the time dimension and $\{t_k\}_{k=1}^{N_P}$ are the time points corresponding to SCR peaks.

From Welch's PSD (1-s-long Hamming window, 50% overlap) of the SC signal, $SC(f)$, the following spectral magnitude area (SMA) and spectral powers (P) in the bands $B = \{[0.05, 0.1],$

$[0.1, 0.2], [0.2, 0.3], [0.3, 0.4], [0.4, 0.5]\}$ Hz were computed as (Ghaderyan & Abbasi, 2016):

$$\begin{aligned} SMA &= \int SC(f)df \\ P_B &= \int_B SC(f)df \end{aligned} \tag{6.17}$$

The Mel-frequency Cepstral Coefficients (MFCCs) $c_j[\ell]$ were computed from $SC(t)$ based on a short-time Fourier transform (STFT) with a 1-s Hamming window and 50% overlap (F. Zheng et al., 2001). For the first two MFCCs, indexed by $j \in \{1, 2\}$, the mean value $\mu_j$, standard deviation $\sigma_j$, and kurtosis $\kappa_j$ were computed across time frames as (Ghaderyan & Abbasi, 2016):

$$\begin{aligned} \mu_j &= \frac{1}{L}\sum_{\ell=1}^{L} c_j[\ell], \\ \sigma_j &= \sqrt{\frac{1}{L-1}\sum_{\ell=1}^{L} (c_j[\ell] - \mu_j)^2}, \\ \kappa_j &= \frac{1}{L}\sum_{\ell=1}^{L} \left(\frac{c_j[\ell] - \mu_j}{\sigma_j}\right)^4. \end{aligned} \tag{6.18}$$

where $L$ is the number of STFT frames and $c_j[\ell]$ is the value of the $j$-th MFCC at frame $\ell$.

The SC feature vector (of dimensionality 15) was obtained by concatenating the 3 temporal, 6 spectral, and 6 time–frequency features, namely:

$$\boldsymbol{x}_{SC} = \left(\overline{SCL}, \overline{pSCR}, AUC, SMA, \{p_B\}_B, \{\mu_j, \sigma_j, \kappa_j\}_{j=1}^{2}\right)^T \in \mathbb{R}^{15} \tag{6.19}$$

### EMG

Each EMG channel $x^C(t)$ was first normalised to the $[0, 1]$ range. As temporal features, the following Mean Absolute Value (MAV), Root Mean Square (RMS) (Farfán et al., 2010), and integral value (iEMG) (Penna et al., 2024) were computed:

$$\begin{aligned} MAV^C &= \text{Mean}\left\{\left|x^C(t)\right|\right\} \\ RMS^C &= \sqrt{\text{Mean}\left\{x^C(t)^2\right\}}, \\ iEMG^C &= \int \left|x^C(t)\right| dt \end{aligned} \tag{6.20}$$

where Mean $\{\cdot\}$ is the sample mean operator computed across the time dimension, and $|\cdot|$ is the absolute value operator.

As spectral features, the following powers $p_B$ in bands $B = \{\delta, \theta, \alpha, \beta, \gamma\}$ were computed as (Tun et al., 2021):

$$p_B^C = \int_B x^C(f)df, \tag{6.21}$$

where $x^C(f)$ is Welch's PSD of the EMG channel $C$ computed using a $1s$-long Hamming window and 50% overlap. Bands $B = \{\delta, \theta, \alpha, \beta, \gamma\}$ were defined by means of the IAF, similarly to the EEG.

The EMG feature vector (of dimensionality 16) was obtained by concatenating the 3 temporal and 5 spectral features associated with each channel $C$, namely:

$$\boldsymbol{x}_{EMG} = \left(\left\{MAV^C, RMS^C, iEMG^C, \{p_B^C\}_B\right\}_C\right)^T \in \mathbb{R}^{16} \tag{6.22}$$

**ET**

From the raw gaze signal $(g_x(t), g_y(t))^T$, fixations were identified using the two-step EyeMMV algorithm (Krassanakis et al., 2014) as those points with limited spatial dispersion – bounded by the thresholds $g_{x_{min}}$ and $g_{y_{min}}$ – over a minimum amount of time $t_{min}$ (Blignaut, 2009). Once identified, the fixation position was computed as the mean position $(\overline{g_x}, \overline{g_y})$ of its constituent gaze points, and its duration $T$ was defined as the elapsed time between the first and last gaze points in the fixation. The spatial thresholds were set to $g_{x_{min}} = 10$ pixels and $g_{y_{min}} = 5$ pixels, while the temporal threshold was set to $t_{min} = 80ms$.

In addition to the number of fixations $N$, from the set of detected fixations $\mathcal{F} = \{(\overline{g_{x_k}}, \overline{g_{y_k}}, T_k)\}_{k=1}^N$, the normalised average fixation duration was computed as (Skaramagkas et al., 2021):

$$\overline{T} = \left(\frac{1}{T_S}\right)\sum_{k=1}^{N} T_k, \tag{6.23}$$

where $T_S$ is the stimulus length.

The ET vector (of dimensionality 2) was obtained by concatenating $N$ and $\overline{T}$, namely:

$$\boldsymbol{x}_{ET} = \left(N, \overline{T}\right)^T \in \mathbb{R}^2 \tag{6.24}$$

### 6.2.2 Datasets and Models

Three multimodal datasets, namely $\mathcal{D}_Q$, $\mathcal{D}_V$, and $\mathcal{D}_A$ ($2016 \times 647$ each), were obtained by early fusing the feature tables and appending the quadrant ($Q$), valence ($V$), and arousal ($A$) labels. All the datasets had balanced class proportions both overall and subject-wise (25% for $Q$ and 50% for $V$ and $A$).

For each dataset, a CASH problem including multiple classifiers, with corresponding hyperparameters, and feature-selection methods was solved.

The classifiers included Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Naive Bayes (NB) based on a Parzen window, k-Nearest Neighbours (kNN), Support Vector Machine (SVM), Decision Tree (DT), shallow Artificial Neural Network (ANN) with one hidden layer, and a random subspace Random Forest (RF).

For LDA, hyperparameters included the regularisation parameter $\gamma$ varying within the $[0, 1]$ range. For QDA, $\gamma \in \{0, 1\}$. NB used a Parzen window with smoothing parameter $h \in [10^{-3}, 10^{3}]$ (log scale) and kernel $\kappa$ selected from uniform (box), triangular (triangle), Gaussian (normal), and parabolic (epanechnikov). The kNN considered the number of neighbours $\eta_K \in [1, 50]$ and the distance metric $d$ selected from Euclidean, city-block, cosine, and Minkowski with order $p \in [1, 5]$. The SVM explored linear, Gaussian (RBF), and polynomial (with order $p \in [2, 5]$) kernels $\kappa$ and box constraint $C \in [10^{-3}, 10^{3}]$ (log scale). For DT, the maximum number of splits $N_S$ and minimum leaf size $S_L$ varied, respectively, in the $[2, 200]$ and $[1, 50]$ ranges, while the split criterion $\mathcal{S}$ was selected from Gini index (gdi), deviance, and twoing. For the ANN, the activation function $\sigma$ of the hidden layer varied across rectified linear unit (relu), hyperbolic tangent (tanh), sigmoid, and linear, the regularisation coefficient $\lambda \in [0, 10^{-3}]$ (log scale) and the hidden layer size $N \in [1, 100]$ were considered. Finally, RF was optimised for the number of trees $N_{DT}$ varying in the $[1, 1000]$ range, leaf size $S_L \in [1, 10]$, and number of features per split $N_F \in [1, 10]$.

The feature-selection algorithms (filter types) included Minimum Redundancy Maximum Relevance (MRMR) (C. Ding & Pend, 2005), $\chi^2$ (Rosidin et al., 2021), Neighbourhood Component Analysis (NCA) (Yang et al., 2012), and ReliefF (Robnik-Šikonja & Kononenko, 2003) with the number of neighbours varying in the $[1, 10]$ range. The ranking threshold $t$, expressed as cumulative percentage of the overall feature score, varied in the $[1, 100]$ range. Table 6.1 summarises the classifiers, corresponding hyperparameters, and search spaces.

Table 6.1: Classifiers, hyperparameters, and search spaces for the CASH problem

| **Classifier** | **Hyperparameter** | **Search Space** |
|---|---|---|
| LDA | $\gamma$ | $[0, 1]$ |
| QDA | $\gamma$ | $\{0, 1\}$ |
| NB | $h$ | $[10^{-3}, 10^{3}]$ (log scale) |
| | $K(\cdot)$ | $\{\text{box}, \text{epanechnikov}, \text{normal}, \text{triangle}\}$ |
| kNN | $k$ | $[1, 50]$ |
| | $d$ | $\{\text{euclidean}, \text{cityblock}, \text{cosine}, \text{minkowski}\}$ |
| | $p$ | $[1, 5]$ |
| SVM | $\kappa(\cdot, \cdot)$ | $\{\text{linear}, \text{rbf}, \text{polynomial}\}$ |
| | $C$ | $[10^{-3}, 10^{3}]$ (log scale) |
| | $p$ | $[2, 5]$ |
| DT | $N_s$ | $[2, 200]$ |
| | $S_L$ | $[1, 50]$ |
| | $\mathcal{S}(\cdot)$ | $\{\text{gdi}, \text{deviance}, \text{twoing}\}$ |
| RF | $N_{DT}$ | $[1, 1000]$ |
| | $S_L$ | $[1, 10]$ |
| | $N_F$ | $[1, 10]$ |
| ANN | $\sigma$ | $\{\text{relu}, \text{tanh}, \text{sigmoid}, \text{none}\}$ |
| | $\lambda$ | $[0, 10^{-3}]$ |
| | $N$ | $[1, 100]$ |

LDA: linear discriminant, QDA: quadratic discriminant analysis, NB: Naive Bayes, kNN: k-nearest neighbour, SVM: support vector machine, DT: decision tree, RF: random forest, ANN: artificial neural network.
The description of the hyperparameters is provided in the text.

### 6.2.3 Model Training and Evaluation

We estimated the expected risk $R(\cdot)$ in the CASH problem through $N = 10$ sets of disjoint partitions of the training set $\mathcal{D}_l^{(i)}$, $\mathcal{D}_v^{(i)}$, and $\mathcal{D}_t^{(i)}$, $\forall i \leq N$. More specifically, we computed the following quantity (see Section 4.4):

$$R\left(A_{\boldsymbol{\lambda}}\left(\mathcal{D}\right)\right) = \frac{1}{N}\sum_{i=1}^{N} R_{emp}\left(A_{\boldsymbol{\lambda}}\left(\mathcal{D}_l^{(i)}\right), \mathcal{D}_v^{(i)}\right), \tag{6.25}$$

where $R_{emp}\left(A_{\lambda}\left(\mathcal{D}_l^{(i)}\right), \mathcal{D}_v^{(i)}\right)$ is the empirical risk over $\mathcal{D}_v^{(i)}$ of the model trained on $\mathcal{D}_l^{(i)}$.

The random splits were made subject-wise, with $\mathcal{D}_l^{(i)}$ including 70% (45) of the subjects, while $\mathcal{D}_v^{(i)}$ and $\mathcal{D}_t^{(i)}$ included 15% (9) each. The sets were sampled from $\mathcal{D}$ at each iteration. We did not follow the exhaustive LPSO scheme, as it would grow factorially. In fact, the number $N$ of different arrangements of $P$ elements into sets of size $P_l$, $P_v$, and $P_t$ is given by (Riley et al., 2006, p. 1136):

$$N = \frac{P!}{P_l!P_v!P_t!}, \tag{6.26}$$

which, for the I DARE dataset ($P = 63$), is lower bounded by $N = 3906$, $(P_l = 61, P_v = P_t = 1)$.

To comply with the requirement of independence between learning and validation/testing sets, feature selection was not performed on the whole dataset but was computed on each $\mathcal{D}_l^{(i)}$ and applied to the corresponding $\mathcal{D}_v^{(i)}$ (Lemm et al., 2011).

The CASH problem was solved through a Bayesian optimisation approach with 150 iterations and the 0–1 loss function. The expected loss $\epsilon$ of the best model–hyperparameter configuration $A^*_{\boldsymbol{\lambda}^*}$ was estimated as (see Section 4.4):

$$\epsilon = \frac{1}{N_R}\sum_{i=1}^{N_R} R_{emp}\left(\mathcal{A}^*_{\boldsymbol{\lambda}^*}\left(\mathcal{D}_l^{(i)} \cup \mathcal{D}_v^{(i)}\right), \mathcal{D}_t^{(i)}\right), \tag{6.27}$$

where $R_{emp}\left(\mathcal{A}^*_{\boldsymbol{\lambda}^*}\left(\mathcal{D}_l^{(i)} \cup \mathcal{D}_v^{(i)}\right), \mathcal{D}_t^{(i)}\right)$ is the empirical risk over $\mathcal{D}_t^{(i)}$ of the model trained on $\mathcal{D}_l^{(i)} \cup \mathcal{D}_v^{(i)}$. Confidence intervals (99%) for $\epsilon$ were additionally computed as described in Section 4.4.

As a benchmark, a classifier constantly predicting, for each instance in $\mathcal{D}_t^{(i)}$, the most frequent class in $\mathcal{D}_l^{(i)} \cup \mathcal{D}_v^{(i)}$ was trained, and the corresponding error $\epsilon_{Me}$ with 99% confidence interval ($CI_{99}$) was computed. Non-overlap between the marginal $CI_{99}$ of the selected model and that of the benchmark was used as a conservative descriptive criterion for separation with significance at the $\alpha = 0.01$ level. Conversely, overlap between marginal confidence intervals was not interpreted

Table 6.2: Performance and complexity of the best models split per dataset.

| Dataset | $\epsilon$ | $CI_{99}$ | $A$ | $C$ |
|---|---|---|---|---|
| $\mathcal{D}_Q$ | 0.238 | [0.173, 0.313] | 76.2% | 177966 |
| $\mathcal{D}_V$ | 0.166 | [0.111, 0.234] | 83.4% | 92952 |
| $\mathcal{D}_A$ | 0.147 | [0.095, 0.212] | 85.3% | 26338 |

$\epsilon$: expected loss, $CI_{99}$: confidence intervals for expected loss, $A$: accuracy, $C$: structural complexity.

Table 6.3: Normalised feature-importance magnitude scores for dataset $\mathcal{D}_Q$, split by modality and domain.

| | **T** | **F** | **TF** | **Sum** |
|---|---|---|---|---|
| **EEG** | 41.709 | 35.522 | 22.583 | 99.815 |
| **SC** | 0.030 | 0.053 | 0.047 | 0.130 |
| **PPG** | 0.018 | 0.015 | – | 0.033 |
| **EMG** | 0.014 | 0.008 | – | 0.022 |
| **ET** | 0.001 | – | – | 0.001 |
| **Sum** | 41.771 | 35.599 | 22.630 | |

T: temporal domain; F: spectral domain; TF: time–frequency domain.

as evidence of equivalence or absence of a difference (Greenland et al., 2016).

## 6.3 Results

For $\mathcal{D}_Q$, the best model was a Random Forest (RF) with $N_{DT} = 288$, $S_L = 1$, and $N_F = 2$, using $\chi^2$ feature selection with threshold $t = 99.915\%$. The model achieved an expected loss of 0.238 (accuracy 76.2%) with a 99% confidence interval $CI_{99}$ = [0.173, 0.313], while the benchmark expected loss was 0.750 (accuracy 25%) with $CI_{99}$ = [0.674, 0.816]. Table 6.2 summarises model performance (expected loss, CI, accuracy and structural complexity – described below) split per dataset. Figure 6.1 shows the histogram with $CI_{99}$ error bars of the expected loss for the models (M) and the corresponding benchmarks (B), split per dataset. EEG accounted for 99.815% of the total normalised importance magnitude, followed by SC (0.130%), PPG (0.033%), EMG (0.022%), and ET (0.001%). Across feature domains, temporal features contributed 41.771%, spectral features 35.599%, and time–frequency features 22.630%. Table 6.3 reports the feature-importance magnitude scores split by modality and domain.

For $\mathcal{D}_V$, the best model was an RF with $N_{DT} = 542$, $S_L = 6$, and $N_F = 9$, using ReliefF

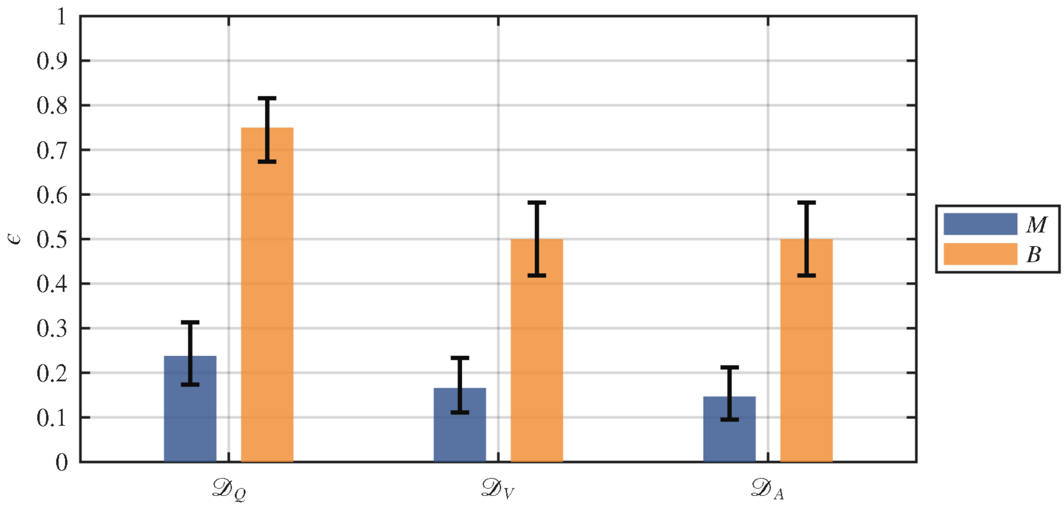


Figure 6.1: Histograms with $CI_{99}$ error bars of the expected loss ($\epsilon$) for the best models (M) and corresponding benchmarks (B), split per dataset.

Table 6.4: Normalised feature-importance magnitude scores for dataset $\mathcal{D}_V$, split by modality and domain.

| | **T** | **F** | **TF** | **Sum** |
|---|---|---|---|---|
| **EEG** | 34.770 | 26.571 | 19.253 | 80.594 |
| **SC** | 1.023 | 2.236 | 2.443 | 5.702 |
| **PPG** | 1.270 | 1.328 | – | 2.597 |
| **EMG** | 3.129 | 6.281 | – | 9.410 |
| **ET** | 1.697 | – | – | 1.697 |
| **Sum** | 41.889 | 36.415 | 21.696 | |

T: temporal domain; F: spectral domain; TF: time–frequency domain.

feature selection with $k = 10$ and threshold $t = 64.385\%$. The model achieved an expected loss of 0.166 (accuracy 83.4%) with a 99% confidence interval $CI_{99} = [0.111, 0.234]$, while the benchmark expected loss was 0.500 (accuracy 50%) with $CI_{99} = [0.418, 0.582]$. Table 6.2 and Figure 6.1 summarise model performance. EEG accounted for the largest share of normalised importance magnitude (80.594%), followed by EMG (9.410%), SC (5.702%), PPG (2.597%), and ET (1.697%). Across feature domains, temporal features contributed 41.889%, spectral features 36.415%, and time–frequency features 21.696%. Table 6.4 reports the feature-importance magnitude scores split by modality and domain.

For $\mathcal{D}_A$, the best model was an RF with $N_{DT} = 119$, $S_L = 3$, and $N_F = 10$, using $\chi^2$ feature selection with threshold $t = 54.047\%$. The model achieved an expected loss of 0.147 (accuracy

Table 6.5: Normalised feature-importance magnitude scores for dataset $\mathcal{D}_A$, split by modality and domain.

| | **T** | **F** | **TF** | **Sum** |
|---|---|---|---|---|
| **EEG** | 42.017 | 35.156 | 22.647 | 99.820 |
| **SC** | 0.022 | 0.040 | 0.038 | 0.099 |
| **PPG** | 0.017 | 0.017 | – | 0.034 |
| **EMG** | 0.021 | 0.025 | – | 0.046 |
| **ET** | 0.001 | – | – | 0.001 |
| **Sum** | 42.078 | 35.237 | 22.685 | |

T: temporal domain; F: spectral domain; TF: time–frequency domain.

85.3%) with a 99% confidence interval $CI_{99} = [0.095, 0.212]$, while the benchmark expected loss was 0.500 (accuracy 50%) with $CI_{99} = [0.418, 0.582]$. Figure 6.1 summarise model performance. EEG accounted for 99.820% of the total normalised importance magnitude, followed by SC (0.099%), EMG (0.046%), PPG (0.034%), and ET (0.001%). Across feature domains, temporal features contributed 42.078%, spectral features 35.237%, and time–frequency features 22.685%. Table 6.5 reports the feature-importance magnitude scores split by modality and domain.

Since all best-performing configurations selected RF classifiers, we further characterised their capacity in relation to their predictive performance. According to previous work that parameterised RF complexity through the number of leaves per tree and the number of trees (Belkin et al., 2019; Curth et al., 2023), we quantified the structural complexity of each trained RF as:

$$C = P_L \cdot P_T, \tag{6.28}$$

where $P_T$ is the number of trees and $P_L$ is the average number of terminal leaves per tree. To compute $C$, the best models were refitted on the corresponding whole datasets.

For $\mathcal{D}_Q$, $\mathcal{D}_V$, and $\mathcal{D}_A$, the RFs consisted, respectively, of $P_T = 288$, 542, and 119 trees, with average numbers of terminal leaves per tree of $P_L = 617.9$, 171.5, and 221.3. They resulted in structural complexities of $C = 177966$, 92952, and 26338, as summarised in Table 6.2.

## 6.4 Discussion

We proposed MATILDE (Multimodal biosignAl-based modaliTy-resilient subject-Independent modeL for affect DEtection), a multimodal affect detection (AD) model based on the standard

I DARE dataset. Features were extracted across temporal, spectral, and time–frequency domains, and three datasets were constructed to predict affective quadrants (4 classes), valence (2 classes), and arousal (2 classes). The Combined Algorithm Selection and Hyperparameter Optimisation (CASH) problem was addressed by evaluating multiple classifiers (LDA, QDA, NB, kNN, SVM, DT, ANN, and RF) and feature-selection methods, solved through Bayesian optimisation over 150 iterations. The best-performing configurations were compared against a majority-class benchmark. Since all selected models were RFs, predictive performance was also described in relation to the structural complexity of the corresponding trained forests.

Across all datasets, RF achieved the highest performance and was, thus, selected, confirming its well-documented superiority against both tML and other ensemble methods (Bamonte et al., 2024; Karim et al., 2025; Rodrigues & Correia, 2023).

All models exceeded their respective benchmarks under the adopted conservative CI-overlap criterion. Descriptively, quadrant classification showed the highest expected loss, followed by valence and arousal. However, since the corresponding $CI_{99}$ intervals overlapped, these differences should be interpreted as descriptive trends rather than as formal pairwise statistical comparisons. The structural complexity analysis supports the interpretation that $\mathcal{D}_Q$ represented the most demanding classification problem, as the corresponding model also achieved the largest complexity. This suggests that quadrant classification required a more articulated partitioning of the feature space, consistent with previous studies that highlighted quadrant classification as more challenging than the individual axes (Alazrai et al., 2018; Galvão et al., 2021; Salankar et al., 2021). In contrast, arousal classification achieved the lowest expected loss with the lowest structural complexity, suggesting a comparatively simpler decision boundary in the selected feature space. This has already been observed with AD models trained on EEG (Piho & Tjahjadi, 2018), peripheral (Gohumpu et al., 2023), and multimodal (A. Sharma & Canavan, 2021) features.

Overall, the obtained accuracies align with those reported in recent literature. Pillalamarri and Shanmugam (2025) found mean accuracies of $84.84 \pm 9.13\%$ for multimodal AD models using traditional classifiers and $88.56 \pm 7.76\%$ for deep networks. In subject-independent settings, Cimtay et al. (2020) reported $74.0 \pm 10.1\%$ for discrete emotions, $57.8 \pm 20.2\%$ for arousal, and $75.3 \pm 3.2\%$ for valence. Later works achieved comparable performance: $75.24 \pm 3.43\%$ for valence and $79.17 \pm 4.99\%$ for arousal (Ma et al., 2025), as well as $83.90 \pm 5.97\%$ across discrete emotions (Fu et al., 2024). The present results, thus, fall within the upper range of traditional tML approaches and are especially competitive for valence and arousal classification.

Feature-importance analyses confirmed the central role of EEG, although its prominence differed across prediction targets. For $\mathcal{D}_Q$ and $\mathcal{D}_A$, EEG accounted for almost all normalised importance magnitude, while Peripheral modalities contributed only marginally In contrast, $\mathcal{D}_V$ showed a more distributed pattern: EEG remained the most informative modality, but peripheral modalities jointly accounted for approximately 1/5 of the total importance magnitude, with EMG and SC providing the largest non-EEG contributions. This suggests that valence classification may benefit more from peripheral physiological information than quadrant or arousal classification. This pattern is compatible with previous findings showing EEG's higher discriminative power over both EMG (J. Chen et al., 2022) and peripheral modalities such as SC and PPG (Torres-Valencia et al., 2016), while also supporting the role of facial muscle activity and autonomic responses in valence-related affective processes (Sato et al., 2020).

The contribution of PPG was small across all datasets, although it was more visible for valence than for quadrant and arousal classification. This may be partly related to the short duration of the stimuli, which can limit the reliability of heart-rate and heart-rate-variability estimates. Studies emphasising heart-rate features, derived from ECG or PPG, over SC (Arthanarisamy Ramaswamy & Palaniswamy, 2022; Gohumpu et al., 2023) typically involved longer stimuli, whereas shorter trials reduce the temporal window available for cardiovascular dynamics. Indeed, for short ($\sim 15s$) stimuli, SC features have been shown to outperform both ECG and PPG (Polo et al., 2024).

Although EMG is traditionally linked to valence (Sato et al., 2020) and has also proven useful for arousal prediction (Tan et al., 2016), its contribution in the present analyses depended strongly on the prediction target. EMG contributed only marginally to quadrant and arousal classification, but it represented the largest non-EEG contribution for valence. This supports the interpretation that valence decoding may be more sensitive to facial muscle activity than the other two prediction targets.

Similarly, ET features showed negligible importance for $\mathcal{D}_Q$ and $\mathcal{D}_A$ and a small contribution for $\mathcal{D}_V$, likely reflecting both their limited predictive strength in the present feature set and the simplicity of the extracted metrics. More complex ET-derived features, including saccades, fixations, blinks, and pupil diameter, have yielded higher accuracies when paired with deep ANNs (Skaramagkas et al., 2023).

Across domains, temporal features contributed most in all three datasets, followed by spectral and time–frequency features. This trend was highly consistent across $\mathcal{D}_Q$, $\mathcal{D}_V$, and $\mathcal{D}_A$, with temporal features accounting for approximately 1/2 of the total normalised importance magnitude. The result confirms previous reports highlighting the importance of temporal features in emotion recognition (Bhattacharyya et al., 2021; Bilucaglia et al., 2021). The success of Convolutional Neural Networks (CNNs) trained on raw spatio-temporal data further supports this observation (Pillalamarri & Shanmugam, 2025).

Despite the promising performance of MATILDE, some limitations should be acknowledged. First, the model was trained and assessed only on I DARE. Although the dataset is standardised and multimodal, it cannot capture the full variability of populations, stimuli, and recording contexts. External validation on independent datasets is therefore necessary to assess the generalisability of the proposed pipeline beyond the present experimental setting (Collins et al., 2014). Second, MATILDE was based on hand-crafted features and feature-level fusion. This choice improves interpretability and is coherent with the available sample size, but it may not fully exploit modality-specific temporal dynamics or cross-modal interactions. Future work could compare this approach with decision-level, hybrid-fusion, or deep-learning architectures (F. Zhao et al., 2024), especially as larger multimodal affective datasets become available. Finally, the model was developed under a subject-independent setting. This design is more relevant for general-purpose AD applications, but it may underestimate the performance achievable with personalised or subject-adaptive models (Arevalillo-Herráez et al., 2019). Future studies should therefore compare subject-independent, subject-dependent, and transfer-learning (W. Li et al., 2022) strategies to clarify the trade-off between generalisability and individual calibration.

# Chapter 7

# MATILDE – Robustness Assessment

This chapter evaluates the robustness of MATILDE through dataset-level ablation studies. Unimodal and Leave-One-Modality-Out (LOMO) configurations are tested using two strategies: a fixed-pipeline setting, where the full-model configuration is preserved, and a re-CASH setting, where model selection and hyperparameter optimisation are repeated for each ablated dataset. Most configurations remain above the benchmark and close to the full models under the adopted descriptive criterion, while ET shows limited standalone and incremental contribution. Re-CASH complexity analyses show non-systematic changes in RF structural complexity, with larger effects for arousal. Overall, the results suggest that the proposed pipeline can adapt to reduced-modality configurations, supporting scalable affect-detection implementations.

## 7.1 Introduction

*Robustness* is one aspect of model testing, alongside the already discussed *correctness* (see Section 6.1.4). It refers to the resilience of a system towards perturbations of its components, which include, among others, the training data and the learning algorithm. It is generally assessed by creating, from the base model $f$, a perturbed version $\tilde{f}$ and measuring the difference in correctness (i.e., expected risk) $\Delta R = R(f) - R(\tilde{f})$ (J. M. Zhang et al., 2022).

Robustness of ML/DL models is commonly addressed through so-called *ablation studies*. The term *ablation* (from Latin *ablatio*, literally meaning *a carrying away*) refers to the removal or destruction of part of a biological system in order to study its function. This loss-of-function logic has a long tradition in experimental physiology – mostly in relation to the brain: nineteenth-century anatomists, pioneered by Jean Pierre Flourens, used to investigate the function of neural structures from the effects of their removal or destruction (Gamboa, 2020).

In the context of ML/DL, ablation studies were proposed in the late 1970s as a method of performance analysis for speech-recognition systems (Newell, 1975). A formal definition of ablation was given by Sheikholeslami et al. (2021). Given a training set $\mathcal{D}$ and a model $f$, an ablation study evaluates the relative contribution of different *configurations* $C$ by training variants of the model or dataset in which selected elements are removed. A configuration can refer either to the dataset or to the model. For a dataset $\mathcal{D}$ with $n$ features, a dataset configuration can be defined as $C_{\mathcal{D}}(\mathcal{D}, \{x_i\})$, where the feature subset $\{x_i\}$ is excluded during training. Similarly, for a model $f$ composed of $k$ components, a model configuration can be written as $C_f(f, \{m_j\})$, where the components $\{m_j\}$, such as layers, neurons, or filters, are removed from the architecture. An ablation study is therefore a set of configurations, either $S_{\mathcal{D}} = \{C^1_{\mathcal{D}}, C^2_{\mathcal{D}}, \ldots, C^z_{\mathcal{D}}\}$ or $S_f = \{C^1_f, C^2_f, \ldots, C^z_f\}$, whose execution defines an *experiment* composed of several *trials*, with each trial corresponding to one configuration.

Ablations over $C_f$ originally trace back to the problem of complexity reduction or *pruning*. Early examples included Skeletonization (Mozer & Smolensky, 1988) and Optimal Brain Damage (LeCun et al., 1989), which addressed the problem of removing (i.e., zeroing) a subset of weights of an ANN, while minimising performance degradation. Since their original formulations, DTs have used ablations to replace selected subtrees with leaf nodes, while minimising the loss in predictive performance. Early examples include Cost-Complexity Pruning (Breiman et al., 1984) and Reduced-Error Pruning (Quinlan, 1987). Recently, ablations over $C_f$ can be observed in architectural studies in DL, where the ablated components are no longer only individual weights, but layers, blocks, connections, or modules. Representative examples include layer-wise ablations (Zeiler & Fergus, 2014) and network depth control in deep CNNs (Simonyan & Zisserman, 2015). Analogous perspectives can also be found in modern tree-ensemble methods, where ablations are used to assess the contribution of algorithmic components, such as regularisation and sparsity-aware learning in Extreme Gradient Boosting (XGBoost) (T. Chen & Guestrin, 2016) and gradient-based sampling in Light Gradient Boosting Machine (LightGBM) (Ke et al., 2017).

Ablations over $C_{\mathcal{D}}$ have historical roots in feature selection, with the aim of identifying a subset $\{x_j\}_{j=1}^{M<N}$ of the original feature vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_M)^T$ that preserves or improves the predictive performance of a model. Early examples include informative receptors in recogni-

tion systems (Marill & Green, 1963), branch-and-bound methods (Narendra & Fukunaga, 1977), and wrapper algorithms (Kohavi & John, 1997). Recently, pure-ablation (Lei et al., 2018) and ablation-like (Lundberg & Lee, 2017; Ribeiro et al., 2016) approaches (e.g., perturbation- and permutation-based) over $C_{\mathcal{D}}$ have been employed for model-agnostic approaches to model explainability and interpretability. Finally, with the recent development of multimodal models and the emerging research direction of incomplete multimodal learning (Zhan et al., 2025), ablations over $C_{\mathcal{D}}$ have been expanded from single features $x_i$ to entire groups of homologous signals or channels.

Ablation studies are now considered a best practice, especially when assessing novel architectures based on multiple interacting components (Fostiropoulos et al., 2023). Without such analyses, empirical gains may be incorrectly attributed to unisolated architectural choices, hyperparameter tuning, or experimental setup (Lipton & Steinhardt, 2019).

### 7.1.1 Ablation Studies in Affect Detection

In physiological-signal-based AD, ablations over $C_f$ largely follow the same rationale as in other domains: they are typically performed on layers, modules, or architectural blocks of DL models (Du et al., 2024; Jia et al., 2024; X. Li et al., 2025; Z. Wang & Wang, 2025). Since they concern the contribution of model components rather than the structure of input data, they are not considered further here.

The present discussion instead focuses on ablations over $C_{\mathcal{D}}$, which are more specific to multimodal learning (Baltrusaitis et al., 2019). Dataset-level ablations are usually not conducted over individual features, but over entire modalities. The full multimodal model is therefore compared either against unimodal baselines, where each modality is considered in isolation, or against $N-1$-modal variants, where one modality is removed at a time. This latter strategy is referred to as Leave-One-Modality-Out (LOMO) ablation (MacRae et al., 2026).

Ablations over $C_{\mathcal{D}}$ are useful for estimating the contribution of each modality and for determining the effectiveness of multimodal fusion. In the literature, reported results are highly heterogeneous: tML approaches often show limited or inconsistent benefits from fusion, whereas recent DL models tend to report larger improvements. The following section reviews representative examples of $C_{\mathcal{D}}$-level ablations in both tML and DL settings.

In the tML context, Koelstra et al. (2012) trained Gaussian Naïve Bayes classifiers using EEG, peripheral, and multimedia content features. Fusion produced only modest gains in some cases: EEG plus multimedia improved arousal by 0.013 F1-score points, while multimedia plus peripheral physiology improved valence by 0.044. Soleymani et al. (2012) trained RBF-SVM classifiers with features extracted from EEG and ET data. Decision-level fusion of EEG and ET improved performance over the best unimodal setting by 0.02 F1-score points for arousal and 0.06 for valence. Miranda-Correa et al. (2021) trained Gaussian Naïve Bayes and decision-level SVM with EEG, ECG, and SC features. In this case, fusion of EEG, ECG, and SC did not improve over EEG alone, instead decreasing performance by 0.004 F1-score points for valence and 0.013 points for arousal. Finally, Gohumpu et al. (2023) trained SVM, KNN, and DTs with PPG/HRV, SC, and skin-temperature features. Their ablation analysis showed only small gains from multimodal fusion: about 0.5% for arousal, 0.3%, 0.3% dominance, and no improvement for liking.

In the DL context, Z. Wang and Wang (2025) trained a novel architecture composed of one-dimensional convolutional layers, attention, and gated recurrent units, using EEG and ECG data. Ablation analysis showed that the multimodal model improved over EEG alone by 9.16% on average, and over ECG alone by 13.36%. Du et al. (2024) proposed a signal-channel attention network using EEG, ECG, SC, and ET data. The complete four-modality model improved over SC alone by 5.69%, over SC-ECG by 10.03%, and over SC-EEG-ECG by 2.01%. X. Li et al. (2025) proposed a gated multi-head cross-modal attention network based on EEG, EOG, and SC data. Ablation analysis showed improvements of 7.68% over EEG alone, 13.99% over EOG alone, and 17.94% over SC alone. Jia et al. (2024) proposed a multi-level disentangling network for cross-subject emotion recognition, using EEG, EOG, and EMG on DEAP, and EEG, ECG, and SC. Ablation analysis showed that removing individual modalities reduced the overall performance, with EEG removal producing the largest drop, approximately 2.6–5.3% for arousal and 3.2–4.3% for valence.

## 7.2 Materials and Methods

We performed dataset-level ablation studies on the full models designed in Chapter 6. These models included EEG, PPG, SC, EMG, and ET modalities and were trained on the valence ($\mathcal{D}_V$), arousal ($\mathcal{D}_A$), and quadrant ($\mathcal{D}_Q$) datasets. The expected loss ($\epsilon$) with corresponding 99% confidence intervals ($CI_{99}$) and accuracy ($A$) values of the models were summarised in Table 6.2.

For the ablation studies, we followed two complementary evaluation strategies. First, we performed a *fixed-pipeline* ablation, in which each ablated model inherited the classifier, hyperparameters, and feature-selection method of the corresponding best-performing full model, previously identified as described in Section 6.2.3. Second, we performed a re-optimised ablation analysis (*re-CASH*), in which each ablated dataset was treated as an independent modelling problem. The same procedure adopted for the full models was repeated. Specifically, we solved the CASH problem through Bayesian optimisation, using 150 iterations and the 0–1 loss function. The search space included the classifiers, feature-selection methods, and hyperparameter configurations described in Table 6.1.

For both ablation strategies, the dataset splitting followed the same repeated and resampled Leave-$P$-Subject-Out (LPSO) scheme used for the full models. The split was repeated $N = 10$ times, with learning $\left(\mathcal{D}_l^{(i)}\right)$, validation $\left(\mathcal{D}_v^{(i)}\right)$, and test $\left(\mathcal{D}_t^{(i)}\right)$ subsets corresponding to 70%, 15%, and 15% of the data, kept fixed across all ablation studies and trials. For the re-CASH ablation studies, $\mathcal{D}_l^{(i)}$ was used for model fitting, $\mathcal{D}_v^{(i)}$ for model selection, and $\mathcal{D}_t^{(i)}$ for error estimation. Conversely, in the fixed-pipeline ablation studies, no validation step was required, since the best configurations were inherited from the corresponding full models. Therefore, model training was performed on $\mathcal{D}_l^{(i)} \cup \mathcal{D}_v^{(i)}$ and error estimation on $\mathcal{D}_t^{(i)}$.

Performance of the ablated models was expressed as expected loss ($\epsilon$) with 99% confidence interval ($CI_{99}$) and corresponding accuracy ($A$), similarly to Chapter 6. Additionally, variations in expected loss ($\Delta_\epsilon$) from both the full ($F$) and baseline ($B$) models were reported, together with descriptive CI-overlap markers. Negative $\Delta_\epsilon$ values indicate improved performance of the ablated model with respect to the reference model. For the re-CASH ablations, we additionally quantified the structural complexity $C$ of the RF models selected after re-optimisation, as described in Section 6.2.3. For each RF-based ablated model, we also computed a relative complexity coefficient:

$$\rho_C = \frac{C}{C_{\text{full}}}, \tag{7.1}$$

where $C_{\text{full}}$ is the structural complexity of the corresponding full multimodal model trained on the same target dataset. Values $\rho_C > 1$ indicate that the ablated model selected a structurally more complex RF than the full model, whereas values $\rho_C < 1$ indicate a structurally simpler RF. Since $C$ is defined only for RF classifiers, configurations in which re-CASH selected a different classifier were not assigned a complexity value.

Fixed-pipeline and re-CASH experiments involved unimodal ($S^{U}_{\mathcal{D}_{Q,V,A}}$) and LOMO ($S^{\text{LOMO}}_{\mathcal{D}_{Q,V,A}}$)

ablation studies. They consisted of the following configurations:

$$\begin{aligned} S^{U}_{\mathcal{D}_{Q,V,A}} &= \left\{ C^{\mathrm{EEG}}_{\mathcal{D}_{Q,V,A}}, C^{\mathrm{SC}}_{\mathcal{D}_{Q,V,A}}, C^{\mathrm{PPG}}_{\mathcal{D}_{Q,V,A}}, C^{\mathrm{EMG}}_{\mathcal{D}_{Q,V,A}}, C^{\mathrm{ET}}_{\mathcal{D}_{Q,V,A}} \right\} \\ S^{\mathrm{LOMO}}_{\mathcal{D}_{Q,V,A}} &= \left\{ C^{\overline{\mathrm{EEG}}}_{\mathcal{D}_{Q,V,A}}, C^{\overline{\mathrm{SC}}}_{\mathcal{D}_{Q,V,A}}, C^{\overline{\mathrm{PPG}}}_{\mathcal{D}_{Q,V,A}}, C^{\overline{\mathrm{EMG}}}_{\mathcal{D}_{Q,V,A}}, C^{\overline{\mathrm{ET}}}_{\mathcal{D}_{Q,V,A}} \right\} \end{aligned} \tag{7.2}$$

where $C^{\overline{X}}_{\mathcal{D}_{Q,V,A}} \doteq \mathcal{D}_{Q,V,A} \setminus C^{X}_{\mathcal{D}_{Q,V,A}}$, with $\setminus$ being the difference set operator.

Computations were performed in MATLAB (The MathWorks, Inc.) using the Signal Processing and Statistics & Machine Learning Toolboxes. They were run on a workstation equipped with an AMD Ryzen™ Threadripper PRO 5975WX CPU (32 cores, 64 threads, 3.6 GHz base clock) and 256 GB DDR4-3200 MHz ECC RAM. No GPU acceleration was used; however, Bayesian optimisation was executed in parallel using 32 workers through the Parallel Computing Toolbox.

## 7.3 Results

### 7.3.1 Fixed-Pipeline Ablation

For each dataset, EEG-, SC-, PPG-, and EMG-unimodal models performed better than the corresponding benchmark under, while no detectable degradation relative to the full model emerged. In contrast, the ET-unimodal model did not show clear separation from the benchmark and showed clear degradation relative to the full models under the same criterion. The increases in expected loss were $\mathcal{D}_Q : \Delta\epsilon_F = +0.493$, $\mathcal{D}_V : \Delta\epsilon_F = +0.327$, and $\mathcal{D}_A : \Delta\epsilon_F = +0.345$, corresponding to accuracy decreases of 49.3%, 32.7%, and 34.5%, respectively. The results of fixed-pipeline unimodal ablation studies, split per dataset and configuration, are summarised in Table 7.1.

For each dataset, all LOMO models outperformed their corresponding benchmarks. However, none showed detectable degradation relative to the corresponding full models. The results of fixed-pipeline LOMO ablation studies, split per dataset and configuration, are summarised in Table 7.2.

The expected loss with 99% confidence intervals of the fixed-pipeline ablations, alongside the corresponding benchmarks and full models for $\mathcal{D}_Q$, $\mathcal{D}_V$, and $\mathcal{D}_A$, are plotted in Figures 7.1, 7.3, and 7.2, respectively.

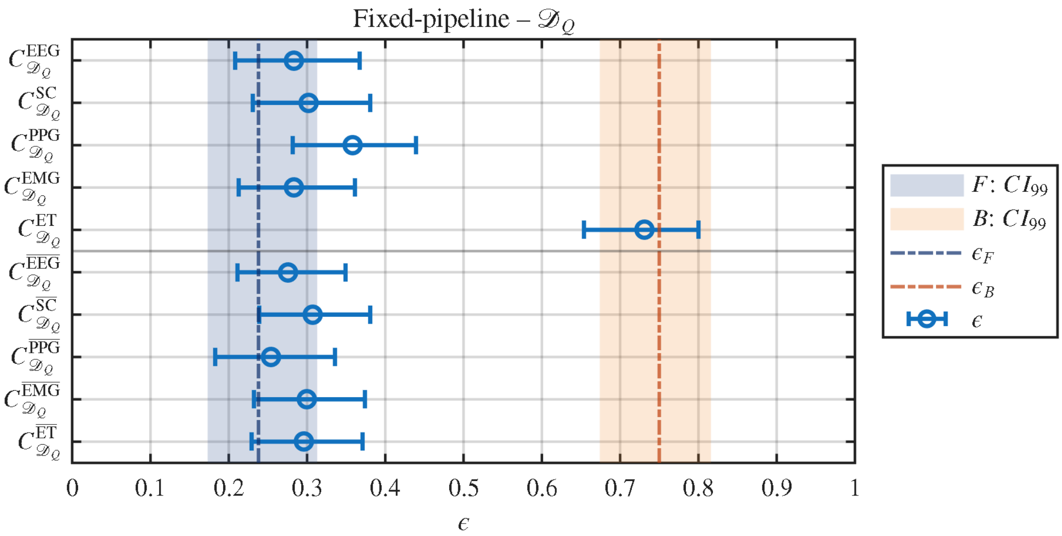


Figure 7.1: Expected loss of fixed-pipeline ablated models for $\mathcal{D}_Q$. Points and horizontal error bars indicate $\epsilon$ and $CI_{99}$, respectively. Shaded bands and vertical lines represent the $CI_{99}$ and expected loss of the full ($F$) and benchmark ($B$) models.

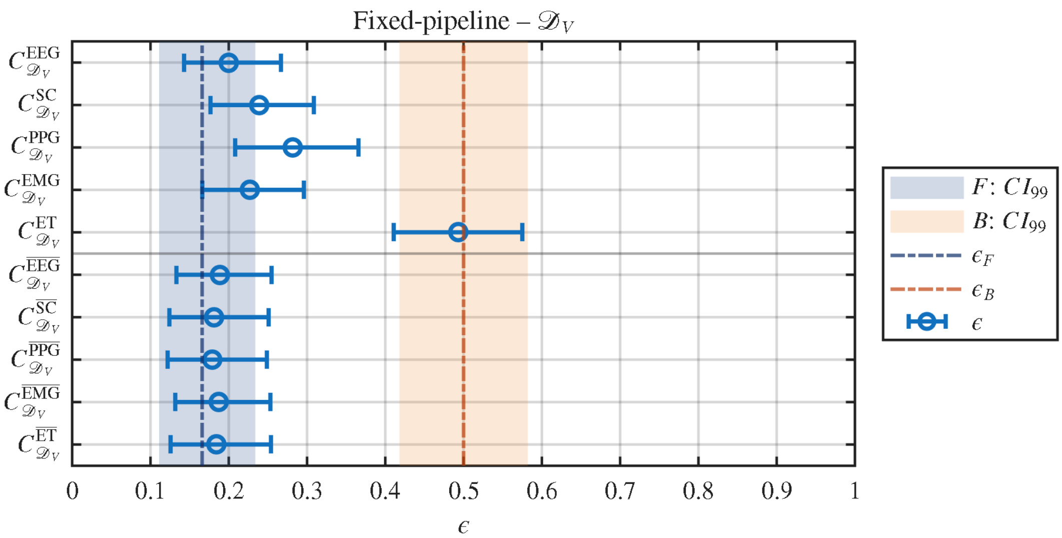


Figure 7.2: Expected loss of fixed-pipeline ablated models for $\mathcal{D}_V$. Points and horizontal error bars indicate $\epsilon$ and $CI_{99}$, respectively. Shaded bands and vertical lines represent the $CI_{99}$ and expected loss of the full ($F$) and benchmark ($B$) models.

Table 7.1: Results of fixed-pipeline unimodal ablation studies split per dataset and model.

| **Dataset** | **Model** | $\boldsymbol{\epsilon}$ | $\boldsymbol{CI_{99}}$ | $\boldsymbol{A}$ | $\boldsymbol{\Delta\epsilon_{\mathrm{B}}}$ | $\boldsymbol{\Delta\epsilon_{\mathrm{F}}}$ |
|---|---|---|---|---|---|---|
| $\mathcal{D}_Q$ | $C^{\mathrm{EEG}}_{\mathcal{D}_Q}$ | 0.283 | [0.208, 0.367] | 71.7% | $-0.467^*$ | +0.045 |
| | $C^{\mathrm{SC}}_{\mathcal{D}_Q}$ | 0.302 | [0.231, 0.381] | 69.8% | $-0.448^*$ | +0.064 |
| | $C^{\mathrm{PPG}}_{\mathcal{D}_Q}$ | 0.358 | [0.282, 0.439] | 64.2% | $-0.392^*$ | +0.120 |
| | $C^{\mathrm{EMG}}_{\mathcal{D}_Q}$ | 0.283 | [0.213, 0.361] | 71.7% | $-0.467^*$ | +0.045 |
| | $C^{\mathrm{ET}}_{\mathcal{D}_Q}$ | 0.731 | [0.654, 0.800] | 26.9% | $-0.019$ | $+0.493^*$ |
| $\mathcal{D}_V$ | $C^{\mathrm{EEG}}_{\mathcal{D}_V}$ | 0.200 | [0.143, 0.267] | 80.0% | $-0.300^*$ | +0.034 |
| | $C^{\mathrm{SC}}_{\mathcal{D}_V}$ | 0.239 | [0.177, 0.309] | 76.1% | $-0.261^*$ | +0.073 |
| | $C^{\mathrm{PPG}}_{\mathcal{D}_V}$ | 0.282 | [0.208, 0.366] | 71.8% | $-0.218^*$ | +0.116 |
| | $C^{\mathrm{EMG}}_{\mathcal{D}_V}$ | 0.227 | [0.166, 0.296] | 77.3% | $-0.273^*$ | +0.061 |
| | $C^{\mathrm{ET}}_{\mathcal{D}_V}$ | 0.493 | [0.411, 0.575] | 50.7% | $-0.007$ | $+0.327^*$ |
| $\mathcal{D}_A$ | $C^{\mathrm{EEG}}_{\mathcal{D}_A}$ | 0.184 | [0.129, 0.250] | 81.6% | $-0.316^*$ | +0.037 |
| | $C^{\mathrm{SC}}_{\mathcal{D}_A}$ | 0.232 | [0.171, 0.302] | 76.8% | $-0.268^*$ | +0.085 |
| | $C^{\mathrm{PPG}}_{\mathcal{D}_A}$ | 0.261 | [0.193, 0.338] | 73.9% | $-0.239^*$ | +0.114 |
| | $C^{\mathrm{EMG}}_{\mathcal{D}_A}$ | 0.153 | [0.102, 0.215] | 84.7% | $-0.347^*$ | +0.006 |
| | $C^{\mathrm{ET}}_{\mathcal{D}_A}$ | 0.492 | [0.411, 0.574] | 50.8% | $-0.008$ | $+0.345^*$ |

$\epsilon$: expected loss; $CI_{99}$: 99% confidence interval for $\epsilon$; $A$: accuracy; $\Delta\epsilon_{\mathrm{B}}$: loss variation with respect to the benchmark; $\Delta\epsilon_{\mathrm{F}}$: loss variation with respect to the full model.
$^*$: non-overlap between marginal $CI_{99}$ intervals.

### 7.3.2 Re-CASH Ablation

For each dataset, EEG-, SC-, PPG-, and EMG-unimodal models performed better than the corresponding benchmark, while no detectable degradation relative to the full model emerged. In contrast, the ET-unimodal model did not show clear separation from the benchmark and showed clear degradation relative to the full models. The increases in expected loss were $\mathcal{D}_Q$ : $\Delta\epsilon_F = +0.496$, $\mathcal{D}_V$ : $\Delta\epsilon_F = +0.320$, and $\mathcal{D}_A$ : $\Delta\epsilon_F = +0.344$, corresponding to accuracy decreases of 49.6%, 32.0%, and 34.4%, respectively.

RF classifiers were selected in 13 out of 15 configurations. The two non-RF configurations were $C^{\mathrm{EEG}}_{\mathcal{D}_Q}$ and $C^{\mathrm{ET}}_{\mathcal{D}_A}$, for which ANN classifiers were selected and $C$ was therefore not computed. Across RF-based unimodal ablations, the relative complexity coefficient was $\rho_C = 2.831 \pm 3.546$, with range [0.112, 10.596]. Seven configurations showed $\rho_C > 1$, whereas six showed $\rho_C < 1$.

The results of the re-CASH unimodal ablation studies, split per dataset and configuration, are summarised in Table 7.3.

For each dataset, all LOMO models outperformed their corresponding benchmarks. However,

Table 7.2: Results of fixed-pipeline LOMO ablation studies split per dataset and model.

| Dataset | Model | $\epsilon$ | $CI_{99}$ | $A$ | $\Delta\epsilon_{\mathrm{B}}$ | $\Delta\epsilon_{\mathrm{F}}$ |
|---|---|---|---|---|---|---|
| $\mathcal{D}_Q$ | $C_{\mathcal{D}_Q}^{\overline{\mathrm{EEG}}}$ | 0.276 | [0.211, 0.349] | 72.4% | $-0.474^*$ | +0.038 |
| | $C_{\mathcal{D}_Q}^{\overline{\mathrm{SC}}}$ | 0.307 | [0.239, 0.381] | 69.3% | $-0.443^*$ | +0.069 |
| | $C_{\mathcal{D}_Q}^{\overline{\mathrm{PPG}}}$ | 0.254 | [0.183, 0.336] | 74.6% | $-0.496^*$ | +0.016 |
| | $C_{\mathcal{D}_Q}^{\overline{\mathrm{EMG}}}$ | 0.300 | [0.232, 0.374] | 70.0% | $-0.450^*$ | +0.062 |
| | $C_{\mathcal{D}_Q}^{\overline{\mathrm{ET}}}$ | 0.296 | [0.229, 0.371] | 70.4% | $-0.454^*$ | +0.058 |
| $\mathcal{D}_V$ | $C_{\mathcal{D}_V}^{\overline{\mathrm{EEG}}}$ | 0.189 | [0.133, 0.255] | 81.1% | $-0.311^*$ | +0.023 |
| | $C_{\mathcal{D}_V}^{\overline{\mathrm{SC}}}$ | 0.181 | [0.124, 0.251] | 81.9% | $-0.319^*$ | +0.015 |
| | $C_{\mathcal{D}_V}^{\overline{\mathrm{PPG}}}$ | 0.179 | [0.122, 0.249] | 82.1% | $-0.321^*$ | +0.013 |
| | $C_{\mathcal{D}_V}^{\overline{\mathrm{EMG}}}$ | 0.187 | [0.132, 0.253] | 81.3% | $-0.313^*$ | +0.021 |
| | $C_{\mathcal{D}_V}^{\overline{\mathrm{ET}}}$ | 0.184 | [0.126, 0.254] | 81.6% | $-0.316^*$ | +0.018 |
| $\mathcal{D}_A$ | $C_{\mathcal{D}_A}^{\overline{\mathrm{EEG}}}$ | 0.219 | [0.159, 0.288] | 78.1% | $-0.281^*$ | +0.072 |
| | $C_{\mathcal{D}_A}^{\overline{\mathrm{SC}}}$ | 0.185 | [0.130, 0.250] | 81.5% | $-0.315^*$ | +0.038 |
| | $C_{\mathcal{D}_A}^{\overline{\mathrm{PPG}}}$ | 0.190 | [0.134, 0.256] | 81.0% | $-0.310^*$ | +0.043 |
| | $C_{\mathcal{D}_A}^{\overline{\mathrm{EMG}}}$ | 0.158 | [0.104, 0.225] | 84.2% | $-0.342^*$ | +0.011 |
| | $C_{\mathcal{D}_A}^{\overline{\mathrm{ET}}}$ | 0.145 | [0.093, 0.210] | 85.5% | $-0.355^*$ | $-0.002$ |

$\epsilon$: expected loss; $CI_{99}$: 99% confidence interval for $\epsilon$; $A$: accuracy; $\Delta\epsilon_{\mathrm{B}}$: loss variation with respect to the benchmark; $\Delta\epsilon_{\mathrm{F}}$: loss variation with respect to the full model.
$^*$: non-overlap between marginal $CI_{99}$ intervals.

none showed detectable degradation relative to the corresponding full models. The results of re-CASH LOMO ablation studies, split per dataset and configuration, are summarised in Table 7.4.

RF classifiers were selected in all 15 configurations. Across these models, the relative complexity coefficient was $\rho_C = 1.974 \pm 2.103$, with range [0.161, 6.643]. Seven configurations showed $\rho_C > 1$, whereas eight showed $\rho_C < 1$.

The expected loss with 99% confidence intervals of the re-CASH ablations, alongside the corresponding benchmarks and full models for $\mathcal{D}_Q$, $\mathcal{D}_V$, and $\mathcal{D}_A$, are plotted in Figures 7.4, 7.5, and 7.6, respectively.

## 7.4 Discussion

We assessed the robustness of MATILDE – developed in Chapter 6 – through dataset-level ablation studies, considering both unimodal and Leave-One-Modality-Out (LOMO) configurations. Two complementary strategies were adopted. In the fixed-pipeline ablation, the classifier, hy-

Table 7.3: Results of re-CASH unimodal ablation studies split per dataset and model.

| **Dataset** | **Model** | $\boldsymbol{\epsilon}$ | $\boldsymbol{CI_{99}}$ | $\boldsymbol{A}$ | $\boldsymbol{\Delta\epsilon_{B}}$ | $\boldsymbol{\Delta\epsilon_{F}}$ | $\boldsymbol{C}$ | $\boldsymbol{\rho_C}$ |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{D}_Q$ | $C^{\text{EEG}}_{\mathcal{D}_Q}$ | 0.249 | [0.183, 0.325] | 75.1% | $-0.501^*$ | $+0.011$ | – | – |
| | $C^{\text{SC}}_{\mathcal{D}_Q}$ | 0.263 | [0.195, 0.340] | 73.7% | $-0.487^*$ | $+0.025$ | 44,704 | 0.251 |
| | $C^{\text{PPG}}_{\mathcal{D}_Q}$ | 0.291 | [0.221, 0.370] | 70.9% | $-0.459^*$ | $+0.053$ | 136,741 | 0.768 |
| | $C^{\text{EMG}}_{\mathcal{D}_Q}$ | 0.254 | [0.187, 0.330] | 74.6% | $-0.496^*$ | $+0.016$ | 291,427 | 1.638 |
| | $C^{\text{ET}}_{\mathcal{D}_Q}$ | 0.734 | [0.657, 0.803] | 26.6% | $-0.016$ | $+0.496^*$ | 27,310 | 0.154 |
| $\mathcal{D}_V$ | $C^{\text{EEG}}_{\mathcal{D}_V}$ | 0.159 | [0.105, 0.226] | 84.1% | $-0.341^*$ | $-0.007$ | 152,847 | 1.644 |
| | $C^{\text{SC}}_{\mathcal{D}_V}$ | 0.169 | [0.113, 0.237] | 83.1% | $-0.331^*$ | $+0.003$ | 313,151 | 3.369 |
| | $C^{\text{PPG}}_{\mathcal{D}_V}$ | 0.202 | [0.142, 0.274] | 79.8% | $-0.298^*$ | $+0.036$ | 10,377 | 0.112 |
| | $C^{\text{EMG}}_{\mathcal{D}_V}$ | 0.155 | [0.102, 0.222] | 84.5% | $-0.345^*$ | $-0.011$ | 249,605 | 2.685 |
| | $C^{\text{ET}}_{\mathcal{D}_V}$ | 0.486 | [0.405, 0.568] | 51.4% | $-0.014$ | $+0.320^*$ | 21,566 | 0.232 |
| $\mathcal{D}_A$ | $C^{\text{EEG}}_{\mathcal{D}_A}$ | 0.163 | [0.108, 0.230] | 83.7% | $-0.337^*$ | $+0.016$ | 142,962 | 5.428 |
| | $C^{\text{SC}}_{\mathcal{D}_A}$ | 0.158 | [0.104, 0.225] | 84.2% | $-0.342^*$ | $+0.011$ | 279,070 | 10.596 |
| | $C^{\text{PPG}}_{\mathcal{D}_A}$ | 0.198 | [0.138, 0.270] | 80.2% | $-0.302^*$ | $+0.051$ | 13,977 | 0.531 |
| | $C^{\text{EMG}}_{\mathcal{D}_A}$ | 0.165 | [0.110, 0.232] | 83.5% | $-0.335^*$ | $+0.018$ | 247,627 | 9.402 |
| | $C^{\text{ET}}_{\mathcal{D}_A}$ | 0.491 | [0.409, 0.573] | 50.9% | $-0.009$ | $+0.344^*$ | – | – |

$\epsilon$: expected loss; $CI_{99}$: 99% confidence interval for $\epsilon$; $A$: accuracy; $\Delta\epsilon_{\text{B}}$: loss variation with respect to the benchmark; $\Delta\epsilon_{\text{F}}$: loss variation with respect to the full model; $C$: RF structural complexity; $\rho_C$: relative RF complexity.
–: non-RF model.
$^*$: non-overlap between marginal $CI_{99}$ intervals.

Table 7.4: Results of re-CASH LOMO ablation studies split per dataset and model.

| Dataset | Model | $\epsilon$ | $CI_{99}$ | $A$ | $\Delta\epsilon_{\mathrm{B}}$ | $\Delta\epsilon_{\mathrm{F}}$ | $C$ | $\rho_C$ |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{D}_Q$ | $C_{\mathcal{D}_Q}^{\overline{\mathrm{EEG}}}$ | 0.241 | [0.175, 0.316] | 75.9% | $-0.509^*$ | +0.003 | 169,613 | 0.953 |
| | $C_{\mathcal{D}_Q}^{\overline{\mathrm{SC}}}$ | 0.244 | [0.178, 0.319] | 75.6% | $-0.506^*$ | +0.006 | 28,659 | 0.161 |
| | $C_{\mathcal{D}_Q}^{\overline{\mathrm{PPG}}}$ | 0.230 | [0.166, 0.304] | 77.0% | $-0.520^*$ | −0.008 | 470,005 | 2.641 |
| | $C_{\mathcal{D}_Q}^{\overline{\mathrm{EMG}}}$ | 0.236 | [0.172, 0.311] | 76.4% | $-0.514^*$ | −0.002 | 56,978 | 0.320 |
| | $C_{\mathcal{D}_Q}^{\overline{\mathrm{ET}}}$ | 0.232 | [0.168, 0.307] | 76.8% | $-0.518^*$ | −0.006 | 145,032 | 0.815 |
| $\mathcal{D}_V$ | $C_{\mathcal{D}_V}^{\overline{\mathrm{EEG}}}$ | 0.165 | [0.110, 0.233] | 83.5% | $-0.335^*$ | −0.001 | 79,124 | 0.851 |
| | $C_{\mathcal{D}_V}^{\overline{\mathrm{SC}}}$ | 0.161 | [0.107, 0.228] | 83.9% | $-0.339^*$ | −0.005 | 54,117 | 0.582 |
| | $C_{\mathcal{D}_V}^{\overline{\mathrm{PPG}}}$ | 0.161 | [0.107, 0.228] | 83.9% | $-0.339^*$ | −0.005 | 78,306 | 0.842 |
| | $C_{\mathcal{D}_V}^{\overline{\mathrm{EMG}}}$ | 0.162 | [0.107, 0.229] | 83.8% | $-0.338^*$ | −0.004 | 105,432 | 1.134 |
| | $C_{\mathcal{D}_V}^{\overline{\mathrm{ET}}}$ | 0.158 | [0.104, 0.225] | 84.2% | $-0.342^*$ | −0.008 | 14,929 | 0.161 |
| $\mathcal{D}_A$ | $C_{\mathcal{D}_A}^{\overline{\mathrm{EEG}}}$ | 0.148 | [0.096, 0.213] | 85.2% | $-0.352^*$ | +0.001 | 33,512 | 1.272 |
| | $C_{\mathcal{D}_A}^{\overline{\mathrm{SC}}}$ | 0.147 | [0.095, 0.213] | 85.3% | $-0.353^*$ | +0.000 | 174,972 | 6.643 |
| | $C_{\mathcal{D}_A}^{\overline{\mathrm{PPG}}}$ | 0.151 | [0.098, 0.217] | 84.9% | $-0.349^*$ | +0.004 | 128,993 | 4.898 |
| | $C_{\mathcal{D}_A}^{\overline{\mathrm{EMG}}}$ | 0.148 | [0.096, 0.214] | 85.2% | $-0.352^*$ | +0.001 | 147,409 | 5.597 |
| | $C_{\mathcal{D}_A}^{\overline{\mathrm{ET}}}$ | 0.141 | [0.090, 0.206] | 85.9% | $-0.359^*$ | −0.006 | 72,323 | 2.746 |

$\epsilon$: expected loss; $CI_{99}$: 99% confidence interval for $\epsilon$; $A$: accuracy; $\Delta\epsilon_{\mathrm{B}}$: loss variation with respect to the benchmark; $\Delta\epsilon_{\mathrm{F}}$: loss variation with respect to the full model; $C$: RF structural complexity; $\rho_C$: relative RF complexity.
–: non-RF model.
$^*$: non-overlap between marginal $CI_{99}$ intervals.

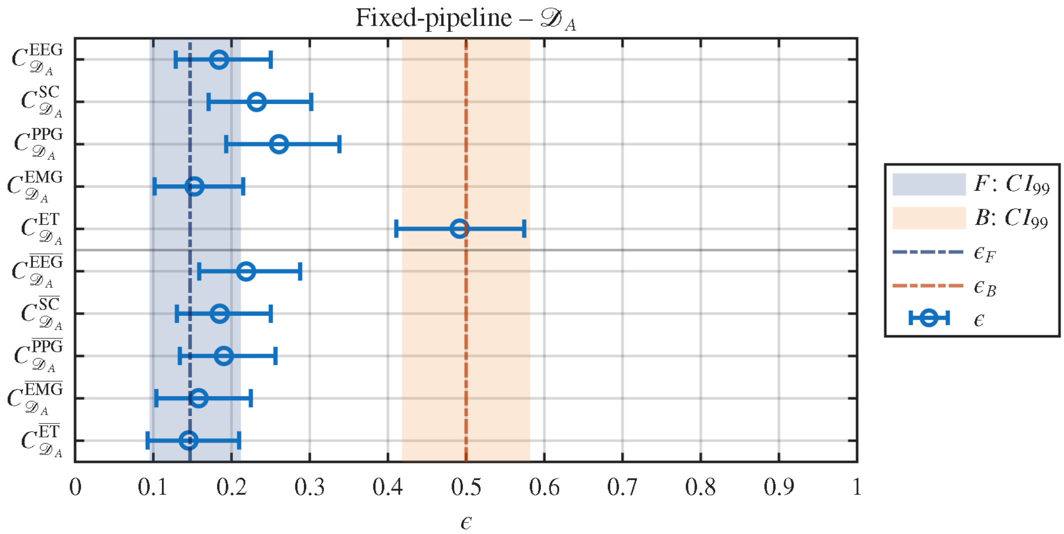


Figure 7.3: Expected loss of fixed-pipeline ablated models for $\mathcal{D}_A$. Points and horizontal error bars indicate $\epsilon$ and $CI_{99}$, respectively. Shaded bands and vertical lines represent the $CI_{99}$ and expected loss of the full ($F$) and benchmark ($B$) models.

perparameters, and feature-selection configuration of the corresponding full models were kept fixed, whereas in the re-CASH ablation the Combined Algorithm Selection and Hyperparameter optimisation (CASH) procedure was repeated independently for each ablated dataset. For RF-based re-CASH models, we additionally computed the structural complexity $C$ and the relative complexity coefficient $\rho_C$ with respect to the corresponding full model.

For both fixed-pipeline and re-CASH experiments, most configurations outperformed the benchmark under the adopted conservative CI-overlap criterion, without showing detectable degradation relative to the corresponding full models. This pattern was observed in both unimodal and LOMO studies, suggesting the robustness of the RF-based pipelines to modality removal. When retrained or re-optimised, RFs may tolerate reduced feature spaces if the remaining modalities retain sufficient information, consistent with their ensemble nature (Breiman, 2001; Hor & Moradi, 2015). It also suggests that discriminative information was evenly distributed across multiple physiological channels, consistent with both multimodal learning theory (Baltrusaitis et al., 2019) and previous tML studies (Gohumpu et al., 2023; Koelstra et al., 2012; Miranda-Correa et al., 2021; Soleymani et al., 2012).

ET represented the only consistent exception. In both fixed-pipeline and re-CASH experiments, ET-unimodal models performed close to the benchmark and worse than the full models under the adopted criterion, whereas removing ET in the LOMO configuration did not degrade

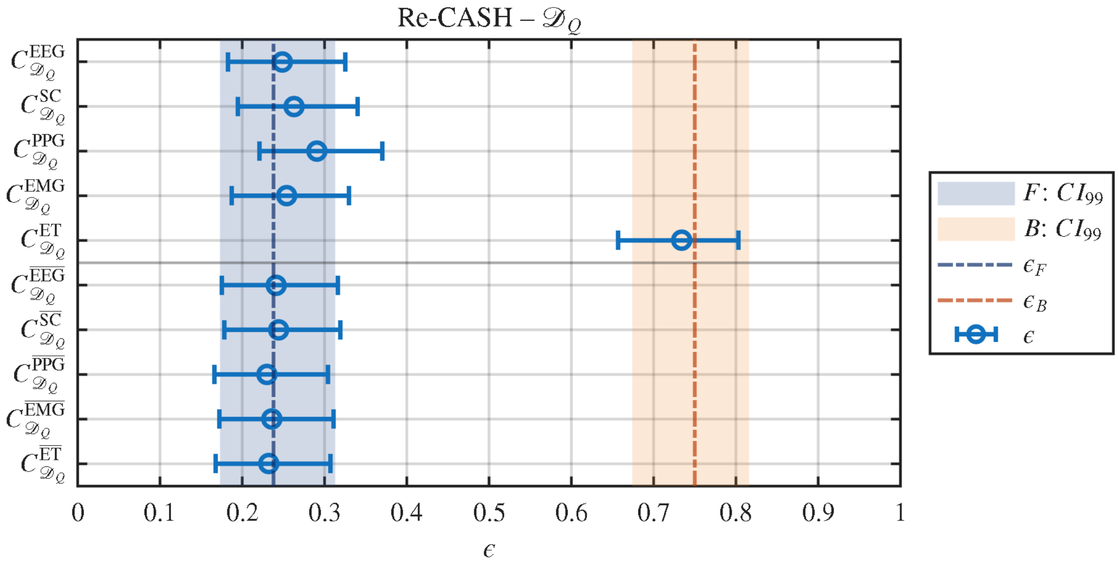


Figure 7.4: Expected loss of re-CASH ablated models for $\mathcal{D}_Q$. Points indicate $\epsilon$, horizontal error bars indicate $CI_{99}$, and shaded bands denote the $CI_{99}$ of the full ($F$) and benchmark ($B$) models.

performance. This supports the limited informativeness of the ET features extracted in this work, already discussed in Chapter 6. It is worth restating that this is feature-specific rather than modality-dependent, as previous studies have shown the contribution of richer eye-tracking descriptors to emotion recognition (Skaramagkas et al., 2023).

The complexity analysis showed that full-model-comparable re-CASH performance was not always obtained through structurally comparable RF solutions. On average, ablations showed higher structural complexity than the corresponding full models; however, this increase was not systematic, since the number of configurations with $\rho_C > 1$ and $\rho_C < 1$ was identical. Except for ET, which showed consistently weak standalone and incremental contribution, no clear modality-level pattern emerged. The clearest dataset-level trend was observed for $\mathcal{D}_A$, where most ablated RFs were more complex than the full model. This increase may suggest that, after modality reduction, re-CASH often selected more articulated RF structures to preserve arousal performance. Such an interpretation is compatible with the distributed representation of arousal, which has been associated with widespread cortical synchronisation (Kim et al., 2021) and multiple autonomic and peripheral responses (C.-A. Wang et al., 2018).

Taken together, the results suggest that the proposed pipeline can be adapted to reduced-modality configurations while preserving predictive performance close to that of the full models, with a non-monotonic increase in complexity. This is relevant from an applied perspective,

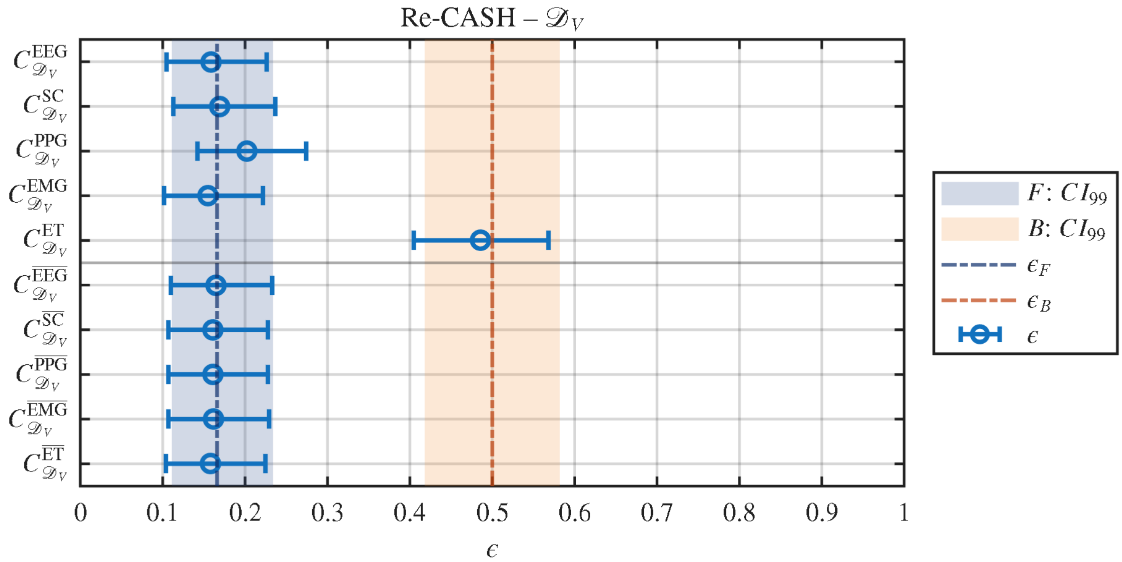


Figure 7.5: Expected loss of re-CASH ablated models for $\mathcal{D}_V$. Points indicate $\epsilon$, horizontal error bars indicate $CI_{99}$, and shaded bands denote the $CI_{99}$ of the full ($F$) and benchmark ($B$) models.

because a reliable AD system may not necessarily require the simultaneous acquisition of all physiological signals. Instead, depending on the target dataset and acceptable performance–complexity trade-off, the system could potentially be simplified towards fewer sensors, or even unimodal configurations. This has implications for scalability, cost, and user burden, especially in applications where full multimodal acquisition may be impractical, such as wearable-based emotion-recognition research (Saganowski et al., 2022; Schmidt et al., 2019; Shu et al., 2018).

Despite these promising results, some limitations should be acknowledged. First, ablations were performed at the modality level. Since EEG dominated the full-model feature-importance analysis, future channel-wise or region-wise ablations could clarify whether performance depends on specific scalp areas or on a distributed EEG contribution, as recently explored by Valderrama and Sheoran (2025). Second, comparisons were based on a limited number of repeated LPSO splits and on the overlap of marginal $CI_{99}$ intervals, which is a conservative approach. Formal comparisons would require inference directly on paired performance differences, using procedures that account for the dependence induced by resampling-based evaluation (Dietterich, 1998; Nadeau & Bengio, 2003). Third, the ablation analyses were performed by retraining or re-optimising models on predefined reduced-modality datasets. They therefore assess the performance of reduced-sensor configurations, rather than robustness to arbitrary missing modalities at inference time. This latter problem would require specific missing-modality learning (Zhan

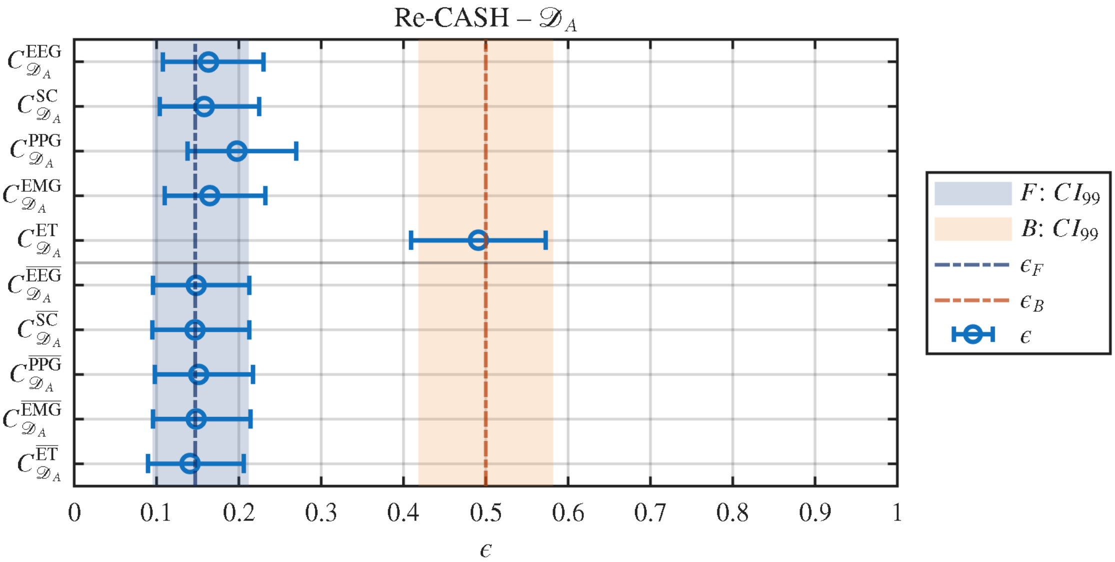


Figure 7.6: Expected loss of re-CASH ablated models for $\mathcal{D}_A$. Points indicate $\epsilon$, horizontal error bars indicate $CI_{99}$, and shaded bands denote the $CI_{99}$ of the full ($F$) and benchmark ($B$) models.

et al., 2025) or modality-dropout strategies (J. Zhao et al., 2021). Finally, the results were not externally validated on independent datasets, partly because directly comparable static-picture multimodal datasets are currently lacking (Bilucaglia et al., 2024).

# Chapter 8

# Conclusions

This work focused on emotions, a key variable in predicting, mediating, and moderating consumption-related behaviours. It addressed the problem of emotion measurement ($\psi$) through multimodal Affect Detection (AD) models based on bioelectrical and biometric modalities ($\phi$), in order to directly estimate $\mathcal{P}(\psi|\phi)$. This approach represents a conceptual shift in Consumer Neuroscience, as it has traditionally relied on reverse inference, based on the assumption that large-scale neuroimaging studies yield $\mathcal{P}(\phi|\psi) \approx 1$.

The first part of the thesis established the theoretical and methodological foundations. It began by reviewing psychological and physiological theories of emotion, clarifying why affective states should be understood as multidimensional phenomena involving subjective, neural, autonomic and behavioural components. It then introduced Consumer Neuroscience and Psychophysiology, highlighting the limitations of reverse inference and motivating AD as a supervised modelling framework for estimating affective states. Finally, it reviewed the ML concepts underlying the proposed modelling approach, including statistical learning, model selection and error estimation.

The first empirical contribution is I DARE (IULM Dataset of Affective REsponses), a novel multimodal dataset of affective responses to static visual stimuli. I DARE included electroencephalogram, electromyogram, skin conductance, photoplethysmogram and eye-tracking recordings from 63 participants exposed to 32 emotional images. The stimuli were selected from standard affective databases through a semi-automatic procedure based on their distribution in the valence–arousal space and on ratings from independent judges. With 5 modalities, 63 participants and 32 stimuli, I DARE provides a dimensionality of 10080 and a minimum detectable effect size of $f = 0.095$, surpassing existing datasets.

The second contribution is MATILDE (Multimodal biosignAl-based modaliTy-resilient subject-Independent modeL for affect DEtection), a multimodal subject-independent AD model developed on I DARE. Its development was formulated as a Combined Algorithm Selection and Hyperparameter optimisation (CASH) problem, exploring multiple classifiers, feature-selection strategies and hyperparameter configurations through Bayesian optimisation. The best MATILDE configurations, based on Random Forest classifiers, achieved accuracies of 76.2% for quadrant classification, 83.4% for valence classification and 85.3% for arousal classification. These results were above the corresponding benchmarks and were consistent with the upper range of traditional ML-based AD systems.

The robustness of MATILDE was then assessed through dataset-level ablation studies. Both unimodal and Leave-One-Modality-Out configurations were tested under two strategies: a fixed-pipeline setting, where the full-model configuration was preserved, and a re-CASH setting, where the optimisation procedure was repeated for each reduced-modality dataset. Most reduced-modality configurations remained above their benchmarks and close to the corresponding full models, suggesting that affective information was distributed across multiple physiological channels. Eye tracking was the main exception, showing limited standalone performance and no evident incremental contribution in Leave-One-Modality-Out configurations. The Random Forest structural analysis further showed that reduced-modality models sometimes required greater complexity to preserve performance, especially in the re-CASH setting.

Overall, the results support a transition from isolated signal interpretation and reverse inference towards data-driven, multimodal and explicitly validated models of affective-state estimation. In this perspective, physiological and biometric signals should not be treated as direct markers of emotion, but as complementary sources of information whose affective meaning emerges through validated modelling frameworks. This shift may support more robust, scalable and methodologically grounded applications of emotion measurement in Consumer Neuroscience.

# Bibliography


Abu Alfeilat, H. A., Hassanat, A. B., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., & Prasath, V. S. (2019). Effects of distance measure choice on k-nearest neighbor classifier performance: A review. *Big Data*, *7*(4), 221–248. https://doi.org/10.1089/big.2018.0175

Aeron, S., & Rahman, Z. (2024). Emotion as cause, effect, mediator, and moderator in marketing: An integrative review and future research directions. *Journal of Consumer Behaviour*, *24*(1), 470–498. https://doi.org/10.1002/cb.2430

Afroza, S., Farruk, A., & Shafiul, A. M. (2023). A systematic review on surface electromyography-based classification system for identifying hand and finger movements. *Healthcare Analytics*, *3*, 100126. https://doi.org/10.1016/j.health.2022.100126

Ahmad, Z., & Khan, N. (2022). A survey on physiological signal-based emotion recognition. *Bioengineering*, *9*(11), 688. https://doi.org/10.3390/bioengineering9110688

Ahmed, N., Aghbari, Z. A., & Girija, S. (2023). A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, *17*, 200171. https://doi.org/10.1016/j.iswa.2022.200171

Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., & Salakoski, T. (2011). An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Computational Statistics & Data Analysis*, *55*(4), 1828–1844. https://doi.org/10.1016/j.csda.2010.11.018

Akhtar, M. T., Jung, T.-P., Makeig, S., & Cauwenberghs, G. (2012). Recursive independent component analysis for online blind source separation. *2012 IEEE International Symposium on Circuits and Systems*, 2013–2016. https://doi.org/10.1109/iscas.2012.6271896

Alazrai, R., Homoud, R., Alwanni, H., & Daoud, M. I. (2018). EEG-based emotion recognition using quadratic time-frequency distribution. *Sensors*, *18*(8), 2739. https://doi.org/10.3390/s18082739

Aljalal, M., Aldosari, S. A., Molinas, M., & Alturki, F. A. (2024). Selecting EEG channels and features using multi-objective optimization for accurate MCI detection: Validation using leave-one-subject-out strategy. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-63180-y

Al-Nafjan, A., Hosny, M., Al-Ohali, Y., & Al-Wabil, A. (2017). Review and classification of emotion recognition based on EEG brain-computer interface system research: A systematic review. *Applied Sciences*, *7*(12), 1239. https://doi.org/10.3390/app7121239

Al-Shawaf, L., Conroy-Beam, D., Asao, K., & Buss, D. M. (2015). Human emotions: An evolutionary psychological perspective. *Emotion Review*, *8*(2), 173–186. https://doi.org/10.1177/1754073914565518

Altman, N., & Krzywinski, M. (2018). The curse(s) of dimensionality. *Nature Methods*, *15*(6), 399–400. https://doi.org/10.1038/s41592-018-0019-x

Aluja-Banet, T., & Nafria, E. (1998). Robust impurity measures in decision trees. In *Data science, classification, and related methods* (pp. 207–214). Springer Japan. https://doi.org/10.1007/978-4-431-65950-1_21

Arevalillo-Herráez, M., Cobos, M., Roger, S., & García-Pineda, M. (2019). Combining inter-subject modeling with a subject-based data transformation to improve affect recognition from EEG signals. *Sensors*, *19*(13), 2999. https://doi.org/10.3390/s19132999

Arnold, M. B. (1960). *Emotion and personality*. Columbia University Press.

Arthanarisamy Ramaswamy, M. P., & Palaniswamy, S. (2022). Subject independent emotion recognition using EEG and physiological signals – a comparative study. *Applied Computing and Informatics*. https://doi.org/10.1108/aci-03-2022-0080

Bach, D. R., Castegnetti, G., Korn, C. W., Gerster, S., Melinscak, F., & Moser, T. (2018). Psychophysiological modeling: Current state and future directions. *Psychophysiology*, *55*(11). https://doi.org/10.1111/psyp.13209

Bagozzi, R. P., Gopinath, M., & Nyer, P. U. (1999). The role of emotions in marketing. *Journal of the Academy of Marketing Science*, *27*(2), 184–206. https://doi.org/10.1177/0092070399272005

Baig, M. Z., & Kavakli, M. (2019). A survey on psycho-physiological analysis & measurement methods in multimodal systems. *Multimodal Technologies and Interaction*, *3*(2), 37. https://doi.org/10.3390/mti3020037

Balconi, M., Venturella, I., Sebastiani, R., & Angioletti, L. (2021). Touching to feel: Brain activity during in-store consumer experience. *Frontiers in psychology*, *12*, 653011. https://doi.org/10.3389/fpsyg.2021.653011

Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443. https://doi.org/10.1109/tpami.2018.2798607

Bamonte, M. F., Risk, M., & Herrero, V. (2024). Determining the optimal window duration to enhance emotion recognition based on galvanic skin response and photoplethysmography signals. *Electronics*, *13*(16), 3333. https://doi.org/10.3390/electronics13163333

Baratchi, M., Wang, C., Limmer, S., van Rijn, J. N., Hoos, H., Bäck, T., & Olhofer, M. (2024). Automated machine learning: Past, present and future. *Artificial Intelligence Review*, *57*(5). https://doi.org/10.1007/s10462-024-10726-1

Bard, P. (1929). Emotion: I. the neuro-humoral basis of emotional reactions. In C. Murchison (Ed.), *The foundations of experimental psychology.* (pp. 449–487). Clark University Press. https://doi.org/10.1037/11018-012

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, *16*, 419–429. https://doi.org/10.1038/nrn3950

Barrett, L. F., & Westlin, C. (2021). Navigating the science of emotion. In H. L. Meiselman (Ed.), *Emotion measurement* (pp. 39–84). Elsevier.

Bartlett, P. L., Harvey, N., Liaw, C., & Mehrabian, A. (2019). Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, *20*(63), 1–17.

Bassett, R., & Deride, J. (2019). Maximum a posteriori estimators as a limit of bayes estimators. *Mathematical Programming*, *174*, 129–144. https://doi.org/10.1007/s10107-018-1241-0

Bekhet, A. K., & Zauszniewski, J. A. (2012). Methodological triangulation: An approach to understanding data. *Nurse Researcher*, *20*(2), 40–43. https://doi.org/10.7748/nr2012.11.20.2.40.c9442

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15849–15854. https://doi.org/10.1073/pnas.1903070116

Bengoechea, A. G., Ruiz, F. J. G., Herrera, M. H., & Crespo, R. A. (2023). Neuromarketing and e-commerce: Analysis of over the top platform homepages. *International Journal of Services Operations and Informatics*, *12*(3), 253–266. https://doi.org/10.1504/ijsoi.2023.132354

Bento, N., Rebelo, J., Barandas, M., Carreiro, A. V., Campagner, A., Cabitza, F., & Gamboa, H. (2022). Comparing handcrafted features and deep neural representations for domain generalization in human activity recognition. *Sensors*, *22*(19), 7324. https://doi.org/10.3390/s22197324

Bertsimas, D., & Dunn, J. (2017). Optimal classification trees. *Machine Learning*, *106*(7), 1039–1082. https://doi.org/10.1007/s10994-017-5633-9

Betella, A., & Verschure, P. F. M. J. (2016). The affective slider: A digital self-assessment scale for the measurement of human emotions. *PLOS ONE*, *11*(2), e0148037. https://doi.org/10.1371/journal.pone.0148037

Bhattacharyya, A., Tripathy, R. K., Garg, L., & Pachori, R. B. (2021). A novel multivariate-multiscale approach for computing EEG spectral and temporal complexity for human emotion recognition. *IEEE Sensors Journal*, *21*(3), 3579–3591. https://doi.org/10.1109/jsen.2020.3027181

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*. https://doi.org/10.3389/fninf.2015.00016

Bigne, E., Boksem, M., Casado-Aranda, L. A., García-Madariaga, J., Gier-Reinartz, N. R., Guerreiro, J., Loureiro, S., Kakaria, S., Smidts, A., & Wedel, M. (2025). How to conduct valuable marketing research with neurophysiological tools. *Psychology & Marketing*, *42*(10), 2616–2649. https://doi.org/10.1002/mar.70002

Bilucaglia, M., Laureanti, R., Circi, R., Zito, M., Bellati, M., Fici, A., Rivetti, F., Mainardi, L. T., & Russo, V. (2022). Spectral differences in resting-state EEG associated to individual emotional styles. *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 4052–4055. https://doi.org/10.1109/embc48229.2022.9871191

Bilucaglia, M., Laureanti, R., Zito, M., Circi, R., Fici, A., Rivetti, F., Valesi, R., Wahl, S., & Russo, V. (2019). Looking through blue glasses: Bioelectrical measures to assess the awakening after a calm situation. *2019 41st Annual International Conference of the*

*IEEE Engineering in Medicine and Biology Society (EMBC)*, 526–529. https://doi.org/10.1109/embc.2019.8856486

Bilucaglia, M., Duma, G. M., Mento, G., Semenzato, L., & Tressoldi, P. (2021). Applying machine learning EEG signal classification to emotion-related brain anticipatory activity. *F1000Research*, *9*, 173. https://doi.org/10.12688/f1000research.22202.3

Bilucaglia, M., Mainardi, L., Ramsøy, T. Z., Zak, P. J., Zito, M., & Russo, V. (2025). Editorial: Machine-learning/deep-learning methods in neuromarketing and consumer neuroscience. *Frontiers in Human Neuroscience*, *19*. https://doi.org/10.3389/fnhum.2025.1638225

Bilucaglia, M., Masi, R., Stanislao, G. D., Laureanti, R., Fici, A., Circi, R., Zito, M., & Russo, V. (2020). ESB: A low-cost EEG synchronization box. *HardwareX*, *8*, e00125. https://doi.org/10.1016/j.ohx.2020.e00125

Bilucaglia, M., Zito, M., Fici, A., Casiraghi, C., Rivetti, F., Bellati, M., & Russo, V. (2024). I DARE: IULM dataset of affective responses. *Frontiers in Human Neuroscience*, *18*. https://doi.org/10.3389/fnhum.2024.1347327

Bishop, C. (2006). *Pattern recognition and machine learning*. Springer.

Blignaut, P. (2009). Fixation identification: The optimum threshold for a dispersion algorithm. *Attention, Perception & Psychophysics*, *71*(4), 881–895. https://doi.org/10.3758/app.71.4.881

Boghossian, P. (2012). What is inference? *Philosophical Studies*, *169*(1), 1–18. https://doi.org/10.1007/s11098-012-9903-x

Borghini, G., Aricò, P., Flumeri, G. D., Sciaraffa, N., & Babiloni, F. (2019). Correlation and similarity between cerebral and non-cerebral electrical activity for user's states assessment. *Sensors*, *19*(3), 704. https://doi.org/10.3390/s19030704

Bota, P. J., Wang, C., Fred, A. L. N., & Placido Da Silva, H. (2019). A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE Access*, *7*, 140990–141020. https://doi.org/10.1109/access.2019.2944001

Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., Roth, W. T., Dawson, M. E., & Filion, D. L. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, *49*(8), 1017–1034. https://doi.org/10.1111/j.1469-8986.2012.01384.x

Boyer, M., Bouyer, L., Roy, J.-S., & Campeau-Lecours, A. (2023). Reducing noise, artifacts and interference in single-channel EMG signals: A review. *Sensors*, *23*(6), 2927. https://doi.org/10.3390/s23062927

Bradley, M. M., Greenwald, M. K., Petry, M. C., & Lang, P. J. (1992). Remembering pictures: Pleasure and arousal in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 379–390. https://doi.org/10.1037/0278-7393.18.2.379

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

Branco, D., Gonçalves, Ö. F., & Badia, S. B. i. (2023). A systematic review of international affective picture system (IAPS) around the world. *Sensors*, *23*(8), 3866. https://doi.org/10.3390/s23083866

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth.

Brodley, C. E., & Utgoff, P. E. (1995). Multivariate decision trees. *Machine Learning*, *19*(1), 45–77. https://doi.org/10.1007/bf00994660

Brown, G., & Ali, R. (2024). Bias/variance is not the same as approximation/estimation. *Transactions on Machine Learning Research*, (2).

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121–167. https://doi.org/10.1023/A:1009715923555

Cabanac, M. (2002). What is emotion? *Behavioural Processes*, *60*(2), 69–83. https://doi.org/10.1016/s0376-6357(02)00078-5

Cacioppo, J. T., & Tassinary, L. G. (1990). Inferring psychological significance from physiological signals. *American psychologist*, *45*(1), 16. https://doi.org/10.1037/0003-066x.45.1.16

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. (2014). *Handbook of psychophysiology*. Cambridge university press. https://doi.org/10.1017/9781107415782

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

Camara, C. F., Filippetti, M. L., & Sel, A. (2025). Neurophysiological dynamics of visceral signals in emotion, self and bodily consciousness. *Proceedings of the Royal Society B: Biological Sciences*, *292*(2048). https://doi.org/10.1098/rspb.2024.2625

Cannon, W. B. (1927). The james-lange theory of emotions: A critical examination and an alternative theory. *The American Journal of Psychology*, *39*(1/4), 106–. https://doi.org/10.2307/1415404

Cannon, W. B. (1931). Again the james-lange and the thalamic theories of emotion. *Psychological Review*, *38*(4), 281–295. https://doi.org/10.1037/h0072957

Casado-Aranda, L.-A., Sánchez-Fernández, J., Bigne, E., & Smidts, A. (2023). The application of neuromarketing tools in communication research: A comprehensive review of trends. *Psychology & Marketing*, *40*(9), 1737–1756. https://doi.org/10.1002/mar.21832

Castellion, G., & Markham, S. K. (2013). New product failure rates: Influence of *Argumentum ad Populum* and self-interest. *Journal of Product Innovation Management*, *30*(5), 976–979. https://doi.org/10.1111/j.1540-5885.2012.01009.x

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chang, C.-Y., Hsu, S.-H., Pion-Tonachini, L., & Jung, T.-P. (2020). Evaluation of artifact subspace reconstruction for automatic artifact components removal in multi-channel EEG recordings. *IEEE Transactions on Biomedical Engineering*, *67*(4), 1114–1121. https://doi.org/10.1109/tbme.2019.2930186

Charlton, P. H., Kotzen, K., Mejïa-Mejïa, E., Aston, P. J., Budidha, K., Mant, J., Pettit, C., Behar, J. A., & Kyriacou, P. A. (2022). Detecting beats in the photoplethysmogram: Benchmarking open-source algorithms. *Physiological Measurement*, *43*(8), 085007. https://doi.org/10.1088/1361-6579/ac826d

Charniak, E. (1983). The bayesian basis of common sense medical diagnosis. *Proceedings of the Third AAAI Conference on Artificial Intelligence (AAAI'83)*, 70–73.

Chen, H., Li, J., He, H., Zhu, J., Sun, S., Li, X., & Hu, B. (2025). Toward the construction of affective brain-computer interface: A systematic review. *ACM Computing Surveys*, *57*(6), 1–56. https://doi.org/10.1145/3712259

Chen, J., Ro, T., & Zhu, Z. (2022). Emotion recognition with audio, video, EEG, and EMG: A dataset and baseline approaches. *IEEE Access*, *10*, 13229–13242. https://doi.org/10.1109/access.2022.3146729

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cherkassky, V., & Ma, Y. (2009). Another look at statistical learning theory and regularization. *Neural Networks*, *22*(7), 958–969. https://doi.org/10.1016/j.neunet.2009.04.005

Cherubino, P., Caratü, M., Modica, E., Rossi, D., Trettel, A., Maglione, A. G., Casa, R. D., Dall'Olio, M., Quadretti, R., & Babiloni, F. (2017). Assessing cerebral and emotional activity during the purchase of fruit and vegetable products in the supermarkets. *Neuroeconomic and Behavioral Aspects of Decision Making: Proceedings of the 2016 Computational Methods in Experimental Economics (CMEE) Conference*, 293–307. https://doi.org/10.1007/978-3-319-62938-4_19

Cherubino, P., Martinez-Levy, A. C., Caratu, M., Cartocci, G., Di Flumeri, G., Modica, E., Rossi, D., Mancini, M., & Trettel, A. (2019). Consumer behaviour through the eyes of neurophysiological measures: State-of-the-art and future trends. *Computational intelligence and neuroscience*, 1–44. https://doi.org/10.1155/2019/1976847

Choi, B. C. K., & Pak, A. W. P. (2004). A catalog of biases in questionnaires. *Preventing Chronic Disease*, *2*(1), A13.

Choi, E., & Lee, C. (2003). Feature extraction based on the bhattacharyya distance. *Pattern Recognition*, *36*(8), 1703–1709. https://doi.org/10.1016/s0031-3203(03)00035-9

Chutia, T., & Baruah, N. (2024). A review on emotion detection by using deep learning techniques. *Artificial Intelligence Review*, *57*(8). https://doi.org/10.1007/s10462-024-10831-1

Ciceri, A., Russo, V., Songa, G., Gabrielli, G., & Clement, J. (2020). A neuroscientific method for assessing effectiveness of digital vs. print ads: Using biometric techniques to measure cross-media ad experience and recall. *Journal of Advertising Research*, *60*(1), 71–86. https://doi.org/10.2501/JAR-2019-015

Cimtay, Y., Ekmekcioglu, E., & Caglar-Ozhan, S. (2020). Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access*, *8*, 168865–168878. https://doi.org/10.1109/access.2020.3023871

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Collins, G. S., de Groot, J. A., Dutton, S., Omar, O., Shanyinde, M., Tajar, A., Voysey, M., Wharton, R., Yu, L.-M., Moons, K. G., & Altman, D. G. (2014). External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Medical Research Methodology*, *14*(1). https://doi.org/10.1186/1471-2288-14-40

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/bf00994018

Craig, A. D. (2002). How do you feel? interoception: The sense of the physiological condition of the body. *Nature Reviews Neuroscience*, *3*(8), 655–666. https://doi.org/10.1038/nrn894

Curth, A., Jeffares, A., & van der Schaar, M. (2023). A u-turn on double descent: Rethinking parameter counting in statistical learning. *Advances in Neural Information Processing Systems*, *36*, 55932–55962.

Dabrowski, A. (2016). Emotions in philosophy. a short introduction. *Studia Humana*, *5*(3), 8–20. https://doi.org/10.1515/sh-2016-0011

Dall'Olio, L., Curti, N., Remondini, D., Safi Harb, Y., Asselbergs, F. W., Castellani, G., & Uh, H.-W. (2020). Prediction of vascular aging based on smartphone acquired PPG signals. *Scientific reports*, *10*(1), 19756. https://doi.org/10.1038/s41598-020-76816-6

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*(1346), 1413–1420. https://doi.org/10.1098/rstb.1996.0125

Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences*, *24*(4), 267–269. https://doi.org/10.1016/j.tics.2020.01.007

Darwin, C. (1872). *The expression of the emotions in man and animals*. John Murray London.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, *41*(7), 909–996. https://doi.org/10.1002/cpa.3160410705

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *41*(1), 1–15. https://doi.org/10.1111/j.2517-6161.1979.tb01052.x

de Cheveigné, A., & Nelken, I. (2019). Filters: When, why, and how (not) to use them. *Neuron*, *102*(2), 280–293. https://doi.org/10.1016/j.neuron.2019.02.039

Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Delplanque, S., & Sander, D. (2021). A fascinating but risky case of reverse inference: From measures to emotions! *Food Quality and Preference*, *92*, 104183. https://doi.org/10.1016/j.foodqual.2021.104183

Dewey, J. (1895). The theory of emotion. *Psychological Review*, *2*(1), 13–32. https://doi.org/10.1037/h0070927

Dietterich, T. G., & Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, *2*, 263–286. https://doi.org/10.1613/jair.105

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197

Ding, C., & Pend, H. (2005). Minimum redundancy feature selection for microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, *03*(02), 185–205. https://doi.org/10.1142/s0219720005001004

Ding, J., Tarokh, V., & Yang, Y. (2018). Model selection techniques: An overview. *IEEE Signal Processing Magazine*, *35*(6), 16–34. https://doi.org/10.1109/msp.2018.2867638

Ding, Z., Ji, Y., Gan, Y., Wang, Y., & Xia, Y. (2024). Current status and trends of technology, methods, and applications of human–computer intelligent interaction (HCII): A bibliometric research. *Multimedia Tools and Applications*, *83*(27), 69111–69144. https://doi.org/10.1007/s11042-023-18096-6

D'Mello, S., Kappas, A., & Gratch, J. (2017). The affective computing approach to affect measurement. *Emotion Review*, *10*(2), 174–183. https://doi.org/10.1177/1754073917696583

D'Mello, S., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*, *47*(3), 1–36. https://doi.org/10.1145/2682899

Dong, L., Li, F., Liu, Q., Wen, X., Lai, Y., Xu, P., & Yao, D. (2017). MATLAB toolboxes for reference electrode standardization technique (REST) of scalp EEG. *Frontiers in Neuroscience*, *11*. https://doi.org/10.3389/fnins.2017.00601

Drummond, C. (2006). Discriminative vs. generative classifiers for cost sensitive learning. *Conference of the Canadian Society for Computational Studies of Intelligence*, 479–490. https://doi.org/10.1007/11766247_41

Du, Z., Ye, X., & Zhao, P. (2024). A novel signal channel attention network for multi-modal emotion recognition. *Frontiers in Neurorobotics*, *18*, 1442080. https://doi.org/10.3389/fnbot.2024.1442080

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification* (2nd ed.). John Wiley & Sons.

Ekman, P. (2009). Darwin's contributions to our understanding of emotional expressions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3449–3451. https://doi.org/10.1098/rstb.2009.0189

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124–129. https://doi.org/10.1037/h0030377

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, *164*(3875), 86–88. https://doi.org/10.1126/science.164.3875.86

Erat, K., Şahin, E. B., Doğan, F., Merdanoğlu, N., Akcakaya, A., & Durdu, P. O. (2024). Emotion recognition with EEG-based brain-computer interfaces: A systematic literature review. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-18259-z

Evgeniou, T., Pontil, M., & Poggio, T. (2000). Regularization networks and support vector machines. *Advances in Computational Mathematics*, *13*(1), 1–50. https://doi.org/10.1023/a:1018946025316

Farfán, F. D., Politti, J. C., & Felice, C. J. (2010). Evaluation of EMG processing techniques using information theory. *BioMedical Engineering OnLine*, *9*(1), 72. https://doi.org/10.1186/1475-925x-9-72

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fazli, S., Popescu, F., Danöczy, M., Blankertz, B., Müller, K.-R., & Grozea, C. (2009). Subject-independent mental state classification in single trials. *Neural networks*, *22*(9), 1305–1312. https://doi.org/10.1016/j.neunet.2009.06.003

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. *Advances in neural information processing systems*, *28*.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378. https://doi.org/10.1037/h0031619

Fostiropoulos, I., Brown, B. N., & Itti, L. (2023). Trustworthy model evaluation on a budget. *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

Fu, B., Chu, W., Gu, C., & Liu, Y. (2024). Cross-modal guiding neural network for multimodal emotion recognition from EEG and eye movement signals. *IEEE Journal of Biomedical and Health Informatics*, *28*(10), 5865–5876. https://doi.org/10.1109/jbhi.2024.3419043

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Galvão, F., Alarcão, S. M., & Fonseca, M. J. (2021). Predicting exact valence and arousal values from EEG. *Sensors*, *21*(10), 3414. https://doi.org/10.3390/s21103414

Gamboa, J. (2020). Goltz against cerebral localization: Methodology and experimental practices. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *84*, 101304. https://doi.org/10.1016/j.shpsc.2020.101304

Ghaderyan, P., & Abbasi, A. (2016). An efficient automatic workload estimation method based on electrodermal activity using pattern classifier combinations. *International Journal of Psychophysiology*, *110*, 91–101. https://doi.org/10.1016/j.ijpsycho.2016.10.013

Gohumpu, J., Xue, M., & Bao, Y. (2023). Emotion recognition with multi-modal peripheral physiological signals. *Frontiers in Computer Science*, *5*. https://doi.org/10.3389/fcomp.2023.1264713

Gomez, P., von Gunten, A., & Danuser, B. (2016). Autonomic nervous system reactivity within the valence–arousal affective space: Modulation by sex and age. *International Journal of Psychophysiology*, *109*, 51–62. https://doi.org/10.1016/j.ijpsycho.2016.10.002

Gratch, J., & Marsella, S. C. (2015). Appraisal models. In *The oxford handbook of affective computing*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199942237.013.015

Greco, A., Valenza, G., Citi, L., & Scilingo, E. P. (2017). Arousal and valence recognition of affective sounds based on electrodermal activity. *IEEE Sensors Journal*, *17*(3), 716–725. https://doi.org/10.1109/jsen.2016.2623677

Greco, A., Valenza, G., Lanata, A., Scilingo, E., & Citi, L. (2016). Cvxeda: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering*, 1–1. https://doi.org/10.1109/tbme.2015.2474131

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 507–520.

Guarrasi, V., Aksu, F., Caruso, C. M., Di Feola, F., Rofena, A., Ruffini, F., & Soda, P. (2025). A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. *Image and Vision Computing*, *158*, 105509. https://doi.org/10.1016/j.imavis.2025.105509

Gucciardi, A., Crotti, M., Itzhak, N. B., Mailleux, L., Ortibus, E., Michelucci, U., Groznik, V., & Sadikov, A. (2022). A new median filter application to deal with large windows of missing data in eye-gaze measurements. *Proceedings of the International Workshop on Neurodevelopmental Impairments in Preterm Children - Computational Advancements*, 1–17.

Guo, R., Guo, H., Wang, L., Chen, M., Yang, D., & Li, B. (2024). Development and application of emotion recognition technology — a systematic literature review. *BMC Psychology*, *12*(1). https://doi.org/10.1186/s40359-024-01581-4

Gupta, R., Kapoor, A. P., & Verma, H. V. (2025). Neuro-insights: A systematic review of neuromarketing perspectives across consumer buying stages. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1542847

Guyon, I., Vapnik, V., Boser, B., Bottou, L., & Solla, S. A. (1991). Structural risk minimization for character recognition. *Proceedings of the 4th International Conference on Neural Information Processing Systems (NIPS)*, 471, 479.

Hakim, A., & Levy, D. J. (2018). A gateway to consumers' minds: Achievements, caveats, and prospects of electroencephalography-based prediction in neuromarketing. *WIREs Cognitive Science*, *10*(2). https://doi.org/10.1002/wcs.1485

Hand, D. J., & Yu, K. (2001). Idiot's bayes: Not so stupid after all? *International Statistical Review / Revue Internationale de Statistique*, *69*(3), 385. https://doi.org/10.2307/1403452

Hernández, V. A. S., Monroy, R., Medina-Pérez, M. A., Loyola-González, O., & Herrera, F. (2021). A practical tutorial for decision tree induction: Evaluation measures for candidate splits and opportunities. *ACM Computing Surveys*, *54*(1), 1–38. https://doi.org/10.1145/3429739

Higham, C. F., & Higham, D. J. (2019). Deep learning: An introduction for applied mathematicians. *SIAM Review*, *61*(4), 860–891. https://doi.org/10.1137/18m1165748

Hjorth, B. (1970). EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, *29*(3), 306–310. https://doi.org/10.1016/0013-4694(70)90143-4

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd international conference on document analysis and recognition*, *1*, 278–282. https://doi.org/10.1109/ICDAR.1995.598994

Holmström, L., & Koistinen, P. (2010). Pattern recognition. *WIREs Computational Statistics*, *2*(4), 404–413. https://doi.org/10.1002/wics.99

Hor, S., & Moradi, M. (2015). Scandent tree: A random forest learning method for incomplete multimodal datasets. In *Medical image computing and computer-assisted intervention – MICCAI 2015* (pp. 694–701). Springer International Publishing. https://doi.org/10.1007/978-3-319-24553-9_85

Hutzler, F. (2014). Reverse inference is not a fallacy per se: Cognitive processes can be inferred from functional imaging data. *NeuroImage*, *84*, 1061–1069. https://doi.org/10.1016/j.neuroimage.2012.12.075

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*(4-5), 411–430. https://doi.org/10.1016/s0893-6080(00)00026-5

Imani, M., & Montazer, G. A. (2019). A survey of emotion recognition methods with emphasis on e-learning environments. *Journal of Network and Computer Applications*, *147*, 102423. https://doi.org/10.1016/j.jnca.2019.102423

Izard, C. E. (1971). *The face of emotion*. Appleton–Century–Crofts.

Izard, C. E. (1972). *Patterns of emotions*. Academic Press.

Izard, C. E. (2010). The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Review*, *2*(4), 363–370. https://doi.org/10.1177/1754073910374661

Jain, A., Duin, P., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4–37. https://doi.org/10.1109/34.824819

Jain, A., & Ramaswami, M. (1988). Classifier design with parzen windows† †research supported by NSF grant ECS-8603541. In *Pattern recognition and artificial intelligence - towards an integration* (pp. 211–228). Elsevier. https://doi.org/10.1016/b978-0-444-87137-4.50021-7

James, W. (1884). What is an emotion? *Mind*, *9*(34), 188–205. https://doi.org/10.1093/mind/os-IX.34.188

Jia, Z., Zhao, F., Guo, Y., Chen, H., & Jiang, T. (2024). Multi-level disentangling network for cross-subject emotion recognition based on multimodal physiological signals. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 3069–3077. https://doi.org/10.24963/ijcai.2024/340

Jung, T.-P., Humphries, C., Lee, T.-W., Makeig, S., McKeown, M., Iragui, V., & Sejnowski, T. J. (1997). Extended ICA removes artifacts from electroencephalographic recordings. *Advances in neural information processing systems*, *10*.

Kar, A. (2020). MLGaze: Machine learning-based analysis of gaze error patterns in consumer eye tracking systems. *Vision*, *4*(2), 25. https://doi.org/10.3390/vision4020025

Karaçalı, B., Ramanath, R., & Snyder, W. E. (2004). A comparative analysis of structural risk minimization by support vector machines and nearest neighbor rule. *Pattern Recognition Letters*, *25*(1), 63–71. https://doi.org/10.1016/j.patrec.2003.09.002

Karim, A., Ryu, S., & Jeong, I. c. (2025). Ensemble learning for biomedical signal classification: A high-accuracy framework using spectrograms from percussion and palpation. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-05027-8

Katsigiannis, S., & Ramzan, N. (2017). DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE journal of biomedical and health informatics*, *22*(1), 98–107. https://doi.org/10.1109/JBHI.2017.2688239

Katz, M. J. (1988). Fractals and the analysis of waveforms. *Computers in Biology and Medicine*, *18*(3), 145–156. https://doi.org/10.1016/0010-4825(88)90041-8

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3146–3154.

Keltner, D., & Gross, J. J. (1999). Functional accounts of emotions. *Cognition & Emotion*, *13*(5), 467–480. https://doi.org/10.1080/026999399379140

Keltner, D., Oatley, K., & Jenkins, J. M. (2018). *Understanding emotions*. John Wiley & Sons.

Kemp, A. H., Krygier, J., & Harmon-Jones, E. (2015). Neuroscientific perspectives of emotion. In R. Calvo, S. D'Mello, J. Gratch, & A. Krappas (Eds.), *The oxford hanbook of affective computing*. Oxford University Press.

Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., & Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, *102*, 102019. https://doi.org/10.1016/j.inffus.2023.102019

Kim, H., Seo, P., Choi, J.-W., & Kim, K.-H. (2021). Emotional arousal due to video stimuli reduces local and inter-regional synchronization of oscillatory cortical activities in alpha- and beta-bands. *PLOS ONE*, *16*(7), e0255032. https://doi.org/10.1371/journal.pone.0255032

Kirby, E. D., Glen, M. J., Sandstrom, N. J., & Williams, C. L. (2024). *Introduction to behavioral neuroscience*. OpenStax.

Kirkland, T., & Cunningham, W. A. (2011). Neural basis of affect and emotion. *WIREs Cognitive Science*, *2*(6), 656–665. https://doi.org/10.1002/wcs.145

Kleckner, I. R., Jones, R. M., Wilder-Smith, O., Wormwood, J. B., Akcakaya, M., Quigley, K. S., Lord, C., & Goodwin, M. S. (2018). Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data. *IEEE Transactions on Biomedical Engineering*, *65*(7), 1460–1467. https://doi.org/10.1109/TBME.2017.2758643

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. *Brain Research Reviews*, *29*(2-3), 169–195. https://doi.org/10.1016/S0165-0173(98)00056-3

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & 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

Koelstra, S., Mühl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & Patras, I. (2012). 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

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1–2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Krassanakis, V., Filippakopoulou, V., & Nakos, B. (2014). Eyemmv toolbox: An eye movement post-analysis tool based on a two-step spatial dispersion threshold for fixation identification. *Journal of Eye Movement Research*, *7*(1). https://doi.org/10.16910/jemr.7.1.4

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

Kumar G S, S., Sampathila, N., & Tanmay, T. (2022). Wavelet based machine learning models for classification of human emotions using EEG signal. *Measurement: Sensors*, *24*, 100554. https://doi.org/10.1016/j.measen.2022.100554

Kurdi, B., Lozano, S., & Banaji, M. R. (2017). Introducing the open affective standardized image set (OASIS). *Behavior research methods*, *49*(2), 457–470. https://doi.org/10.3758/s13428-016-0715-3

Lam, L., & Suen, S. (1997). Application of majority voting to pattern recognition: An analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *27*(5), 553–568. https://doi.org/10.1109/3468.618255

Lampropoulos, G., Fernández-Arias, P., Antón-Sancho, Á., & Vergara, D. (2024). Affective computing in augmented reality, virtual reality, and immersive learning environments. *Electronics*, *13*(15), 2917. https://doi.org/10.3390/electronics13152917

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Lang, P., Bradley, M., & Cuthbert, B. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual* (tech. rep.). University of Florida, Gainesville.

Lange, C. G. (1885). *Om sindsbevægelser. et psyko-fysiologisk studie*. Lund.

Lazarus, R. S. (1966). *Psychological stress and the coping process*. McGraw-Hill.

Le, T., & Clarke, B. (2018). On the interpretation of ensemble classifiers in terms of bayes classifiers. *Journal of Classification*, *35*(2), 198–229. https://doi.org/10.1007/s00357-018-9257-y

LeCun, Y., Denker, J., & Solla, S. (1989). Optimal brain damage. In D. Touretzky (Ed.), *Advances in neural information processing systems* (pp. 598–605, Vol. 2). Morgan-Kaufmann. https://proceedings.neurips.cc/paper_files/paper/1989/file/6c9882bbac1c7093bd25041881277658-Paper.pdf

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, *113*(523), 1094–1111. https://doi.org/10.1080/01621459.2017.1307116

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *NeuroImage*, *56*(2), 387–399. https://doi.org/10.1016/j.neuroimage.2010.11.004

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology*, *66*(1), 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

Li, W., Huan, W., Hou, B., Tian, Y., Zhang, Z., & Song, A. (2022). Can emotion be transferred?—a review on transfer learning for EEG-based emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(3), 833–846. https://doi.org/10.1109/tcds.2021.3098842

Li, X., Li, Y., Li, Y., & Yang, Y. (2025). GMHCA-MCBILSTM: A gated multi-head cross-modal attention-based network for emotion recognition using multi-physiological signals. *Algorithms*, *18*(10), 664. https://doi.org/10.3390/a18100664

Lipton, Z. C., & Steinhardt, J. (2019). Troubling trends in machine learning scholarship. *ACM Queue*, *17*(1), 45–77. https://doi.org/10.1145/3317287.3328534

Liu, B., Zhao, Q.-C., Ren, Y.-Y., Wang, Q.-J., & Zheng, X.-L. (2018). An elaborate algorithm for automatic processing of eye movement data and identifying fixations in eye-tracking experiments. *Advances in Mechanical Engineering*, *10*(5), 168781401877367. https://doi.org/10.1177/1687814018773678

Llurba, C., & Palau, R. (2024). Real-time emotion recognition for improving the teaching–learning process: A scoping review. *Journal of Imaging*, *10*(12), 313. https://doi.org/10.3390/jimaging10120313

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., Ly, A., Gronau, Q. F., Šmíra, M., Epskamp, S., et al. (2019). JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software*, *88*, 1–17. https://doi.org/10.18637/jss.v088.i02

Luck, S. J. (2022). *Applied event-related potential data analysis*. LibreTexts. https://doi.org/10.18115/D5QG92

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

Ma, Z., Li, A., Tang, J., Zhang, J., & Yin, Z. (2025). Multimodal emotion recognition by fusing complementary patterns from central to peripheral neurophysiological signals across feature domains. *Engineering Applications of Artificial Intelligence*, *143*, 110004. https://doi.org/10.1016/j.engappai.2025.110004

Machery, E. (2014). In defense of reverse inference. *The British Journal for the Philosophy of Science*, *65*(2), 251–267. https://doi.org/10.1093/bjps/axs044

MacLean, P. D. (1949). Psychosomatic disease and the" visceral brain": Recent developments bearing on the papez theory of emotion. *Biopsychosocial Science and Medicine*, *11*(6), 338–353. https://doi.org/10.1097/00006842-194911000-00003

MacLean, P. D. (1952). Some psychiatric implications of physiological studies on frontotemporal portion of limbic system (visceral brain). *Electroencephalography and Clinical Neurophysiology*, *4*(4), 407–418. https://doi.org/10.1016/0013-4694(52)90073-4

MacLean, P. D. (1970). The triune brain, emotion, and scientific bias. In F. O. Schmidt (Ed.), *The neuro-sciences second study program* (pp. 336, 349). The Rockfeller University Press.

MacRae, D., van der Hoek, L., de Vette, S., Neh, H., Moreno, A., Fuller, C., Langendijk, J., Valdenegro-Toro, M., Sijtsema, N., van Ooijen, P., & van Dijk, L. (2026). A multi-toxicity deep learning approach for normal tissue complication probability modelling in head and neck cancer patients receiving radiotherapy. *Radiotherapy and Oncology*, *219*, 111486. https://doi.org/10.1016/j.radonc.2026.111486

Malezieux, M., Klein, A. S., & Gogolla, N. (2023). Neural circuits for emotion. *Annual Review of Neuroscience*, *46*(1), 211–231. https://doi.org/10.1146/annurev-neuro-111020-103314

Malik, M., Bigger, J. T., Camm, A. J., Kleiger, R. E., Malliani, A., Moss, A. J., & Schwartz, P. J. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, *17*(3), 354–381. https://doi.org/10.1093/oxfordjournals.eurheartj.a014868

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 674–693. https://doi.org/10.1109/34.192463

Mancini, M., Cherubino, P., Martinez, A., Vozzi, A., Menicocci, S., Ferrara, S., Giorgi, A., Aricò, P., Trettel, A., & Babiloni, F. (2023). What is behind in-stream advertising on YouTube? a remote neuromarketing study employing eye-tracking and facial coding techniques. *Brain Sciences*, *13*(10), 1481. https://doi.org/10.3390/brainsci13101481

Mansour, Y., & McAllester, D. A. (2000). Generalization bounds for decision trees. *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory (COLT'00)*, 69–74.

Marchesini, R. (2021). Animales, etología y filosofía. *Revista Colombiana de Filosofía de la Ciencia*, *21*(43). https://doi.org/10.18270/rcfc.v43i21.3792

Marill, T., & Green, D. M. (1963). On the effectiveness of receptors in recognition systems. *IEEE Transactions on Information Theory*, *9*(1), 11–17. https://doi.org/10.1109/TIT.1963.1057810

Martens, H. H. (1959). Two notes on machine “learning”. *Information and Control*, *2*(4), 364–379. https://doi.org/10.1016/s0019-9958(59)80014-0

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition & Emotion*, *23*(2), 209–237. https://doi.org/10.1080/02699930802204677

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia medica*, *22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Mejbri, N., Essalmi, F., Jemni, M., & Alyoubi, B. A. (2022). Trends in the use of affective computing in e-learning environments. *Education and Information Technologies*, *27*(3), 3867–3889. https://doi.org/10.1007/s10639-021-10769-9

Merletti, R., & Cerone, G. (2020). Tutorial. surface EMG detection, conditioning and preprocessing: Best practices. *Journal of Electromyography and Kinesiology*, *54*, 102440. https://doi.org/10.1016/j.jelekin.2020.102440

Miranda-Correa, J. A., Abadi, M. K., Sebe, N., & Patras, I. (2018). 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

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

Moors, A. (2009). Theories of emotion causation: A review. *Cognition & Emotion*, *23*(4), 625–662. https://doi.org/10.1080/02699930802645739

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review*, *5*(2), 119–124. https://doi.org/10.1177/1754073912468165

Mozer, M. C., & Smolensky, P. (1988). Skeletonization: A technique for trimming the fat from a network via relevance assessment. In D. Touretzky (Ed.), *Advances in neural information processing systems* (pp. 107–115, Vol. 1). Morgan-Kaufmann. https://proceedings.neurips.cc/paper_files/paper/1988/file/07e1cd7dca89a1678042477183b7ac3f-Paper.pdf

Mullen, T. R., Kothe, C. A. E., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., Jung, T.-P., & Cauwenberghs, G. (2015). Real-time neuroimaging and cognitive monitoring using wearable dry EEG. *IEEE Transactions on Biomedical Engineering*, *62*(11), 2553–2567. https://doi.org/10.1109/tbme.2015.2481482

Muller, K.-R., Mika, S., Ratsch, G., Tsuda, K., & Scholkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, *12*(2), 181–201. https://doi.org/10.1109/72.914517

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*(3), 239–281. https://doi.org/10.1023/A:1024068626366

Nakatsu, R. T. (2021). An evaluation of four resampling methods used in machine learning classification. *IEEE Intelligent Systems*, *36*(3), 51–57. https://doi.org/10.1109/mis.2020.2978066

Nakkiran, P., Kaplun, G., Kalimeris, D., Yang, T., Edelman, B. L., Zhang, F., & Barak, B. (2019). SGD on neural networks learns functions of increasing complexity. *Advances in Neural Information Processing Systems*, *32*.

Nalwanga, D., & Henning, L. (2016). If it looks like a duck, swims like a duck, and quacks like a duck—does it have to be a duck? (J. M. Vinetz, Ed.). *PLOS Neglected Tropical Diseases*, *10*(4), e0004430. https://doi.org/10.1371/journal.pntd.0004430

Narendra, P. M., & Fukunaga, K. (1977). A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, *C-26*(9), 917–922. https://doi.org/10.1109/TC.1977.1674939

Natarajan, B. K. (1989). On learning sets and functions. *Machine Learning*, *4*(1), 67–97. https://doi.org/10.1007/bf00114804

Nathan, M. J., & Del Pinal, G. (2017). The future of cognitive neuroscience? reverse inference in focus. *Philosophy Compass*, *12*(7). https://doi.org/10.1111/phc3.12427

Neiss, M. B., Leigland, L. A., Carlson, N. E., & Janowsky, J. S. (2009). Age differences in perception and awareness of emotion. *Neurobiology of Aging*, *30*(8), 1305–1313. https://doi.org/10.1016/j.neurobiolaging.2007.11.007

Nemorin, S. (2016). Neuromarketing and the “poor in world” consumer: How the animalization of thinking underpins contemporary market research discourses. *Consumption Markets & Culture*, *20*(1), 59–80. https://doi.org/10.1080/10253866.2016.1160897

Newell, A. (1975). A tutorial on speech understanding systems. *Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium*, 4–54.

Neyman, J. (1960). In determinism in science and new demands on statisticians. *Journal of the American Statistical Association*, *55*(292), 625–639. https://doi.org/10.1080/01621459.1960.10483363

Ng, A., & Jordan, M. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, *14*.

Nuwer, M. R. (2018). 10-10 electrode system for EEG recording. *Clinical Neurophysiology*, *129*(5), 1103. https://doi.org/10.1016/j.clinph.2018.01.065

Oneto, L. (2020). *Model selection and error estimation in a nutshell*. Springer. https://doi.org/10.1007/978-3-030-24360-5

Panksepp, J. (1982). Toward a general psychobiological theory of emotions. *Behavioral and Brain Sciences*, *5*(3), 407–422. https://doi.org/10.1017/S0140525X00012759

Panksepp, J. (1992). A critical role for "affective neuroscience" in resolving what is basic about basic emotions. *Psychological Review*, *99*(3), 554–560. https://doi.org/10.1037/0033-295x.99.3.554

Panksepp, J. (2011). Cross-species affective neuroscience decoding of the primal affective experiences of humans and related animals (A. Sirigu, Ed.). *PLoS ONE*, *6*(9), e21236. https://doi.org/10.1371/journal.pone.0021236

Pant, S., Yang, H.-J., Lim, E., Kim, S.-H., & Yoo, S.-B. (2023). PhyMER: Physiological dataset for multimodal emotion recognition with personality as a context. *IEEE Access*, *11*, 107638–107656. https://doi.org/10.1109/access.2023.3320053

Papez, J. W. (1937). A proposed mechanism of emotion. *Archives of Neurology & Psychiatry*, *38*(4), 725–743. https://doi.org/10.1001/archneurpsyc.1937.02260220069003

Peelen, M. V., & Downing, P. E. (2023). Testing cognitive theories with multivariate pattern analysis of neuroimaging data. *Nature Human Behaviour*, *7*(9), 1430–1441. https://doi.org/10.1038/s41562-023-01680-z

Penna, M. F., Giordano, L., Tortora, S., Astarita, D., Amato, L., Dell'Agnello, F., Menegatti, E., Gruppioni, E., Vitiello, N., Crea, S., & Trigili, E. (2024). A muscle synergies-based controller to drive a powered upper-limb exoskeleton in reaching tasks. *Wearable Technologies*, *5*. https://doi.org/10.1017/wtc.2024.16

Pepa, L., Spalazzi, L., Capecci, M., & Ceravolo, M. G. (2023). Automatic emotion recognition in clinical scenario: A systematic review of methods. *IEEE Transactions on Affective Computing*, *14*(2), 1675–1695. https://doi.org/10.1109/taffc.2021.3128787

Percival, D. B., & Walden, A. T. (2000). *Wavelet methods for time series analysis*. Cambridge University Press. https://doi.org/10.1017/CBO9780511841040

Pereira, L. M., Salazar, A., & Vergara, L. (2023). On comparing early and late fusion methods. In *Advances in computational intelligence* (pp. 365–378). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43085-5_29

Picard, R. W. (1995). *Affective computing* (tech. rep. No. 321). M.I.T Media Laboratory Perceptual Computing Section.

Picard, R. W. (1997). *Affective computing*. MIT press.

Piho, L., & Tjahjadi, T. (2018). A mutual information based adaptive windowing of informative EEG for emotion recognition. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2018.2840973

Pillalamarri, R., & Shanmugam, U. (2025). A review on EEG-based multimodal learning for emotion recognition. *Artificial Intelligence Review*, *58*(5). https://doi.org/10.1007/s10462-025-11126-9

Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, *88*(6), 2297–2301. https://doi.org/10.1073/pnas.88.6.2297

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197. https://doi.org/10.1016/j.neuroimage.2019.05.026

Plassmann, H., Venkatraman, V., Huettel, S., & Yoon, C. (2015). Consumer neuroscience: Applications, challenges, and possible solutions. *Journal of Marketing Research*, *52*(4), 427–435. https://doi.org/10.1509/jmr.14.0048

Platt, J., Cristianini, N., & Shawe-Taylor, J. (1999). Large margin DAGs for multiclass classification. *Advances in neural information processing systems*, *12*.

Plutchik, R. (1958). Outlines of a new theory of emotion. *Transactions of the New York Academy of Sciences*, *20*, 394–403. https://doi.org/10.1111/j.2164-0947.1958.tb00600.x

Plutchik, R. (1962). *The emotions: Facts, theories, and a new model*. Random House.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research, and experience. vol. 1: Theories of emotion* (pp. 3–33). Academic Press.

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron*, *72*(5), 692–697. https://doi.org/10.1016/j.neuron.2011.11.001

Poldrak, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63. https://doi.org/10.1016/j.tics.2005.12.004

Polo, E. M., Farabbi, A., Mollura, M., Mainardi, L., & Barbieri, R. (2024). Understanding the role of emotion in decision making process: Using machine learning to analyze physiological responses to visual, auditory, and combined stimulation. *Frontiers in Human Neuroscience*, *17*. https://doi.org/10.3389/fnhum.2023.1286621

Porges, S. W. (1995). Orienting in a defensive world: Mammalian modifications of our evolutionary heritage. a polyvagal theory. *Psychophysiology*, *32*(4), 301–318. https://doi.org/10.1111/j.1469-8986.1995.tb01213.x

Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, *27*(3), 221–234. https://doi.org/10.1016/S0020-7373(87)80053-6

Ramsøy, T. Z. (2019). Building a foundation for neuromarketing and consumer neuroscience research. *Journal of Advertising Research*, *59*(3), 281–294. https://doi.org/10.2501/jar-2019-034

Raudys, S., & Duin, R. P. (1998). Expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognition Letters*, *19*(5–6), 385–392. https://doi.org/10.1016/s0167-8655(98)00016-6

Rawnaque, F. S., Rahman, K. M., Anwar, S. F., Vaidyanathan, R., Chau, T., Sarker, F., & Mamun, K. A. A. (2020). Technological advancements and opportunities in neuromarketing: A systematic review. *Brain Informatics*, *7*(1), 1–19. https://doi.org/10.1186/s40708-020-00109-x

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Riegel, M., Wierzba, M., Wypych, M., Żurawski, Ł., Jednoróg, K., Grabowska, A., & Marchewka, A. (2017). Nencki affective picture system: Cross-cultural study in europe and iran. *Frontiers in Psychology*, *8*, 274. https://doi.org/10.3389/fpsyg.2017.00274

Riley, K. F., Hobson, M. P., & Bence, S. J. (2006). *Mathematical methods for physics and engineering*. Cambridge University Press. https://doi.org/10.1017/cbo9780511810763

Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, *53*(1–2), 23–69. https://doi.org/10.1023/a:1025667309714

Rodrigues, F., & Correia, H. (2023). Semi-supervised and ensemble learning to predict work-related stress. *Journal of Intelligent Information Systems*, *62*(1), 77–90. https://doi.org/10.1007/s10844-023-00806-z

Rokach, L. (2009). Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. *Computational Statistics & Data Analysis*, *53*(12), 4046–4072. https://doi.org/10.1016/j.csda.2009.07.017

Rolls, E. T. (2023). Emotion, motivation, decision-making, the orbitofrontal cortex, anterior cingulate cortex, and the amygdala. *Brain Structure and Function*, *228*(5), 1201–1257. https://doi.org/10.1007/s00429-023-02644-9

Rosidin, S., Muljono, Fajar Shidik, G., Zainul Fanani, A., Al Zami, F., & Purwanto. (2021). Improvement with chi square selection feature using supervised machine learning approach on COVID-19 data. *2021 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 32–36. https://doi.org/10.1109/isemantic52711.2021.9573196

Rosso, O. A., Blanco, S., Yordanova, J., Kolev, V., Figliola, A., Schürmann, M., & Başar, E. (2001). Wavelet entropy: A new tool for analysis of short duration brain electrical signals. *Journal of Neuroscience Methods*, *105*(1), 65–75. https://doi.org/10.1016/s0165-0270(00)00356-3

Rubinstein, Y. D., Hastie, T., et al. (1997). Discriminative vs informative learning. *Proceedings of the The Third International Conference on Knowledge Discovery and Data Mining*, *5*, 49–53.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, *76*(5), 805–819. https://doi.org/10.1037/0022-3514.76.5.805

Russell, J. A., & Mehrabian, A. (1974). Distinguishing anger and anxiety in terms of emotional response factors. *Journal of Consulting and Clinical Psychology*, *42*(1), 79–83. https://doi.org/10.1037/h0035915

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, *11*(3), 273–294. https://doi.org/10.1016/0092-6566(77)90037-x

Russo, V., Bilucaglia, M., Casiraghi, C., Chiarelli, S., Columbano, M., Fici, A., Rivetti, F., Rossi, C., Valesi, R., & Zito, M. (2023). Neuroselling: Applying neuroscience to selling for a new business perspective. an analysis on teleshopping advertising. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1238879

Russo, V., Bilucaglia, M., Circi, R., Bellati, M., Valesi, R., Laureanti, R., Licitra, G., & Zito, M. (2022). The role of the emotional sequence in the communication of the territorial cheeses: A neuromarketing approach. *Foods*, *11*(15), 2349. https://doi.org/10.3390/foods11152349

Russo, V., Milani Marin, L. E., Fici, A., Bilucaglia, M., Circi, R., Rivetti, F., & Zito, M. (2021). Strategic communication and neuromarketing in the fisheries sector: Generating ideas from the territory. *Frontiers in Communication*, *6*, 49. https://doi.org/10.3389/fcomm.2021.659484

Russo, V., Valesi, R., Gallo, A., Laureanti, R., & Zito, M. (2020). "the theater of the mind": The effect of radio exposure on TV advertising. *Social Sciences*, *9*(7), 123. https://doi.org/10.3390/socsci9070123

Rutkowska, J. M., Ghilardi, T., Vacaru, S. V., van Schaik, J. E., Meyer, M., Hunnius, S., & Oostenveld, R. (2024). Optimal processing of surface facial EMG to identify emotional expressions: A data-driven approach. *Behavior Research Methods*. https://doi.org/10.3758/s13428-024-02421-4

Saganowski, S., Komoszyńska, J., Behnke, M., Perz, B., Kunc, D., Klich, B., Kaczmarek, Ł. D., & Kazienko, P. (2022). Emognition dataset: Emotion recognition with self-reports, facial expressions, and physiology using wearables. *Scientific Data*, *9*(1). https://doi.org/10.1038/s41597-022-01262-0

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4). https://doi.org/10.1002/widm.1249

Salankar, N., Mishra, P., & Garg, L. (2021). Emotion recognition from EEG signals using empirical mode decomposition and second-order difference plot. *Biomedical Signal Processing and Control*, *65*, 102389. https://doi.org/10.1016/j.bspc.2020.102389

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

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

Sayed Ismail, S. N. M., Ab. Aziz, N. A., Ibrahim, S. Z., & Mohamad, M. S. (2024). A systematic review of emotion recognition using cardio-based signals. *ICT Express*, *10*(1), 156–183. https://doi.org/10.1016/j.icte.2023.09.001

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379–399. https://doi.org/10.1037/h0046234

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. In *Appraisal processes in emotion* (pp. 92–120). Oxford University PressNew York, NY. https://doi.org/10.1093/oso/9780195130072.003.0005

Scherer, K. R., & Moors, A. (2019). The emotion process: Event appraisal and component differentiation. *Annual Review of Psychology*, *70*(1), 719–745. https://doi.org/10.1146/annurev-psych-122216-011854

Schmalz, X., Biurrun Manresa, J., & Zhang, L. (2023). What is a bayes factor? *Psychological Methods*, *28*(3), 705–718. https://doi.org/10.1037/met0000421

Schmidt, P., Reiss, A., Dürichen, R., & Laerhoven, K. V. (2019). Wearable-based affect recognition—a review. *Sensors*, *19*(19), 4079. https://doi.org/10.3390/s19194079

Segri, G. (1894). *Principi di psicologia. dolore e piacere. storia naturale dei sentimenti*. Fratelli Duomo Editori.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17*(11), 565–573. https://doi.org/10.1016/j.tics.2013.09.007

Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, *5*, 258. https://doi.org/10.3389/fpubh.2017.00258

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, *104*(1), 148–175. https://doi.org/10.1109/jproc.2015.2494218

Sharma, A., & Canavan, S. (2021). Multimodal physiological-based emotion recognition. *Pattern Recognition. ICPR International Workshops and Challenges*, *12662*, 101–113. https://doi.org/10.1007/978-3-030-68790-8_9

Sharma, K., Trott, S., Sahadev, S., & Singh, R. (2023). Emotions and consumer behaviour: A review and research agenda. *International Journal of Consumer Studies*, *47*(6), 2396–2416. https://doi.org/10.1111/ijcs.12937

Sheikholeslami, S., Meister, M., Wang, T., Payberah, A. H., Vlassov, V., & Dowling, J. (2021). Autoablation: Automated parallel ablation studies for deep learning. *Proceedings of the 1st Workshop on Machine Learning and Systems*, 55–61. https://doi.org/10.1145/3437984.3458834

Shields, S. A., & Zawadzki, M. J. (2012). Theories of emotion, history of. In R. W. Rieber (Ed.), *Encyclopedia of the history of psychological theories - volume i* (pp. 1088–1105). Springer.

Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., & 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. (2021). 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

Siddiqui, M. F. H., Dhakal, P., Yang, X., & Javaid, A. Y. (2022). A survey on databases for multimodal emotion recognition and an introduction to the VIRI (visible and InfraRed image) database. *Multimodal Technologies and Interaction*, *6*(6), 47. https://doi.org/10.3390/mti6060047

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1409.1556

Sinha, S. R., Sullivan, L., Sabau, D., San-Juan, D., Dombrowski, K. E., Halford, J. J., Hani, A. J., Drislane, F. W., & Stecker, M. M. (2016). American clinical neurophysiology society guideline 1: Minimum technical requirements for performing clinical electroencephalography. *Journal of Clinical Neurophysiology*, *33*(4), 303–307. https://doi.org/10.1097/WNP.0000000000000308

Sinnott-Armstrong, W., & Simmons, C. (2021). Some common fallacies in arguments from m/EEG data. *NeuroImage*, *245*, 118725. https://doi.org/10.1016/j.neuroimage.2021.118725

Skaramagkas, V., Giannakakis, G., Ktistakis, E., Manousos, D., Karatzanis, I., Tachos, N., Tripoliti, E., Marias, K., Fotiadis, D. I., & Tsiknakis, M. (2021). Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Reviews in Biomedical Engineering*, *16*, 260–277. https://doi.org/10.1109/rbme.2021.3066072

Skaramagkas, V., Ktistakis, E., Manousos, D., Kazantzaki, E., Tachos, N. S., Tripoliti, E., Fotiadis, D. I., & Tsiknakis, M. (2023). Esee-d: Emotional state estimation based on eye-tracking dataset. *Brain Sciences*, *13*(4), 589. https://doi.org/10.3390/brainsci13040589

Smidts, A. (2002). *Kijken in het brein. over de mogelijkheden van neuromarketing* (tech. rep. No. 321). Erasmus Research Institute of Management.

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., Lichtenauer, J., Pun, T., & Pantic, M. (2012). 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

Song, T., Zheng, W., Lu, C., Zong, Y., Zhang, X., & Cui, Z. (2019). MPED: A multi-modal physiological emotion database for discrete emotion recognition. *IEEE Access*, *7*, 12177–12191. https://doi.org/10.1109/ACCESS.2019.2891579

Spezialetti, M., Placidi, G., & Rossi, S. (2020). Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, *7*. https://doi.org/10.3389/frobt.2020.532279

Stasi, A., Songa, G., Mauri, M., Ciceri, A., Diotallevi, F., Nardone, G., & Russo, V. (2018). Neuromarketing empirical approaches and food choice: A systematic review. *Food Research International*, *108*, 650–664. https://doi.org/10.1016/j.foodres.2017.11.049

Stern, J. A. (1964). Toward a definition of psychophysiology. *Psychophysiology*, *1*(1), 90–91. https://doi.org/10.1111/j.1469-8986.1964.tb02626.x

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

Tan, J.-W., Andrade, A. O., Li, H., Walter, S., Hrabal, D., Rukavina, S., Limbrecht-Ecklundt, K., Hoffman, H., & Traue, H. C. (2016). Recognition of intensive valence and arousal affective states via facial electromyographic activity in young and senior adults (L. Hadjileontiadis, Ed.). *PLOS ONE*, *11*(1), e0146691. https://doi.org/10.1371/journal.pone.0146691

Terven, J., Cordova-Esparza, D.-M., Romero-González, J.-A., Ramírez-Pedraza, A., & Chávez-Urbiola, E. A. (2025). A comprehensive survey of loss functions and metrics in deep learning. *Artificial Intelligence Review*, *58*(7). https://doi.org/10.1007/s10462-025-11198-7

Tharwat, A. (2016). Linear vs. quadratic discriminant analysis classifier: A tutorial. *International Journal of Applied Pattern Recognition*, *3*(2), 145–180. https://doi.org/10.1504/IJAPR.2016.079050

Thayer, J. F., & Lane, R. D. (2000). A model of neurovisceral integration in emotion regulation and dysregulation. *Journal of Affective Disorders*, *61*(3), 201–216. https://doi.org/10.1016/s0165-0327(00)00338-4

Thayer, R. E. (1989). *The biopsychology of mood and arousal*. Oxford University Press.

Theodoridis, S., & Koutroumbas, K. (2009). *Pattern recognition*. Elsevier.

Tian, Y., Zhang, Y., & Zhang, H. (2023). Recent advances in stochastic gradient descent in deep learning. *Mathematics*, *11*(3), 682. https://doi.org/10.3390/math11030682

Tomkins, S. S. (1962). *Affect imagery consciousness, vol. i: The positive affects*. Springer Publishing Company.

Tomkins, S. S. (1963). *Affect imagery consciousness, vol. II: The negative affects*. Springer Publishing Company.

Torres-Valencia, C., Alvarez-Lopez, M., & Orozco-Gutiérrez, A. (2016). SVM-based feature selection methods for emotion recognition from multimodal data. *Journal on Multimodal User Interfaces*, *11*(1), 9–23. https://doi.org/10.1007/s12193-016-0222-y

Trampe, D., Quoidbach, J., & Taquet, M. (2015). Emotions in everyday life (A. Avenanti, Ed.). *PLOS ONE*, *10*(12), e0145450. https://doi.org/10.1371/journal.pone.0145450

Traue, H., Ohl, F., Brechmann, A., Schwenker, F., Kessler, H., Limbrecht, K., Hoffmann, H., Scherer, S., Kotzyba, M., Scheck, A., & Walter, S. (2013). A framework for emotions and dispositions in man-companion interaction. In *Coverbal synchrony in human-machine interaction* (pp. 99–140). CRC Press. https://doi.org/10.1201/b15477-6

Tun, N., Sanuki, F., & Iramina, K. (2021). Electroencephalogram-electromyogram functional coupling and delay time change based on motor task performance. *Sensors*, *21*(13), 4380. https://doi.org/10.3390/s21134380

Valderrama, C. E., & Sheoran, A. (2025). Identifying relevant EEG channels for subject-independent emotion recognition using attention network layers. *Frontiers in Psychiatry*, *16*. https://doi.org/10.3389/fpsyt.2025.1494369 van Kleef, G. A., & Côté, S. (2022). The social effects of emotions. *Annual Review of Psychology*, *73*(1), 629–658. https://doi.org/10.1146/annurev-psych-020821-010855

Van Dongen, H. P. A., Olofsen, E., VanHartevelt, J. H., & Kruyt, E. W. (1999). Searching for biological rhythms: Peak detection in the periodogram of unequally spaced data. *Journal of Biological Rhythms*, *14*(6), 617–620. https://doi.org/10.1177/074873099129000984

Vapnik, V. (1991). Principles of risk minimization for learning theory. *Proceedings of the 4th International Conference on Neural Information Processing Systems (NIPS)*, 831–838.

Vapnik, V. (2000). *The nature of statistical learning theory*. Springer.

Vapnik, V. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999. https://doi.org/10.1109/72.788640

Vázquez, M. A., Maghsoudi, A., & Mariño, I. P. (2021). An interpretable machine learning method for the detection of schizophrenia using EEG signals. *Frontiers in Systems Neuroscience*, *15*. https://doi.org/10.3389/fnsys.2021.652662

Verma, G. K., & Tiwary, U. S. (2014). Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage*, *102*, 162–172. https://doi.org/10.1016/j.neuroimage.2013.11.007

Verzelli, P., Tchumatchenko, T., & Kotaleski, J. H. (2024). Editorial overview: Computational neuroscience as a bridge between artificial intelligence, modeling and data. *Current Opinion in Neurobiology*, *84*, 102835. https://doi.org/10.1016/j.conb.2023.102835

Von Luxburg, U., & Schölkopf, B. (2011). Statistical learning theory: Models, concepts, and results. In D. M. Gabbay, S. Hartmann, & J. Woods (Eds.), *Handbook of the history of logic - volume 10: Inductive logic* (pp. 651–706, Vol. 10). Elsevier. https://doi.org/10.1016/B978-0-444-52936-7.50016-1

Wang, C.-A., Baird, T., Huang, J., Coutinho, J. D., Brien, D. C., & Munoz, D. P. (2018). Arousal effects on pupil size, heart rate, and skin conductance in an emotional face task. *Frontiers in Neurology*, *9*, 1029. https://doi.org/10.3389/fneur.2018.01029

Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W., & Zhang, W. (2022). A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, *83-84*, 19–52. https://doi.org/10.1016/j.inffus.2022.03.009

Wang, Z., & Wang, Y. (2025). Emotion recognition based on multimodal physiological electrical signals. *Frontiers in Neuroscience*, *19*, 1512799. https://doi.org/10.3389/fnins.2025.1512799

Wassmann, C. (2008). Physiological optics, cognition and emotion: A novel look at the early work of wilhelm wundt. *Journal of the History of Medicine and Allied Sciences*, *64*(2), 213–249. https://doi.org/10.1093/jhmas/jrn058

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), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, *8*(7), 1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341

Yang, W., Wang, K., & Zuo, W. (2012). Neighborhood component feature selection for high-dimensional data. *Journal of Computers*, *7*(1). https://doi.org/10.4304/jcp.7.1.161-168

Yoon, C., Gutchess, A. H., Feinberg, F., & Polk, T. A. (2006). A functional magnetic resonance imaging study of neural dissociations between brand and person judgments. *Journal of Consumer Research*, *33*(1), 31–40. https://doi.org/10.1086/504132

Yoonesi, S., Abedi Azar, R., Arab Bafrani, M., Yaghmayee, S., Shahavand, H., Mirmazloumi, M., Moazeni Limoudehi, N., Rahmani, M., Hasany, S., Idjadi, F. Z., Aalipour, M. A., Gharedaghi, H., Salehi, S., Asadi Anar, M., & Soleimani, M. S. (2025). Facial expression deep learning algorithms in the detection of neurological disorders: A systematic review and meta-analysis. *BioMedical Engineering OnLine*, *24*(1). https://doi.org/10.1186/s12938-025-01396-3

Younis, E. M. G., Mohsen, S., Houssein, E. H., & Ibrahim, O. A. S. (2024). Machine learning for human emotion recognition: A comprehensive review. *Neural Computing and Applications*, *36*(16), 8901–8947. https://doi.org/10.1007/s00521-024-09426-2

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, 818–833. https://doi.org/10.1007/978-3-319-10590-1_53

Zhan, Y., Yang, R., You, J., Huang, M., Liu, W., & Liu, X. (2025). A systematic literature review on incomplete multimodal learning: Techniques and challenges. *Systems Science & Control Engineering*, *13*(1). https://doi.org/10.1080/21642583.2025.2467083

Zhang, J. M., Harman, M., Ma, L., & Liu, Y. (2022). Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, *48*(1), 1–36. https://doi.org/10.1109/tse.2019.2962027

Zhang, L., Walter, S., Ma, X., Werner, P., Al-Hamadi, A., Traue, H. C., & Gruss, S. (2016). BioVid Emo DB: A multimodal database for emotion analyses validated by subjective ratings. *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–6. https://doi.org/10.1109/SSCI.2016.7849931

Zhang, X., Zhang, T., Sun, L., Zhao, J., & Jin, Q. (2025). Exploring interpretability in deep learning for affective computing: A comprehensive review. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *21*(7), 1–28. https://doi.org/10.1145/3723005

Zhao, F., Zhang, C., & Geng, B. (2024). Deep multimodal data fusion. *ACM Computing Surveys*, *56*(9), 1–36. https://doi.org/10.1145/3649447

Zhao, J., Li, R., & Jin, Q. (2021). Missing modality imagination network for emotion recognition with uncertain missing modalities. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2608–2618. https://doi.org/10.18653/v1/2021.acl-long.203

Zhao, S., Jia, G., Yang, J., Ding, G., & Keutzer, K. (2021). Emotion recognition from multiple modalities: Fundamentals and methodologies. *IEEE Signal Processing Magazine*, *38*(6), 59–73. https://doi.org/10.1109/MSP.2021.3106895

Zheng, F., Zhang, G., & Song, Z. (2001). Comparison of different implementations of MFCC. *Journal of Computer Science and Technology*, *16*(6), 582–589. https://doi.org/10.1007/bf02943243

Zheng, W.-L., Zhu, J.-Y., & Lu, B.-L. (2017). Identifying stable patterns over time for emotion recognition from EEG. *IEEE Transactions on Affective Computing*, *10*(3), 417–429. https://doi.org/10.1109/TAFFC.2017.2712143

Zito, M., Fici, A., Bilucaglia, M., Ambrogetti, F. S., & Russo, V. (2021). Assessing the emotional response in social communication: The role of neuromarketing. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.625570