2 Analysing Embodied Emotions through Artificial Psychology

Nataša Kovač, Faculty of Applied Sciences, University of Donja Gorica, Oktoih 1, Podgorica, 81000, Montenegro, https://orcid.org/ [0000-0002-6671-2938](https://orcid.org/0000-0002-6671-2938)

Hojjatollah Farahani, Department of Psychology, Faculty of Humanities, Tarbiat Modares University, Tehran, Iran, <https://orcid.org/0000-0002-9799-7008>

Peter Watson, Cognition and Brain Sciences Unit, University of Cambridge, 15 Chaucer Road, Cambridge, CB2 7EF, United Kingdom, https://orcid.org/0000-0002-9436-0693

**Abstract:** This chapter begins with an analysis of what we have learned about the interfaces between neuroscience and emotions. This topic provides key lessons for the study and understanding of Embodied Emotions, as it considers the pivotal role of neuroscientific research methods for the study of emotions. Then, the relevance of biomarkers for the study of Embodied Emotions is discussed, emphasising the breakthroughs and thresholds implicated in the search for biomarkers. Moreover, it lays a foundational understanding of Python for delving deeper into Emotional Artificial Intelligence, and provides the critical steps involved in preparing data for the comprehensive analysis of embodied emotions. Beginning with the transition from raw data to embodied emotion insights, it explores various data acquisition techniques tailored for modelling emotions within AI. The chapter then proceeds to address the essential stages of data cleaning and feature extraction, ensuring the quality and relevance of emotional data. It extends its focus to specific data types, including time series, visual, audio, and text data, each requiring unique preprocessing techniques to uncover emotional insights within. Additionally, the chapter discusses gaze data analysis as a vital component of embodied emotion recognition and explores the power of multimodal data fusion, enabling a holistic understanding of emotions as they manifest across different data modalities.

**Keywords:** Time series data; audio signals; text data; gaze data;

# 2.1 Introduction

Every emotion begins with a signal but that signal, whether a flicker in an EEG trace, a shift in gaze, a subtle muscular tension, or a whisper in the voice, is rarely clean. It arrives tangled in noise, in missing fragments, in the idiosyncrasies of bodies and machines alike. To extract meaning from this cacophony, we must first prepare the data to hear what the emotion might be saying.

Over recent years, the study of human emotion has undergone a profound conceptual transformation. Rather than treating affect as a purely cognitive or physiological phenomenon, researchers are increasingly viewing it as an embodied process, an inseparable interplay between brain, body, and environment. The past five years (2020–2025) have seen the scientific community definitively move past traditional boundaries of cognition and physiology, embracing an embodied perspective that grounds emotion in the dynamic interchange between the brain, the body, and the environment (Pinna & Edwards, 2020). This transformation establishes the foundation of Artificial Psychology (AP), a discipline dedicated to modelling and replicating human emotional processes through computational mechanisms that account for the body's active role.

Central to this paradigm is interoception, the brain’s continuous sensing and interpretation of internal bodily states, such as heart rate and respiration. A synthesis of systematic reviews confirms that interoception provides the pre-reflective ground for affective awareness (Klein et al., 2025). High interoceptive accuracy correlates with greater emotional intensity and improved regulatory control, while deficits are commonly implicated in psychological challenges, underscoring the body's role as an emotional generator, not merely a vessel for expression (Pinna & Edwards, 2020).

Complementing internal sensing, the external bodily expression of emotion, posture, gesture, and movement is confirmed as a universal channel for social communication. Systematic analyses of kinematic data show that specific, measurable patterns, such as expansive postures associated with joy or rigid postures linked to fear, reliably predict emotional valence across diverse groups (Lin & Zhang, 2022; Dael et al., 2012). These findings confirm that the emotional state is readable through the entire physical form, reinforcing the adage that we feel with our bodies as much as with our minds.

This holistic understanding provides the operational blueprint for contemporary AP and Affective Computing (AC). Multimodal deep learning systems now achieve superior accuracy by directly integrating physiological signals, posture, and facial expressions, successfully mimicking the human brain's own integrative, embodied processing of affect (Lin & Zhang, 2022; Wu, Mi, & Gao, 2025). This fusion of data streams goes beyond simple facial recognition to include the subtle cues from the entire physical self. It allows artificial systems to interact with humans with unprecedented nuance and contextual awareness (Wu, Mi, & Gao, 2025).

The clinical applications of this framework are already transformative. Systematic reviews of AI-driven mental health interventions, such as conversational agents and Extended Reality (XR) therapies, show promising results in achieving symptom reduction (Botes, 2025). The next generation of these tools is moving towards true Embodied AI (EAI), leveraging virtual agents and immersive environments that respond not just to user dialogue, but to their posture and psychophysiological state, thereby creating more emotionally resonant and therapeutically effective interactions.

Despite these technical advances, the integrative reviews identify persistent, human-centred challenges that define the next phase of research. A primary concern is the potential for algorithmic bias stemming from a lack of cross-cultural diversity in the datasets used to train emotion recognition systems (Barker et al., 2025). Suppose EAI is to realise its potential in diverse global contexts. In that case, researchers must prioritise cultural inclusivity and ethical development to prevent the misinterpretation or misrepresentation of embodied emotional cues in non-Western populations (Lin & Zhang, 2022). Furthermore, clinical reviews call for increased methodological rigour and long-term clinical validation to ensure the sustained effectiveness and ethical sustainability of these EAI interventions (Botes, 2025). The future of AP rests not just on perfecting the technical replication of emotion, but on ensuring that its models are scientifically sound and genuinely reflective of the full, diverse spectrum of human embodied experience.

# 2.2 Python Basics for Emotional Artificial Intelligence

We are recognising that artificial intelligence (AI) significantly impacts the study of human emotion by enabling the interpretation of complex physiological signals, facial expressions, vocal cues, and text data, a crossover often described as Artificial Psychology (PsAIchology) (Farahani et al., 2024). Analysing this embodied emotional information requires equally advanced computer tools for researchers and practitioners. To fully unlock the potential of emotional artificial intelligence, robust computing tools capable of managing, analysing, and visualising emotionally rich data are essential. Due to its simplicity, readability, and comprehensive libraries, Python has become the dominant programming language in this interdisciplinary field (VanderPlas, 2016). Python has become a leading language in emotional AI applications for several convincing reasons. Its straightforward syntax closely resembles English, making it accessible even for newcomers from psychology, neuroscience, or other non-technical fields. Beneath its simplicity, Python offers advanced computational capabilities, from statistical analysis and signal processing to sophisticated machine learning and deep learning techniques (Kovač et al., 2024).

Python offers a broad array of open-source libraries that enable researchers to rapidly prototype, analyse, visualise, and deploy advanced emotional models. These libraries manage various aspects of the emotional AI pipeline, from initial data management and preprocessing to emotional feature extraction and visualisation, culminating in the deployment of robust predictive models. Consequently, the successful implementation of Emotional AI with Python requires understanding and effective utilisation of its key libraries:

1. **NumPy** provides robust support for numerical computation and managing arrays of physiological and behavioural data. Its ability to handle large-scale numeric data makes it suitable for emotional analysis tasks involving EEG, heart rate, and other biometric signals:

import numpy as np

eeg\_data = np.random.normal(0, 1, 5000)

mean\_eeg = np.mean(eeg\_data)

print("Mean EEG signal:", mean\_eeg)

1. **Pandas** simplifies data cleaning and preprocessing, which are vital steps in ensuring data quality. Researchers frequently handle datasets from multiple emotional modalities, requiring structured dataframes to simplify complex analysis pipelines:

import pandas as pd

emotions\_df = pd.DataFrame({

"Subject": [1, 2, 3],

"Emotion": ["happy", "sad", "neutral"],

"Intensity": [7, 5, 3]

})

print(emotions\_df)

1. Visualisation is essential for intuitively interpreting emotional data. **Matplotlib** and **Seaborn** convert raw numerical data into meaningful visuals, such as graphs or heatmaps, making it easier to understand emotional patterns:

import seaborn as sns

import matplotlib.pyplot as plt

intensity = np.random.rand(100)

sns.histplot(intensity, kde=True, color="green")

plt.title("Distribution of Emotional Intensity")

plt.show()

1. Voice carries rich emotional signals. **Librosa** enables easy extraction of emotional audio features like pitch, rhythm, and timbre, essential for understanding affective vocal characteristics:

import librosa

y, sr = librosa.load("voice\_emotion.wav")

tempo, \_ = librosa.beat.beat\_track(y, sr=sr)

print("Detected tempo:", tempo, "BPM")

1. Facial expression analysis involves detecting and interpreting subtle movements. Libraries such as **OpenCV** and **dlib** offer powerful tools to detect faces, track facial landmarks, and extract emotion-sensitive features:

import cv2

import dlib

detector = dlib.get\_frontal\_face\_detector()

image = cv2.imread("face.jpg")

gray = cv2.cvtColor(image, cv2.COLOR\_BGR2GRAY)

faces = detector(gray)

print("Number of faces detected:", len(faces))

1. Emotion recognition and prediction rely heavily on machine learning, and similar supervised pipelines are increasingly used across psychological health domains, including the prediction of chronic pain severity (Kovač et al., 2025). **Scikit-learn** provides an easy-to-use toolkit for traditional machine learning, while **TensorFlow** supports sophisticated deep learning methods. A complementary line of work uses calibrated regression to predict internal shame, with XGBoost performing best and distress tolerance emerging as the strongest predictor (Kovač, Ratković, Farahani, & Watson, 2025b).

from sklearn.ensemble import RandomForestClassifier

X = np.random.rand(50, 10)

y = np.random.randint(0, 2, 50)

model = RandomForestClassifier()

model.fit(X, y)

accuracy = model.score(X, y)

print("Emotion classification accuracy:", accuracy)

These libraries create a cohesive framework, enabling researchers to move from raw emotional data collection through advanced analysis to meaningful emotional insights. By mastering these core tools, readers will be well-prepared to navigate and apply powerful emotional AI applications effectively.

In addition to technical knowledge of Python, it is important to note that ethical considerations are equally important. Emotional AI often deals with sensitive and personal data, which requires careful handling to protect the privacy and dignity of individuals. Python’s transparent and readable code, as well as its commitment to open source, encourage ethical clarity, transparency, and accountability in research and applications (Jobin, Ienca, & Vayena, 2019). We believe that researchers, in addition to regularly documenting Python code, should also maintain data privacy protocols, secure consent, and implement strict anonymisation procedures to respect participant confidentiality.

Embarking on Python programming for emotional AI might seem daunting initially. Yet, the supportive community, abundance of tutorials, and accessible documentation dramatically reduce entry barriers. Beginners should start by mastering basic Python syntax, then gradually explore libraries, and eventually move into specialised emotion-focused tools (Librosa, OpenCV, dlib), and machine learning libraries (Scikit-learn, TensorFlow). Through consistent practice and incremental learning, even those without prior technical backgrounds can achieve proficiency, enabling sophisticated analyses of complex emotional phenomena.

# 2.3 From Data to Embodied Emotion

Emotion is not stored in a single variable. It pulses through muscle tensions, breath rhythms, glance trajectories, and vocal microfluctuations. Each of these signals can be digitised, sampled, measured, annotated, and transformed into features that AI systems can understand. The journey from unstructured signals to computational representations is neither trivial nor purely technical. It demands sensitivity to the psychological context, to the nature of embodiment, and to the epistemological assumptions we make when we convert movement into numbers.

Capturing embodied emotion is more than placing sensors on a participant’s body or turning on a camera. It is a philosophical commitment, a methodological stance, and a technical operation wrapped into one. The data we acquire defines the boundaries of what we can know, what we can model, and what we are willing to assume about the nature of emotion itself. If we record only facial expressions, we risk privileging visible emotion while ignoring the body's internal story. If we focus only on brain signals, we may interpret affect as neurological when it is equally social, physiological, and behavioural.

At its core, data preprocessing for embodied emotion begins with a commitment to **integrity**, not just of the data, but of its meaning. This involves:

* **Retaining affective variance** while removing irrelevant noise,
* **Aligning multimodal sources** without distorting temporal dynamics,
* **Extracting features** that have interpretive resonance in both psychological theory and algorithmic performance.

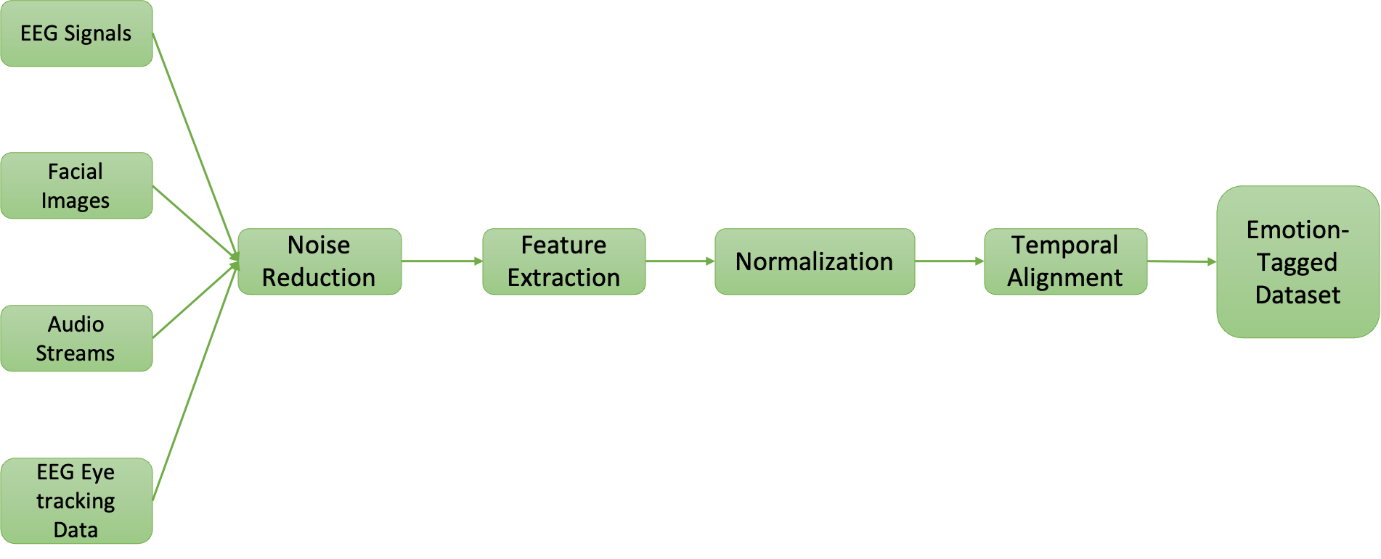


Figure 1. From data to embodied emotion

Figure 1 encapsulates the logic of embodiment. Every arrow hides hundreds of decisions: how to threshold noise, what interpolation strategy to use, how to respect time when aligning sources.

What we call preprocessing is, in essence, a filtering of the world. The data we collect is always partial, always noisy, always filtered through human design choices. As Gendron and Barrett (2009) have argued, *emotion data is not found; it is made*. The act of preprocessing is therefore both statistical and philosophical: it frames which aspects of human experience are legible to machines.

Recent work in embodied AI (Calvo & D’Mello, 2010; Schuller & Batliner, 2013) has emphasised the need for **context-aware data pipelines**, where preprocessing is guided by the psychological properties of the emotion under study. For example, preprocessing gaze data for disgust differs vastly from preprocessing voice data for sadness, not just in technique, but in what counts as a meaningful signal.

Capturing embodied emotion is more than placing sensors on a participant’s body or turning on a camera. It is a philosophical commitment, a methodological stance, and a technical operation wrapped into one. The data we acquire defines the boundaries of what we can know, what we can model, and what we are willing to assume about the nature of emotion itself. If we record only facial expressions, we risk privileging visible emotion while ignoring the body's internal story. If we focus only on brain signals, we may interpret affect as neurological when it is equally social, physiological, and behavioural.

No single sensor can fully capture the complexity of an emotional experience. That complexity unfolds across multiple dimensions, such as facial muscle movements, vocal inflexions, gaze shifts, heart rhythms, and even the electrical patterns of cortical oscillations. Each of these dimensions offers a different window into emotional processing, and each requires a distinct mode of capture. The question, then, is not whether one should use EEG or voice data or eye tracking. Rather, it is how one chooses from these modalities based on the **psychological construct under investigation**. For example, if the research focuses on emotional reactivity in social contexts, gaze data might reveal patterns of approach or avoidance. In contrast, if one is examining affect regulation, heart rate variability, or frontal alpha asymmetry could provide more meaningful insight. Before data collection, researchers have to recognise the nature of the emotion signal and the limitations each sensor brings. Sensors do not simply record; they interpret. Physiological signals have long been central in emotion research, largely due to their involuntary nature. Unlike self-report or facial expressions, bodily reactions often escape conscious modulation, offering a more reliable path to affective truth. At least in theory.

**Electroencephalography (EEG)** remains one of the most frequently used methods for investigating the brain’s dynamic response to emotional stimuli. Capable of capturing millisecond-level changes in electrical activity across the scalp, EEG provides a lens into attentional focus, affective valence, and cognitive engagement. For example, the asymmetry of alpha band power over the frontal cortex has been repeatedly associated with the direction of affect. Greater left-side activity is often interpreted as indicative of approach-related emotions, while right-side activity may signal avoidance or withdrawal tendencies (Davidson, 2004; Coan & Allen, 2004).

However, EEG is sensitive to a wide range of artefacts, including eye movements, muscle contractions, and environmental electrical interference. This sensitivity demands careful preparation: high-conductivity electrodes, proper placement using the 10–20 system, and often a quiet, shielded room. While dry electrode systems like OpenBCI allow for greater mobility, they may sacrifice some signal quality when compared to gel-based systems such as the *g.HIamp* from *g.tec*.

Another commonly used physiological modality is **electrodermal activity (EDA)**, sometimes referred to as galvanic skin response (GSR). This signal measures the skin’s ability to conduct electricity, which varies with sweat gland activity and is linked to sympathetic nervous system arousal. Emotional stimuli typically provoke transient spikes in skin conductance. These can be used to detect engagement, stress, or fear. Devices such as the Empatica E4 or Shimmer GSR+ make real-time measurement of EDA feasible, even in ambulatory contexts.

**Cardiac activity** provides deeper insight into the regulation and pacing of emotional states. Standard metrics include heart rate (HR) and heart rate variability (HRV), which are typically recorded using electrocardiography (ECG) or photoplethysmography (PPG). HRV in particular has gained attention as an index of emotional resilience and self-regulation, where greater variability is linked to flexible adaptation in emotionally challenging scenarios. While ECG remains the gold standard for capturing precise inter-beat intervals, wearable PPG devices (e.g., Polar H10, Biopac, or Empatica’s wrist-worn monitors) offer a reasonable compromise between accuracy and ecological practicality.

Each of these physiological signals carries a temporal footprint that shapes how it must be collected. EEG, with its high temporal resolution, demands high sampling rates (256–1024 Hz), while EDA and HRV are often sampled more modestly (10–100 Hz). Understanding these constraints is essential for multimodal integration, which we discuss later.

Where physiology speaks in rhythms, **behavioural signals speak in patterns**. These include overt facial expressions, vocal prosody, gestures, and gaze dynamics. They are often the most visible aspects of emotion, and also the most culturally and contextually mediated.

**Facial expressions** are frequently treated as the canonical medium of emotional communication. Decades of work, especially from Ekman and Friesen (1978), have led to standardised coding systems such as the Facial Action Coding System (FACS). However, manual FACS annotation is time-consuming and often inconsistent across annotators. As a result, automated systems like OpenFace or commercial solutions like FaceReader have become popular. These systems analyse video data frame-by-frame to extract facial action units (AUs) and head pose metrics, providing continuous estimates of expression intensity. The accuracy of these tools is influenced by camera placement, lighting conditions, frame rate, and even the participant’s facial morphology. Thus, researchers must attend not just to the software they use but also to the physical design of the recording environment.

In parallel, **voice** has emerged as a powerful channel for emotion recognition. Unlike facial expressions, vocal emotion may be detectable even in the absence of visual cues. Acoustic features such as pitch (fundamental frequency), jitter, shimmer, formant structure, and Mel-frequency cepstral coefficients (MFCCs) form the backbone of most vocal affect models. Python libraries like librosa enable researchers to extract these features and visualise affective contours in speech.

The **gaze** offers another entry point into the affective state. Eye-tracking systems such as Tobii Pro or Pupil Labs record where and for how long a person fixates during a task. Metrics like fixation duration, saccade velocity, and blink rate can signal arousal, cognitive load, and social interest. Importantly, gaze also reflects the structure of attention, a process tightly intertwined with emotional appraisal.

Emotion-aware research paradigms often benefit from combining these behavioural modalities. A spoken word paired with a gaze aversion may imply sarcasm, discomfort, or shame. Facial tension combined with a flat voice may reveal suppressed anger. Without multimodal capture, such subtleties remain invisible. When capturing multiple emotion signals, the challenge becomes **integration**. This doesn’t mean simply recording everything at once, but ensuring that all data streams are meaningfully aligned. A multimodal setup might include an EEG system, a video camera for facial tracking, a microphone for voice capture, a wristband for skin conductance, and a wearable for HRV. But if these devices are not synchronised in time, their combined interpretive power collapses. There are three principal methods used to synchronise emotion data streams. The first involves **hardware triggers**, where physical signals such as TTL pulses are sent to each device at known timepoints. These serve as temporal anchors, allowing all recorded signals to align with precision down to the millisecond. Though highly accurate, this approach demands technical expertise and device compatibility, making it best suited for lab-based studies.

A more adaptable method relies on a shared timing protocol implemented via **middleware**, such as the **Lab Streaming Layer (LSL)**. LSL offers a software-based infrastructure to time-align data streams from different hardware sources. It does so by synchronising each data source to a master clock, usually the host computer’s system time, and then recording precise timestamps for every sample. This allows researchers to collect EEG, EDA, eye tracking, audio, and video within the same experimental timeline. Unlike TTL triggers, LSL is scalable and device-agnostic, making it especially valuable for real-time applications and mobile research environments.

In cases where neither hardware triggers nor shared clocks are available, researchers may resort to **software-based post hoc alignment**. This method involves identifying shared events and then interpolating timestamps across devices. While less precise, this approach allows for flexible integration of heterogeneous data sources, particularly when using commercial tools with limited interoperability. Each synchronisation method embodies a different perspective on what emotion research prioritises: precision, scalability, or adaptability. The essential task is to choose an integration strategy that aligns with both the scientific question and the practical constraints of the data collection environment.

No matter how advanced a multimodal system becomes, it remains an artificial construct unless designed to honour the **ecological validity** of emotion. Emotions are not confined to sterile labs. They occur while navigating a street, receiving a text, standing before an audience, or lying awake at night. Capturing them authentically requires extending our sensors into these lived contexts. Wearable EEG caps, smartwatches, smartphone microphones, and unobtrusive video recording can bring emotion research into the wild. But doing so comes at a cost: more noise, more missing data, more unpredictable behaviour. At the same time, this data is *truer*, messier, yes, but closer to how emotion is experienced and expressed in natural life.

With this ecological expansion comes **ethical responsibility**. Data collected from the body must be handled with exceptional care. Participants should be informed not only about what is recorded, but also what may be inferred. Researchers must ensure that the data pipeline respects privacy and avoids reductionist or pathologising interpretations.

Emotion is not a biometric. It is a lived, layered, socially negotiated experience. Any attempt to capture it must do justice to this complexity. To acquire emotion is not to extract it; it is to **co-create a setting** where the body speaks, and the machine listens. The setup, sensors, and protocols we choose shape what counts as emotion in our datasets. They shape what AI will learn as emotion. This responsibility cannot be offloaded to the hardware engineer or outsourced to a software package. It is the ethical, epistemological, and psychological foundation of every model we build.

# 2.4 Data Cleaning and Feature Extraction for Emotion Analysis

Once the raw emotional data has been collected through multimodal sensors, the next vital step involves transforming these raw, often chaotic streams of information into meaningful inputs for computational models. This stage, known as data cleaning and feature extraction, greatly affects the quality and accuracy of emotional insights obtained from computational analyses. Figure 2 shows a high-level schematic of how physiological signals (EEG, EDA, HRV) and behavioural signals (voice, facial expressions, gaze) flow into the preprocessing stack: noise reduction/filtering, feature engineering grounded in theory, and ethical compliance checks that govern every stage.



Figure 2. Data cleaning and feature extraction for embodied emotion

Real-world emotion data, whether recorded in a controlled laboratory or collected via mobile sensors, inevitably contains noise and artefacts. These disturbances can result from sensor limitations, environmental interferences, human movements, or even physiological activities unrelated to the emotion of interest. Ignoring such contamination not only diminishes the accuracy of subsequent analyses but may also lead to fundamentally incorrect emotional interpretations (Soleymani et al., 2017; Aguilera et al., 2023).

Data cleaning, therefore, serves as an essential safeguard which can carefully discriminate between authentic emotional signals and unwanted artefacts. Such processes include filtering, artefact removal, baseline correction, and temporal synchronisation, each performed with careful attention to retaining genuine emotional variability (Calvo & D'Mello, 2010). Physiological signals require particularly meticulous cleaning strategies due to their high susceptibility to noise and artefacts. EEG signals (Figure 3), for example, typically undergo preprocessing steps such as band-pass filtering, artefact detection (eye-blinks, muscle movements), and epoch segmentation (Davidson, 2004; Delorme & Makeig, 2004).

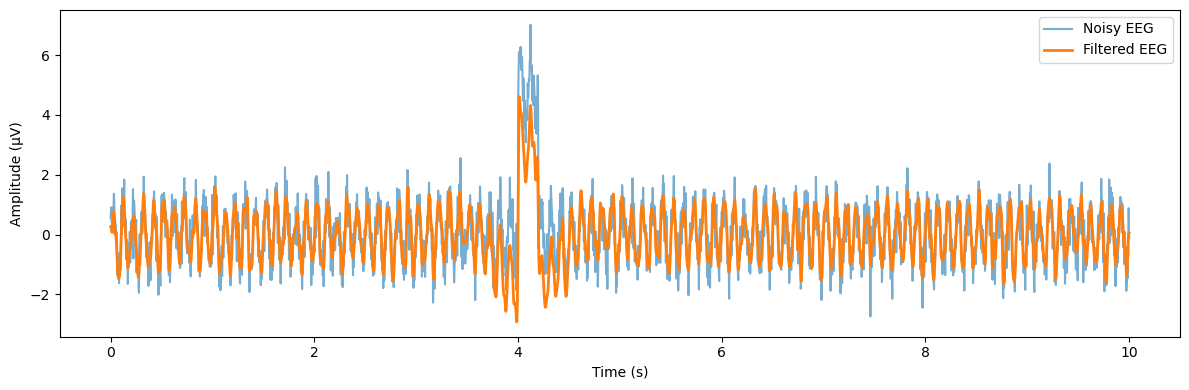


Figure 3. EEG signal cleaning: before and after filtering

Light Blue Line contains electrical noise: spikes, irregular patterns, and other disruptions that are not from the brain. After signal cleaning (orange line), the waves are more stable and clearer. This makes it easier to detect patterns associated with different emotional or cognitive states. Just like a blurry photo needs sharpening, an EEG needs cleaning before we can extract meaningful features like alpha power or emotional arousal indicators.

Similarly, Electrodermal Activity (EDA) or GSR data cleaning typically includes smoothing (Figure 4) and baseline removal, aimed at isolating event-related peaks from noise (Boucsein, 2012). Figure 4 shows the original recording as a blue line, which contains electrical noise and other spikes not related to real physiological changes. After applying filters, the signal is smoother, and the peaks (which indicate moments of arousal) become clearer. Cleaning ensures that we're not misinterpreting noise as emotional arousal. Only cleaned data should be used in modelling real emotional responses.

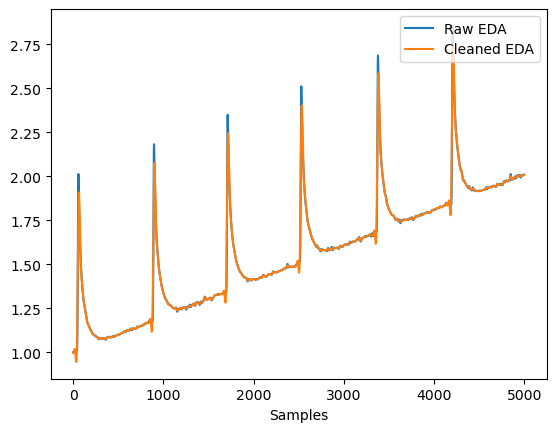


Figure 4. Raw vs. cleaned EDA signal

After ensuring the cleanliness of the data, the next step is feature extraction, performed to select and calculate attributes that meaningfully represent emotional states. Effective feature extraction bridges raw data and theoretical emotion models, allowing computational methods to more accurately infer affective states (Scherer, 2009). Key strategies for feature extraction include:

* **Temporal features** (mean, variance, peaks, valleys);
* **Spectral features** (power in EEG bands, MFCCs for voice);
* **Event-based features** (latency and amplitude of emotion-relevant events).

For instance, EEG-based features commonly include alpha-band asymmetry (Figure 5), known to be indicative of emotional valence (Davidson, 2004), and event-related potentials (ERPs) reflecting attention and cognitive-emotional appraisal.

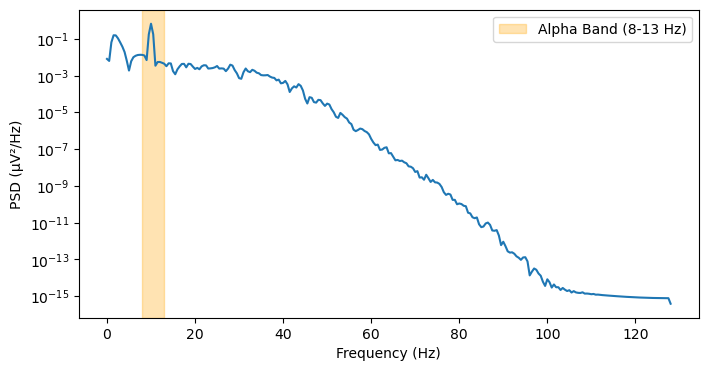


Figure 5. EEG spectral power for alpha band extraction

This plot displays how much *power* (signal strength) exists in various frequency ranges of EEG data. Power Spectral Density (PSD) measures how strong each brainwave frequency is. Frequency (Hz) ranges from slow waves (e.g., delta) to fast waves (e.g., gamma). The highlighted region (orange band) focuses on Alpha waves (8–13 Hz), often linked to relaxation and calm attention. Different emotions can shift the balance between these frequencies. For example, increased beta power might suggest tension, while strong alpha may suggest serenity. Extracting frequency-domain features is a core task in emotional state classification from EEG.

Voice data offers rich emotional information, embedded in pitch variations, loudness fluctuations, and spectral contours. Features such as Mel-frequency cepstral coefficients (MFCCs) (see Figure 6), pitch (fundamental frequency), and speech rate are robust predictors of emotional states (Schuller & Batliner, 2013).

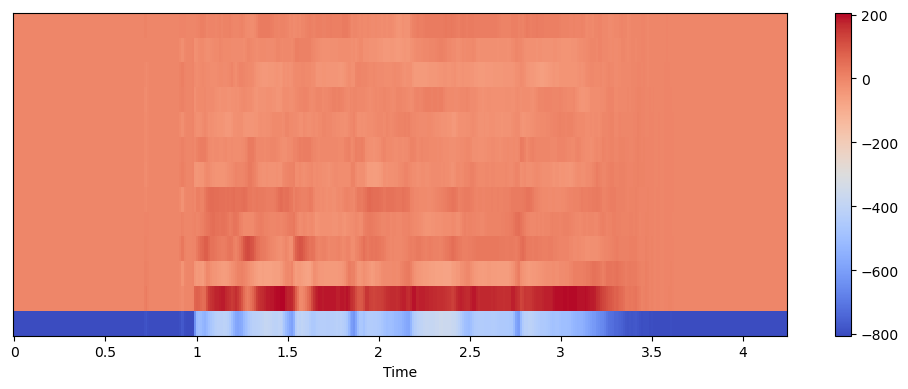


Figure 6. MFCCs for vocal emotion analysis

The plot in Figure 6 may look abstract, but it's a powerful way to visualise how the sound of a person’s voice carries emotional meaning. MFCC Values Over Time are features extracted from the voice recording and are sensitive to pitch, tone, and other properties that convey affect. Blue areas show lower energy, while red areas show higher energy. Emotions like anger, fear, or calmness change how we speak. MFCCs are used by computers to *hear* emotions even if they don’t understand language. They’re widely used in speech emotion recognition models.

The moment we choose features, we decide what kind of *listener* our model will be. A raw EEG trace or an audio waveform is too unruly to reason with directly; features translate those messy biographies of the body into compact, comparable descriptors. But in affective computing, features are never just arithmetic. They crystallise a theory about how appraisal unfolds in time, how regulation shows up across channels, and how context reshapes display. Put differently: a feature set is a set of commitments (Can, Mahesh, & André, 2023).

A good place to begin is by imagining the kinds of stories we want our system to understand. Some are **temporal**: a quick spike in sympathetic arousal, a slow return to the baseline after a social stressor, an episode that flickers more than it roars. Others are **spectral**: a shift in EEG alpha, a change in voice fundamental frequency (F0) and spectral tilt, a modulation in the low–high balance of heart-rate variability. Some are **event-related**: the way a skin-conductance response blooms a second or two after an appraisal, or the way an ERP component indicates attention. Still others are **spatial or kinematic**: the micro-geometry of facial action units (AUs), head pitch when a person leans in, the peculiar rhythm of a hurried gait. What makes this more than taxonomy is that each family of features embodies a different psychological hypothesis: arousal occurs in rapid sympathetic bursts and parasympathetic withdrawal; approach–withdrawal appears as frontal EEG asymmetry; valence often influences prosody and facial expression; regulation modifies how much the face *tells* compared to the body or the autonomic system (Gross, 2015; Can et al., 2023).

Figure 7 visualises this landscape. It is deliberately compact: one view, many pathways. You can read it in any direction. Start at *Feature Space* and move toward *Temporal* if your phenomena are short and phasic; follow *Spectral* if your signal-to-noise is better in frequency than in time; walk the *Cross-modal* edge if you expect divergence between what people show and what they feel.



Figure 7. Feature taxonomy for embodied emotion signals

The question that follows is not **which features are best**? but **which features are faithful to *this* study’s assumptions and ecology?** In a lab study with clean stimulus onsets, event-locked windows can support ERP-like EEG features and precise skin-conductance response (SCR) metrics. In the wild, where stimuli are mundane and irregular, sliding windows are more appropriate: you let the model learn *when* the signal changes without forcing it to align to an artificial timeline. The same logic governs window length. If you care about quick surges (think startle or social friction), you need short windows and generous overlap. If you are modelling mood drift over an hour, long windows that smooth local turbulence make more sense. Even the best feature can underperform if it inhabits the wrong timescale (Can et al., 2023).

There is also a humbler point: **redundancy is a feature, not a flaw, when it crosses modalities**. People regulate. Suppression flattens the face yet leaves EDA active; reappraisal can quiet autonomic arousal while leaving a fleeting frown. If you rely on a single channel, you are essentially trusting a single *witness*. Fusing voice with EDA, or head pose with HRV, makes the system resilient when one witness goes quiet (Gross, 2015). This is why, in practice, we often keep a small, interpretable set from each modality rather than eking out a marginal gain from an exotic transform in just one.

How do we check that a chosen vocabulary still matches the data we collected? Not by faith. A simple, revealing exercise is to look at the geometry of your feature table before you train anything heavy. In Figure 8, we project a windowed feature matrix into the first two principal components. Even though the example is didactic, the practice scales: a low-dimensional plot often exposes drift (e.g., early sessions vs. late sessions), confounds (e.g., motion vs. speaking), or the gratifying separation of conditions that tells you the representation is on the right track (see Figure 8). When the plot shows only a cloud, it is not a failure; it is an instruction: adjust windows, revisit normalisation, or prune features that are louder than they are helpful.

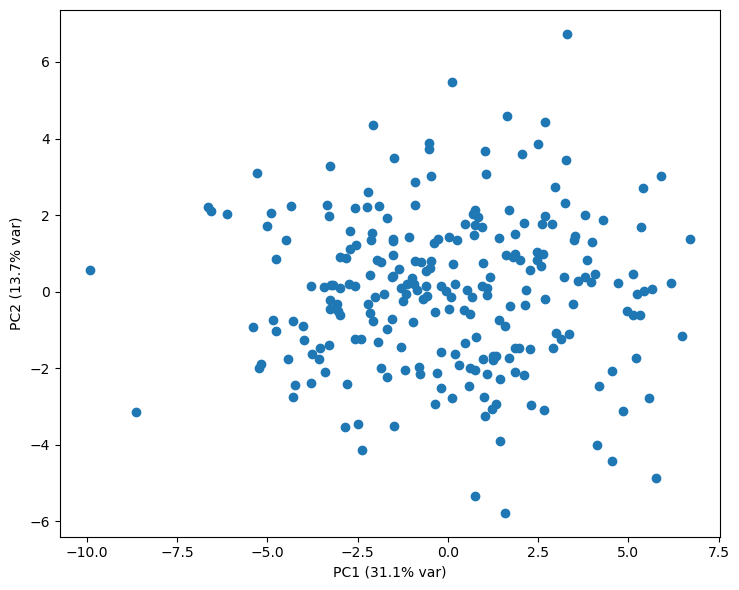


Figure 8. PCA of windowed multimodal features (first two components)

Pruning is its own craft. A common anti-pattern in our field is to hoard features *just in case* and then hide behind a regulariser. A more sensible approach is to ask each feature to justify its presence. We can do this with ablations in a baseline model or even with simple, transparent statistics. Figure 9 uses the latter method and ranks features by their absolute correlation with a synthetic target. In real projects, you will prefer permutation importance or ablations in a small, calibrated model (ridge or logistic regression plus Platt scaling), but the principle remains the same: demonstrate that a feature adds value. The benefits are twofold. First, you shed unnecessary weight and improve calibration under distribution shift (the real world is unpredictable). Second, you maintain an audit trail that a clinician, an IRB, or a sceptical colleague can understand without needing a decoder ring (Can et al., 2023).

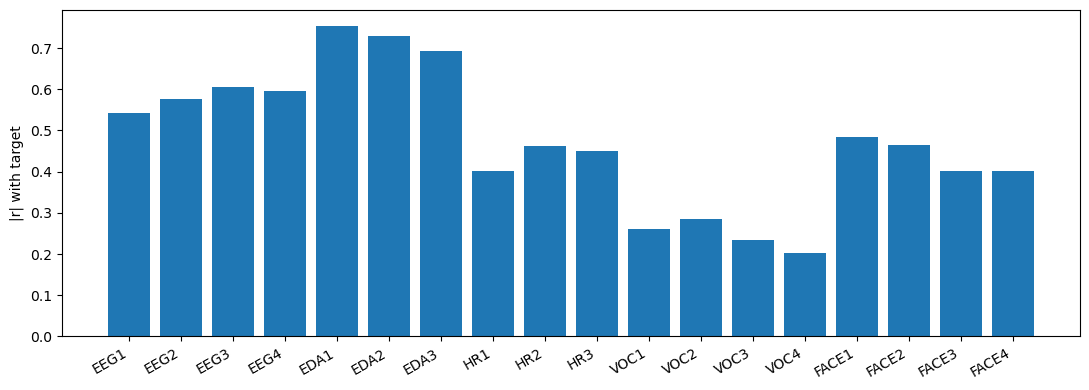


Figure 9. Simple feature importance via |correlation| with target

All this sounds technical, and it is, but the through-line is psychological. Features are claims about processes. Tonic EDA asserts something different from SCR rate; frontal alpha asymmetry speaks to approach/withdrawal in a way that total alpha power does not; F0 variance and spectral tilt tell a different story about vocal effort and affect than raw intensity. When the model’s output says *high arousal* and you can point to elevated SCRs, decreased RMSSD, and a livelier F0 contour, your explanation isn’t a post-hoc story; it is the design working as intended (Russell, 1980; Gross, 2015; Can et al., 2023).

At some point, handcrafted descriptors meet their limits. If you collect a lot of data, or if your contexts change faster than you can curate features, **representation learning** becomes attractive. Self-supervised approaches can distil robust embeddings without dense labels. The trick is to preserve auditability. We should not care only that the embedding predicts but that it behaves consistently across subgroups and doesn’t smuggle in sensitive attributes. A pragmatic pattern is to learn a compact embedding and then pass it to a simple, well-calibrated head (ridge/GBDT) so you can still interrogate the system and publish a legible model card (Can et al., 2023).

Despite technical sophistication, the psychological validity of extracted features must remain central. Features must be theoretically grounded to reflect known emotional constructs. For instance, HRV metrics relate strongly to emotional regulation capacity (Thayer et al., 2012), while pupil dilation has been linked to emotional arousal and cognitive effort (Laeng et al., 2012). Hand-crafted features remain strong baselines because they carry **theoretical anchors** and are easy to audit. But when you have enough data, or when your domain shifts often, **representation learning** helps:

* **Self-supervised** objectives (e.g., predicting masked segments or future frames) learn compact embeddings that capture structure without dense labels, which are rare in affective data (Can et al., 2023).
* **Cross-modal contrastive** learning aligns, say, audio–physiology or face–EDA, so the embedding space encodes synchrony and lag relations; this can stabilise fusion when clocks drift.
* **Hybrid stacks** that feed learned embeddings into simple, calibrated models (ridge/logistic with Platt scaling) often beat deeper end-to-end systems in the wild, where missingness and context shifts dominate.

Even here, keep the **ethics** close: learned representations can smuggle sensitive attributes or dataset-specific quirks. Audit for subgroup performance, do feature-attribution sanity checks, and publish model cards (Can et al., 2023). The sensitive nature of emotional data necessitates ethical caution. Researchers must ensure data anonymisation, respect participant consent about inferred states, and maintain transparency regarding the emotional interpretations made through computational methods (WHO, 2024; Hanisch et al., 2021).

We can say that **features** are the vocabulary we use to retell the story to a model. The trap is to think of features as a shopping list. In embodied emotion, features are **commitments** to a theory of how appraisal unfolds, to a choice of time-scale, to a stance on regulation, and to a definition of context. A good feature set is therefore small, legible, and **explicitly tied** to the processes you care about (Can, Mahesh, & André, 2023). So, the main question is, **which features carry psychological meaning?** Here is the brief overview:

**EEG: Bandpowers** (delta–gamma), **frontal alpha asymmetry**, **event-related dynamics** (ERPs/ERD–ERS) capture attention and approach–withdrawal tendencies at millisecond resolution (Can et al., 2023).Keep features **window-consistent** with the phenomena you target (short windows for phasic reactivity, longer for sustained states);

**EDA: Tonic level (SCL)** indexes baseline arousal; **phasic SCR** count, amplitude, and latency index event-linked sympathetic bursts. Proper decomposition (tonic vs. phasic) is essential (Can et al., 2023);

**Cardiac (HR/PPG/HRV): Time-domain HRV** (SDNN, RMSSD) and **frequency-domain HRV** (LF, HF, LF/HF) reflect autonomic balance; combine with **respiratory** measures to contextualise respiratory sinus arrhythmia (RSA) (Can et al., 2023);

**Voice: prosodic** (F0 mean/variance, intensity), **spectral** (tilt, centroid), and **voice-quality** (jitter, shimmer) features often align with valence and arousal changes and remain robust in real-world audio, especially when the face is occluded;

**Gaze, Face, Body: fixation rate**, **saccade amplitude**, and **pupil dilation** (gaze) plus **AU intensities**, **head pose**, and **posture/gait** (face/body) externalize appraisal and action readiness; suppression can dampen visible cues, so pairing with physiology helps resolve ambiguity (Gross, 2015; Can et al., 2023).

# 2.5 Time Series Data Preprocessing

Time is an inherent dimension of emotions. Emotional states are rarely static. They evolve, intensify, diminish, and shift dynamically over time. Consequently, embodied emotion data frequently emerges in the form of a **time series**. These sequences of data points are collected at regular intervals. Preprocessing these temporal sequences for emotion analysis demands special considerations, significantly differing from approaches applicable to static data (Kuppens & Verduyn, 2017). Emotions naturally change and flow, responding to external events and internal cognitive appraisals (Scherer, 2009). Heartbeats speed up, skin conductance rises, EEG rhythms shift, and facial expressions change. All of these patterns carry essential emotional meaning. Thus, properly capturing, aligning, and interpreting emotional data requires treating time not as an afterthought but as a fundamental organising dimension.

Real-world emotion recordings are rarely perfect. Missing data can occur due to sensor malfunction, human movement, or environmental interference. We ensure that each stream has monotonic timestamps, a known sampling rate, and explicit units. Only then do we touch the signal itself. In physiology, the first enemy is *loss of continuity*. Real recordings drop packets, electrodes loosen, wearables shift on the skin, and people move. We identify missing segments and annotate their cause when known. For time series data, missing values are commonly addressed using interpolation methods (linear or spline) to maintain the continuity of emotional trajectories (Little & Rubin, 2019). Emotion data often includes noise from sensor fluctuations or external disturbances. Time series smoothing methods such as moving averages, median filters, or advanced methods like Gaussian smoothing help preserve the genuine emotional signals by reducing unwanted variability.

Multimodal emotion analysis commonly involves combining multiple sensor streams recorded at different sampling rates (e.g., EEG at 256Hz, GSR at 20Hz, video at 30fps). Temporal alignment ensures that data streams represent the same emotional events simultaneously, preventing analytical confusion or false interpretations (Calvo & D'Mello, 2010). Feature extraction translates raw temporal signals into structured inputs suitable for machine learning. Commonly extracted features include statistical measures (mean, variance, skewness), frequency domain features (spectral bands in EEG), and event-related features (peak detection, latency).

The following script is a complete, ready-to-run pipeline for preparing real-time series data for analysis in emotion research. It reads your existing files for a primary signal and a secondary modality, infers the sampling rate, repairs only very short gaps, smooths noisy measurements, resamples everything to a common rate, and aligns both streams in time. It then cuts the data into short windows, extracts simple statistical and spectral features, removes low-quality windows, standardises features using training data only to avoid leakage, and saves both a feature table and clear diagnostic figures. You only need to point the configuration at the top to your files and, if needed, set the correct column names.

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from scipy.ndimage import gaussian\_filter1d

from scipy.signal import welch, resample\_poly

from numpy.linalg import lstsq

from fractions import Fraction

from math import gcd

import json

import os

PRIMARY\_CSV = "data\_final.csv" # your primary time-series file

PRIMARY\_TIME\_COL = None # e.g., "t", "time", "timestamp"; set None to auto-detect

PRIMARY\_VALUE\_COL = None # e.g., "x", "value", "signal"; set None to auto-detect

SECONDARY\_CSV = "speech\_modality.csv" # your secondary modality file

SECONDARY\_TIME\_COL = None # set None to auto-detect

SECONDARY\_VALUE\_COL = None # e.g., "speech\_prob", "prob"; set None to auto-detect

FS\_TARGET = 4.0 # Hz target resample rate

GAP\_INTERP\_THRESHOLD\_S = 2.0 # s; interpolate only gaps <= this

SMOOTH\_SIGMA\_S = 0.5 # s; Gaussian sigma in seconds

ALIGNMENT\_TOL\_S = 0.25 # s; nearest-neighbor alignment tolerance

WINDOW\_S = 8.0

HOP\_S = 4.0

TRAIN\_FRACTION = 0.7

def pick\_col(df, preferred, fallbacks):

if preferred is not None and preferred in df.columns:

return preferred

for c in fallbacks:

if c in df.columns:

return c

raise ValueError(f"None of the columns { [preferred] + fallbacks if preferred else fallbacks } found in {list(df.columns)}")

def to\_seconds(series):

"""Return (seconds\_array, is\_datetime). Tries datetime first; falls back to numeric."""

try:

dt = pd.to\_datetime(series, utc=True, errors="raise")

secs = (dt.view("int64") / 1e9).to\_numpy() # POSIX seconds

return secs, True

except Exception:

vals = pd.to\_numeric(series, errors="coerce").to\_numpy(dtype=float)

if not np.isfinite(vals).any():

raise ValueError("Time column could not be parsed as datetime or numeric seconds.")

return vals, False

def normalize\_time\_for\_alignment(t\_primary, t\_secondary, abs\_primary, abs\_secondary):

"""

If both are absolute timestamps, align on absolute seconds.

Otherwise, shift both to start at zero and align relatively.

"""

if abs\_primary and abs\_secondary:

return t\_primary, t\_secondary

return t\_primary - np.nanmin(t\_primary), t\_secondary - np.nanmin(t\_secondary)

def interpolate\_short\_gaps(t, x, max\_gap\_seconds=2.0):

"""

Interpolate linearly only across gaps with duration <= max\_gap\_seconds.

Longer gaps remain NaN.

"""

t = np.asarray(t, dtype=float)

x = np.asarray(x, dtype=float)

isnan = ~np.isfinite(x)

# Identify contiguous NaN runs

runs = []

in\_run = False

start = None

for i, flag in enumerate(isnan):

if flag and not in\_run:

in\_run = True

start = i

elif not flag and in\_run:

runs.append((start, i - 1))

in\_run = False

if in\_run:

runs.append((start, x.size - 1))

x\_interp = x.copy()

for s, e in runs:

gap\_duration = t[e] - t[s] if e > s else 0.0

if gap\_duration <= max\_gap\_seconds:

left = s - 1

right = e + 1

if left >= 0 and right < x.size and np.isfinite(x[left]) and np.isfinite(x[right]):

x\_interp[s:e+1] = np.interp(t[s:e+1], [t[left], t[right]], [x[left], x[right]])

return x\_interp

def bandpower\_welch(x, fs, fmin, fmax):

f, Pxx = welch(x, fs=fs, nperseg=min(len(x), int(fs \* 8)))

idx = (f >= fmin) & (f <= fmax)

return np.trapz(Pxx[idx], f[idx]) if np.any(idx) else np.nan

def linear\_trend(x):

y = np.asarray(x)

X = np.vstack([np.arange(y.size), np.ones(y.size)]).T

beta, \*\_ = lstsq(X, y, rcond=None)

return float(beta[0])

def infer\_fs(t\_seconds):

dt = np.diff(t\_seconds)

dt = dt[np.isfinite(dt) & (dt > 0)]

if dt.size == 0:

raise ValueError("Cannot infer sampling rate (insufficient or invalid time data).")

return 1.0 / np.median(dt)

p\_df = pd.read\_csv(PRIMARY\_CSV)

t\_col = pick\_col(p\_df, PRIMARY\_TIME\_COL, ["t", "time", "timestamp", "datetime", "Time", "Timestamp"])

x\_col = pick\_col(p\_df, PRIMARY\_VALUE\_COL, ["x", "value", "signal", "y", "amp", "amplitude"])

t\_primary\_raw, primary\_is\_abs = to\_seconds(p\_df[t\_col])

x\_raw = pd.to\_numeric(p\_df[x\_col], errors="coerce").to\_numpy(dtype=float)

# Sort by time and drop exact-duplicate time stamps if any

order = np.argsort(t\_primary\_raw)

t\_primary\_raw = t\_primary\_raw[order]

x\_raw = x\_raw[order]

keep = np.ones\_like(t\_primary\_raw, dtype=bool)

keep[1:] = t\_primary\_raw[1:] > t\_primary\_raw[:-1]

t\_primary\_raw, x\_raw = t\_primary\_raw[keep], x\_raw[keep]

fs\_raw = infer\_fs(t\_primary\_raw)

# Normalize time for plotting convenience (relative seconds from start)

t\_primary\_rel = t\_primary\_raw - np.nanmin(t\_primary\_raw)

# ---------- 2) Short-gap interpolation; long-gap masking ----------

x\_interp = interpolate\_short\_gaps(t\_primary\_rel, x\_raw, max\_gap\_seconds=GAP\_INTERP\_THRESHOLD\_S)

valid\_mask = np.isfinite(x\_interp).astype(float)

# ---------- 3) Denoise / Smooth ----------

sigma\_samples = max(1e-6, SMOOTH\_SIGMA\_S \* fs\_raw) # Gaussian sigma in samples

x\_fill = np.where(np.isfinite(x\_interp), x\_interp, np.nanmean(x\_interp))

x\_smooth = gaussian\_filter1d(x\_fill, sigma=sigma\_samples)

# ---------- 4) Resample with anti-aliasing ----------

ratio = Fraction(FS\_TARGET / fs\_raw).limit\_denominator(1000)

up, down = ratio.numerator, ratio.denominator

x\_for\_resample = np.where(np.isfinite(x\_smooth), x\_smooth, np.nanmean(x\_smooth))

x\_resampled = resample\_poly(x\_for\_resample, up, down)

mask\_resampled = resample\_poly(valid\_mask, up, down)

# Resampled time vector (relative seconds)

n\_res = x\_resampled.size

t\_resampled = np.arange(n\_res) / FS\_TARGET

mask\_resampled = np.clip(mask\_resampled, 0, 1)

# ---------- 5) Load and align second modality ----------

s\_df = pd.read\_csv(SECONDARY\_CSV)

sb\_t\_col = pick\_col(s\_df, SECONDARY\_TIME\_COL, ["t", "time", "timestamp", "datetime", "Time", "Timestamp"])

sb\_v\_col = pick\_col(s\_df, SECONDARY\_VALUE\_COL, ["speech\_prob", "prob", "probability", "value", "y"])

t\_secondary\_raw, secondary\_is\_abs = to\_seconds(s\_df[sb\_t\_col])

b\_vals = pd.to\_numeric(s\_df[sb\_v\_col], errors="coerce").to\_numpy(dtype=float)

# Sort, deduplicate

order\_b = np.argsort(t\_secondary\_raw)

t\_secondary\_raw = t\_secondary\_raw[order\_b]

b\_vals = b\_vals[order\_b]

keep\_b = np.ones\_like(t\_secondary\_raw, dtype=bool)

keep\_b[1:] = t\_secondary\_raw[1:] > t\_secondary\_raw[:-1]

t\_secondary\_raw, b\_vals = t\_secondary\_raw[keep\_b], b\_vals[keep\_b]

# Align time bases (absolute if both timestamped; else relative)

t\_primary\_align, t\_secondary\_align = normalize\_time\_for\_alignment(

t\_primary\_raw, t\_secondary\_raw, primary\_is\_abs, secondary\_is\_abs

)

# Build dataframes for merge\_asof (use resampled primary timeline)

series\_df = pd.DataFrame({"t": t\_resampled, "x": x\_resampled, "valid": mask\_resampled})

speech\_df = pd.DataFrame({"t": (t\_secondary\_align - np.nanmin(t\_primary\_align)), "speech\_prob": b\_vals})

series\_df.sort\_values("t", inplace=True)

speech\_df.sort\_values("t", inplace=True)

aligned = pd.merge\_asof(series\_df, speech\_df, on="t", direction="nearest", tolerance=ALIGNMENT\_TOL\_S)

# ---------- 6) Windowing & Feature Extraction ----------

starts = np.arange(0, float(series\_df["t"].iloc[-1]) - WINDOW\_S + 1e-9, HOP\_S)

rows = []

for s in starts:

e = s + WINDOW\_S

m = (aligned["t"] >= s) & (aligned["t"] < e)

seg = aligned.loc[m]

if seg.empty:

continue

xw = seg["x"].to\_numpy()

qw = seg["valid"].to\_numpy()

sw = seg["speech\_prob"].to\_numpy()

coverage = float(np.mean(qw > 0.5))

feats = {

"t\_start": float(s),

"t\_end": float(e),

"n": int(xw.size),

"coverage": coverage,

"mean": float(np.mean(xw)),

"std": float(np.std(xw)),

"min": float(np.min(xw)),

"max": float(np.max(xw)),

"trend": linear\_trend(xw),

# Spectral bands relative to FS\_TARGET = 4 Hz

"bp\_vlow": bandpower\_welch(xw, FS\_TARGET, 0.01, 0.08),

"bp\_low": bandpower\_welch(xw, FS\_TARGET, 0.08, 0.5),

"bp\_high": bandpower\_welch(xw, FS\_TARGET, 0.5, 1.5),

"speech\_mean": float(np.nanmean(sw)) if np.isfinite(sw).any() else np.nan,

}

rows.append(feats)

features = pd.DataFrame(rows)

# ---------- 7) Quality filtering and train/test standardization ----------

features\_qc = features[features["coverage"] >= 0.8].reset\_index(drop=True)

cut\_time = features\_qc["t\_end"].max() \* TRAIN\_FRACTION if not features\_qc.empty else 0.0

train\_idx = features\_qc["t\_end"] <= cut\_time

test\_idx = ~train\_idx

num\_cols = ["mean", "std", "min", "max", "trend", "bp\_vlow", "bp\_low", "bp\_high", "speech\_mean"]

train\_means = features\_qc.loc[train\_idx, num\_cols].mean()

train\_stds = features\_qc.loc[train\_idx, num\_cols].std().replace(0, 1.0)

features\_scaled = features\_qc.copy()

features\_scaled[num\_cols] = (features\_qc[num\_cols] - train\_means) / train\_stds

# ---------- 8) Save artifacts ----------

out\_csv = "ts\_preprocessing\_features.csv"

features\_scaled.to\_csv(out\_csv, index=False)

params = {

"fs\_raw": float(fs\_raw),

"fs\_target": FS\_TARGET,

"window\_seconds": WINDOW\_S,

"hop\_seconds": HOP\_S,

"gap\_interpolation\_threshold\_s": GAP\_INTERP\_THRESHOLD\_S,

"alignment\_tolerance\_s": ALIGNMENT\_TOL\_S,

"train\_split\_end\_time": float(cut\_time),

"primary\_csv": os.path.abspath(PRIMARY\_CSV),

"secondary\_csv": os.path.abspath(SECONDARY\_CSV),

"primary\_time\_col": t\_col,

"primary\_value\_col": x\_col,

"secondary\_time\_col": sb\_t\_col,

"secondary\_value\_col": sb\_v\_col,

}

with open("ts\_preprocessing\_params.json", "w") as f:

json.dump(params, f, indent=2)

The goal of this pipeline is to turn raw sensor recordings into tidy, comparable numbers that a model or a human can understand. Real sensors sometimes stop briefly or lose contact, so the script first fills only very short gaps by drawing a straight line between the nearest valid points, and it leaves longer outages as missing in order not to invent data. Because sensors are also noisy, a gentle Gaussian smoothing is applied so that rapid jitters are reduced while the broader shape of the signal remains intact. Different devices often record at different speeds, which makes direct comparison difficult, so the script converts the primary signal to a common target rate using a method that prevents artificial distortion. It then lines up the primary signal with the second data stream in real time within a small tolerance, so values from the same moments can be compared.

Next, the data are sliced into short, partly overlapping windows. For each window, the script calculates simple summaries such as average level, variability, minimum and maximum, the overall upward or downward trend, and how much energy sits in slow, medium, or faster fluctuations. It also records the average value of the second modality over the same period. Each window carries a quality score based on how much real, non-missing data it contains, and windows below the threshold are excluded so that downstream analysis is not driven by gaps. To evaluate models fairly, the script separates the earlier portion of time as training and keeps the later portion for testing, then scales the features using only the training part so that future data is not allowed to influence the past. Finally, it writes a clean feature table to a CSV file and stores a small JSON file with the exact settings for reproducibility.

# 2.6 Preprocessing Visual Data for Emotion Analysis

When we work with video, the most helpful thing we can do is keep the picture simple and steady. A camera that records smoothly (around thirty frames per second), a face that is well lit from the front, and a view that shows the head and a bit of the shoulders usually give us clean signals. Heavy compression can also be misleading: blocky artefacts in a low-quality file sometimes look like real facial movements to an algorithm, so moderate, consistent compression is preferable (Das, 2023).

For analysis, specialised software is useful. It first finds the face in each frame and keeps following it as the person moves. Then it pinpoints a set of stable reference points, for example, the corners of the eyes and mouth, the tip of the nose, and the outline of the jaw (see Figure 10). *Find–follow–straighten* routine, implemented by widely used tools such as *RetinaFace* and *MediaPipe Face Mesh*, helps the later emotion models focus on meaningful changes in the face rather than on head pose or camera position (Bazarevsky et al., 2020; Deng et al., 2020; Li & Deng, 2020). If the person turns too far to the side or tilts their head a lot, the system simply marks those moments as less reliable instead of pretending it knows more than it does (Kollias, 2022).



Figure 10. Face detection and landmark extraction

Before we analyse anything, we adjust brightness and contrast so that faces look as if they were filmed under similar conditions. Gentle, local contrast methods (for example, adaptive histogram equalisation) tend to do this well (Li & Deng, 2020). Alongside these corrections, we should save a few simple quality notes for every frame. For example, it is useful to record how sharp the image is, how large the face is on screen, whether glasses or hands are covering key areas, and how much the head is turned. Those notes are later used to down-weight doubtful frames rather than to throw them away, which helps systems behave more robustly *in the wild* (Stappen et al., 2021).

Researchers often describe expressions in **action units**, which are small, named movements such as lifting the inner eyebrows or pulling the lip corners (Ekman & Rosenberg, 2005). In practice, the computer measures both shape (how far landmarks move) and appearance (how local texture changes) around these regions. These little time patterns can separate a deliberate smile from a relieved one or a fleeting frown from a sustained one, and they are especially helpful when we later combine vision with physiology and voice (Kossaifi et al., 2017; Can, Mahesh, & André, 2023).

Faces are not the whole story. People also *speak* with their bodies. Modern pose estimators can track major joints and hands in real time, giving us simple yet informative measures such as upper-body expansion, sway, and blink or saccade rates. These cues are valuable when faces are partially covered, culturally restrained, or simply off camera (Bazarevsky et al., 2020; Davey et al., 2021; Pereira et al., 2024).

When we feed images to deep models, we keep things consistent and honest. All face crops are resized to the same dimensions and normalised in the same way so that the network does not learn accidental differences between cameras. We also imitate the noise that we expect in daily life (rotations, brightness, blur), so the model learns to ignore them. At the same time, we keep an *interpretable line* running next to the deep features: we continue to compute action-unit signals and landmark movements and check that the network’s internal representation still lines up with those human-readable cues. If it does not, that is a warning that the model may be keying on identity, background or other shortcuts (Li & Deng, 2020; Minderer et al., 2021).

We must also remember that people do not feel in single frames. Frame-by-frame predictions jump around and can be noisy. We therefore smooth the evidence over short intervals and group them into meaningful segments, which can then be aligned with peaks in skin conductance, dips in heart-rate variability, or voice changes. The result is a timeline that reads like a story rather than a flicker book (Can et al., 2023; Ringeval et al., 2017).

# 2.7 Audio Signals for Embodied Emotion Analysis

When people speak, their voices often reveal more about their feelings than their words. A catch in the throat that appears only on certain topics, a brighter pitch when mentioning a friend, or a flattening of loudness after a long day are bodily traces of appraisal and arousal transmitted through the respiratory system, the larynx, and the vocal tract. Because emotion influences breathing patterns, laryngeal tension, and articulatory timing, the acoustic signal provides a practical window into autonomic shifts and regulatory efforts (MacIntyre, 2022; Schuller & Batliner, 2013). Voice is a non-intrusive, richly embodied channel that travels effectively across everyday contexts, from therapy rooms to cars and mobile phones (Latif et al., 2020; Haq & Jackson, 2011).

Listeners instinctively rely on four clusters of cues when they wish to discover what the voice conveys. **Prosody** (variations in pitch or fundamental frequency), **loudness**, and **rhythm** reflect arousal and urgency; rapid speech and rising pitch often indicate agitation or excitement, while slower, quieter speech is typical in sadness and fatigue (Cowen et al., 2019; Akçay & Oğuz, 2020). **Voice quality**, how *breathy* or *tense* the voice sounds, stems from vocal-fold vibration and glottal airflow, and frequently changes with stress or relief (Gobl & Chasaide, 2003; Hashem et al., 2023). **Spectral–timbral** patterns (described by MFCCs or similar features) mirror the shape of the vocal tract and articulatory accuracy, which can diminish under stress (Eyben & Schuller, 2015; Khalil et al., 2019). Lastly, **paralinguistic events** (sighs, laughs, hesitations) serve as emotional punctuation and are useful indicators when language is culturally restrained (Trigeorgis et al., 2016; Haq et al., 2011).

Same as in the case of other types of datasets, field recordings are rarely clean. They come with air-conditioning hum, traffic, room reverberation, and changing microphone distance. A well-documented front end, therefore, matters more than exotic models. When it comes to the recording of raw data, we start by **keeping capture simple and consistent**: a single channel at 16 kHz or 24 kHz, stable input gain that avoids clipping, and a short calibration phrase to check levels. On phones and wearables, it helps to log the device model so that later analyses can account for microphone differences (Latif et al., 2020). Proceed with **cleaning that preserves the meaning**. Light spectral gating or Wiener filtering can lower stationary background noise; mild dereverberation enhances intelligibility in reverberant spaces; and **voice-activity detection (VAD)** removes long silences so that subsequent statistics reflect the voice, not the room (Eyben et al., 2015; Tan et al., 2021). When multiple people speak, **speaker diarisation** separates streams so that features represent one speaker at a time (Park et al., 2022). We then normalise signals on a comparable scale (peak- or loudness-normalisation within safe limits) so that differences in microphone gain are not mistaken for differences in emotion (Schuller & Batliner, 2013).

**Segmentation** is used to cut the stream into small, overlapping windows (often 2–4 s) to capture the arc of an utterance, while keeping **episodes** (moments of clear rise and recovery) in view for later alignment with physiology or gaze. Fixed windows are convenient, but event-based windows (at breaths or pauses) often tell the cleaner story (Eyben et al., 2015; Ringeval et al., 2019). From each window, we compute **prosodic** features (F0 level and range, energy statistics, speaking rate), **voice-quality** measures (jitter, shimmer, harmonic-to-noise ratio; sometimes glottal parameters via COVAREP-style analysis), and **spectral** descriptors (MFCCs, spectral centroid and tilt). These are simple but effective markers of arousal and control when averaged, differenced, and smoothed over time (Eyben et al., 2015; Latif et al., 2020; Khalil et al., 2019). We add **temporal derivatives** (how fast pitch or energy is changing) because emotion is often carried in the *movement* rather than the static value (Akçay & Oğuz, 2020).

Alongside handcrafted features, **self-supervised acoustic embeddings** have become the default backbone. Models such as **wav2vec 2.0**, **HuBERT**, **BYOL-S**, and large audio networks trained on spoken-word and sound events yield robust representations that transfer well to emotion tasks with minimal fine-tuning (Baevski, Zhou, Mohamed, & Auli, 2020; Hsu et al., 2021; Kong et al., 2020). These embeddings are not magic; we still stabilise them with per-speaker normalisation, small amounts of smoothing, and **data augmentation** (noise, reverberation, small pitch and tempo shifts, SpecAugment-style masking) to avoid overfitting to recording conditions (Park et al., 2019; Latif et al., 2020).

Two confounds deserve explicit treatment: (1) in the case of **content leakage**, models can cheat by learning *what was said* (lexical content) rather than *how it was said*. We mitigate this with phoneme-balanced prompts, adversarial training against transcripts, or evaluations that hold content constant (Poria et al., 2021); (2) it is important to take care about **speaker identity**, because many features are stable traits (average pitch, formant structure). We therefore evaluate across unseen speakers and, when possible, decorrelate identity using adversarial losses or embedding anonymisation (Champion, 2023; Stappen et al., 2021).

Reliable systems are **humble** about uncertainty. They attach quality indicators (VAD confidence, SNR, clipping rate) to each window, smooth predictions over short spans, and abstain when evidence is weak. They are tested beyond a single lab: cross-corpus and cross-device evaluations are the norm in community challenges such as **AVEC** and **MuSe**, which repeatedly show that careful preprocessing and calibration often matter more than one more model layer (Ringeval et al., 2019; Stappen et al., 2021). For continuous targets (arousal/valence), we report the **concordance correlation coefficient**; for imbalanced categories, the **unweighted average recall** avoids a misleading overall accuracy (Chamishka et al., 2022).

Audio rarely stands alone. Its greatest value appears when we align it with **physiology** (EDA peaks, HRV dips) and **vision** (micro-expressions, gaze shifts): subtle tremor in the voice that coincides with a skin-conductance surge is stronger evidence than either cue by itself (Can, Mahesh, & André, 2023).

To demystify what *audio preprocessing* does, Figure 11 walks through the entire pipeline using **the same recording**. That matters: because the clip never changes, you can see exactly how each step clarifies the emotional traces in the voice rather than introducing new artefacts. Think of it as cleaning a lens, then switching between views that highlight different parts of the same scene.

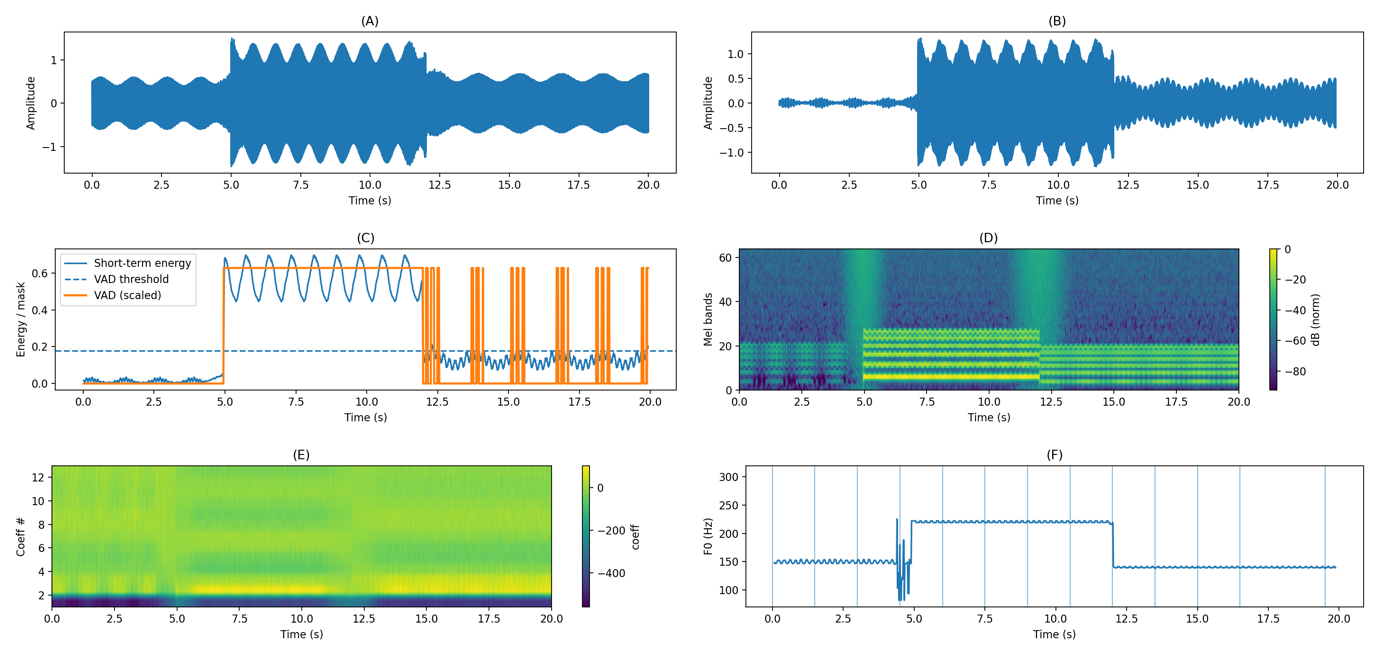


Figure 11. End-to-end audio preprocessing on a single recording. (A) Raw waveform with natural noise and breathing. (B) Denoised waveform after light spectral gating, preserving expressive dynamics. (C) Short-term energy with a simple VAD threshold; the mask indicates speech-dominant regions. (D) Log-mel spectrogram, revealing richer high-frequency energy during the central, higher-arousal segment. (E) MFCCs (13 coefficients) summarising timbre over time. (F) Pitch contour (F0) with fixed 3-second analysis windows used for episode-level modelling.

In panel A, you’re looking at the raw waveform. It’s messy but alive: you can already glimpse the broad arc of the performance, with a gentler start, a brighter middle, and a softer ending. Panel B shows the denoised signal. The background haze is lifted just enough to make the speech stand out, without flattening the expressive rises and falls. This step doesn’t *beautify* the speaker; it simply reduces distractions so later measurements describe the voice rather than the room. Panel C adds a practical layer: short-term energy (how strong the voice is at each moment) and a simple VAD. The dashed line is the operating point; the mask highlights where speech dominates. VAD stops long silences or rustling from sneaking into your summary statistics. In panel D, the same signal is shown as a log-mel spectrogram, i.e., a time–frequency map that gently resembles what the human ear pays attention to. Brighter bands in the middle of the clip reflect higher arousal: the voice becomes more energetic, richer in overtones, and therefore *brighter* on the page. Panel E condenses that picture into MFCCs, compact timbral descriptors that many emotion models use as inputs. You can see them drift and tighten across the three phases of the clip; it’s a shorthand for how the vocal tract shapes the sound under different emotional loads. Panel F traces the pitch contour (F0) and draws the analysis windows used downstream. The contour rises and varies more in the central portion (exactly where the spectrogram and energy also looked more intense), giving a coherent multisided view of increased arousal. The vertical lines mark fixed windows that later models will treat as *mini-episodes*.

# 2.8 Preparing Text Data for Understanding Emotions in the Body

People *lean* on words to register how they feel. Unlike facial muscle activity or heart rate, text carries **appraisal** and **context**: who or what the feeling is about, why it matters, and how it unfolds over time. Decades of psycholinguistic work show that function words, pronoun use, tense, and affect terms correlate with mood, coping, and health outcomes (Pennebaker & Chung, 2011; Boyd, Ashokkumar, Seraj, & Pennebaker, 2022). Modern NLP adds scale: large corpora and pretrained language models can surface subtle emotional signals beyond simple positive/negative labels (Acheampong et al., 2020; Buechel & Hahn, 2016; Demszky et al., 2020).

Text fills in what physiology and behaviour cannot say on their own. A spike in skin conductance tells you *that* something mattered; a sentence often tells you *what* and *why*. When synchronized with timestamps text makes it possible to link **appraisals** (*I felt cornered when the question came up*) with **bodily responses** (arousal peak + brief gaze aversion) (Zadeh et al., 2018; Tsai et al., 2019; Hazarika et al., 2020). Text also travels well: when cameras or wearables are impractical, reflective journals and brief check-ins can still capture shifts in affect and regulation strategies (Pennebaker & Chung, 2011; Boyd et al., 2022).

The goal is to make text **machine-readable without washing out affect**. Over-aggressive cleaning (stripping emojis, elongations, punctuation, casing) can erase cues that carry strong emotional meaning online (Barbieri et al., 2016; Mohammad, 2022). A practical rule:

* normalize obvious noise (HTML, boilerplate),
* standardize whitespace and Unicode,
* **preserve** emojis, repeated letters (“soooo”), exclamation marks, and laughter tokens because they are emotion features, not dirt (Novak, Smailović, Sluban, & Mozetič, 2015; Mohammad, 2022).

Below is a Python illustration that keeps expressive cues intact while preparing tokens for analysis.

import re

def clean\_keep\_affect(text):

text = re.sub(r'https?://\S+|www\.\S+', ' <URL> ', text) # keep a marker

text = re.sub(r'@\w+', ' <USER> ', text) # anonymize mentions

text = re.sub(r'#(\w+)', r'\1', text) # drop hash, keep tag

text = text.replace('\u200d', '') # zero-width joiner

text = re.sub(r'\s+', ' ', text).strip()

return text

demo = Path("my\_text\_file.txt").read\_text(encoding="utf-8")

print(clean\_keep\_affect(demo))

Output: Sooo happy!!! 😭➡️😊 Thanks <USER> — yesterday was awful: <URL>

There are two dependable ways to turn language into signals your models can read. The first is **lexicon-guided scoring,** where curated inventories map words to emotions (e.g., NRC Emotion Lexicon; LIWC-22 categories). They are transparent and useful for small datasets or clinical explainability (Mohammad & Turney, 2013; Boyd et al., 2022). A lexicon example in Python can be represented as:

from nrclex import NRCLex

text = "I feel relieved and proud today, but yesterday I was anxious."

emo = NRCLex(text)

print(emo.top\_emotions)

Output: [('anticipation', 0.2857142857142857)]

The second route uses **contextual embeddings**. Models like BERT/RoBERTa encode words in context and can be fine-tuned for multi-label emotion (e.g., GoEmotions’ 27 classes), capturing irony, negation, and who-did-what-to-whom better than bag-of-words (Devlin, Chang, Lee, & Toutanova, 2019; Liu et al., 2019; Demszky et al., 2020).

Emotion labels depend on the **unit of analysis**. Sentence-level labels work for short posts; therapy transcripts often need turn-level or episode-level summaries (Buechel & Hahn, 2016). For multimodal work, align text spans to time (utterance start/stop) so they can be compared with concurrent changes in voice or physiology (Zadeh et al., 2018; Tsai et al., 2019). When domains differ (clinical vs. social media), expect domain shift and fine-tuning or few-shot adaptation usually helps (Demszky et al., 2020; Acheampong et al., 2020).

# 2.9 Gaze Data Analysis in Embodied Emotion Recognition

We reveal a lot with our eyes even when we say nothing. People in a tense meeting glance repeatedly at the exit; a parent lingers on a child’s face longer than on anyone else’s; a student’s pupils open a little wider during a hard problem and relax again when it’s solved. These small movements are not random wanderings. They reflect shifts in **attention**, **motivation**, and **arousal**, linking cognition and physiology in real time (Holmqvist et al., 2015; Joshi & Gold, 2020). That is why gaze belongs in an embodied-emotion toolkit alongside physiology, speech, and facial behaviour: it lets us watch *where* the mind goes and *how strongly* the body is engaged.

Emotion changes what we look at and how we look. Threat cues pull the eyes quickly and often keep them there; curiosity broadens the search and produces more exploratory scanpaths; sadness slows the pace and shortens saccades (Carvalho et al., 2014; Orquin & Holmqvist, 2018; Wen et al., 2022). Pupil diameter adds a direct physiological handle: it tracks sympathetic arousal and mental effort, rising within a second of increased load or affect and falling with relief (Mathôt, 2018; Joshi & Gold, 2020). These properties make gaze especially powerful when faces are masked or speech is sparse: it provides a running marker of engagement that we can align with skin conductance, heart-rate variability, or vocal tremor.

Good gaze data starts with a **quiet, careful setup**. Whether you use a remote tracker or head-mounted glasses, brief calibration and a quick validation check pay off later; even small miscalibrations can bias conclusions about what attracted attention (Holmqvist et al., 2015; Orquin & Holmqvist, 2018). In the wild, you will encounter blinks, eyelashes, and brief tracking losses. Rather than discarding whole trials, we (i) **detect blinks**, (ii) **interpolate short gaps**, and (iii) apply **gentle smoothing** so that fixation detection is not driven by sensor jitter (Kret & Sjak-Shie, 2019). For **pupil size**, we remove blink-related spikes, correct for slow drift, and control or record **luminance**, because lighting changes can mimic emotion (Mathôt, 2018).

You can see the core idea in Figure 12. We convert the raw gaze path (x–y over time) into **velocity** (how fast the point of regard is moving), and then mark intervals that stay below a sensible, data-driven threshold as *fixations*. The shaded bands show those moments when the eyes truly rest on something rather than fluttering from noise. In affective contexts, longer or repeated fixations on threat- or loss-related areas are common in anxiety, while broader, more exploratory patterns appear with curiosity and interest (Armstrong & Olatunji, 2012; Kidd & Hayden, 2015). Practically, we smooth the velocity trace, set the threshold from a robust percentile of the distribution, and apply a **minimum-duration** rule so that blinks or micro-movements are not misread as fixations (Kret & Sjak-Shie, 2019; Holmqvist et al., 2015).

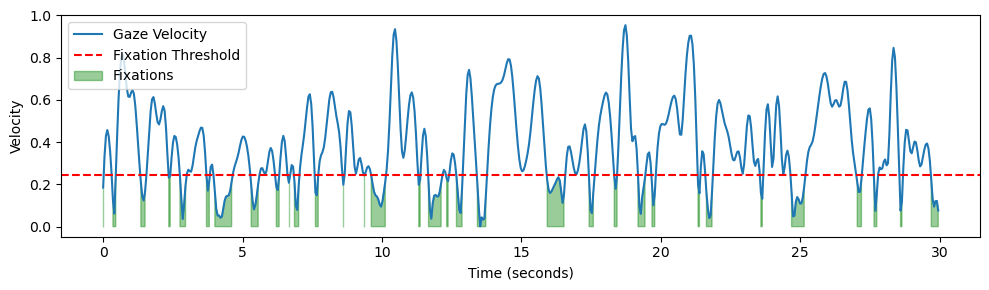


Figure 12. Fixation detection from gaze velocity

From cleaned traces, we extract features that map neatly onto psychological processes:

* **Fixation metrics** (mean duration, total dwell time in Areas of Interest (AOIs), and transitions between AOIs), index **engagement** and **attentional bias** (Holmqvist et al., 2015; Orquin & Holmqvist, 2018).
* **Saccade metrics** (rate, amplitude, and peak velocity) reflect **search strategy** and arousal; anxious scanning often looks fast and shallow, whereas confident search is slower and more targeted (Armstrong & Olatunji, 2012; Kidd & Hayden, 2015).
* **Pupil metrics**, such as tonic level and phasic dilation (peak, latency, recovery), track **autonomic activation** and **effort** with good temporal fidelity (Mathôt, 2018; Joshi & Gold, 2020).
* **Scanpath structure** (entropy, sequence patterns, and AOI transition matrices) captures how organised the exploration is; high entropy can indicate either curiosity or distraction, depending on context (Holmqvist et al., 2015).

In experiments, fixation classification is often done with simple velocity- or dispersion-threshold rules (I-VT or I-DT). For mobile recordings, short-window Hidden Markov Models or Bayesian filters improve robustness to head motion and brief losses (Holmqvist et al., 2015).

Gaze is deeply personal. It reveals what attracts us and what we avoid, and it can serve as a biometric identifier. We therefore minimise raw storage (features instead of full streams), obtain **explicit consent** for any inference of emotion, and disclose the limits of what the system can say (WHO, 2024). On-device processing is preferred for mobile systems, where raw video must be stored, and access is tightly controlled. Gaze can also leak sensitive traits, which makes **privacy-preserving design** and simple user controls to pause logging non-negotiable (Kröger, Lutz, & Müller, 2020; Steil et al., 2019). cultural variation matters: direct eye contact can signal attention in many Western settings but deference or politeness-avoidance in parts of East Asia; gaze scanning strategies also differ across cultures, so cross-population validation is essential (Senju et al., 2013; Uono & Hietanen, 2015; Pang, Zhou, & Chu, 2024).

# 2.10 Multimodal Data Fusion for Embodied Emotion

The promise of artificial psychology is not that one sensor will finally *see* emotion; it’s that **multiple, imperfect witnesses**, such as EEG rhythms, EDA bursts, HRV balance, vocal prosody, facial micro-gestures, gaze, posture, and textual context, can be made to **agree enough** about what happened and for whom. Technically, that agreement is a fusion problem; scientifically, it is a commitment to the idea that emotion is distributed across body and setting rather than localised in a single channel (Poria et al., 2018; Can, Mahesh, & André, 2023). The question for this section is practical: *how do we model the joint story in a way that survives the wild missing packets, clock drift, suppression, masks, noise, and still earns the trust of a clinician, a teacher, or a user?*

A first design choice is **when** to fuse (Figure 13). In **early fusion**, we concatenate features from synchronised windows and learn one model. This is attractive when clocks are tight and coverage is high; it lets the learner capture cross-modal covariance directly, such as a rise in EDA coinciding with a slight F0 increase and a head pitch change (Can et al., 2023). In **intermediate fusion**, we learn a shared representation that forces modalities to meet in a latent space while keeping their specifics nearby; this tends to be robust when sampling rates differ or when windows do not align perfectly (Hazarika et al., 2020; Lahat et al., 2015). In **late fusion**, each modality votes separately and a meta-learner or rule set reconciles them; it’s the most modular and usually the most resilient to device failure or dropout (Poria et al., 2018; Can et al., 2023). There is no dogma here since it is well known that strong systems often mix styles: feature-level fusion for stable pairs (voice+face), decision-level fusion to hedge against sensors that disappear mid-session (Mai, Hu, Xing, & Hsu, 2020).

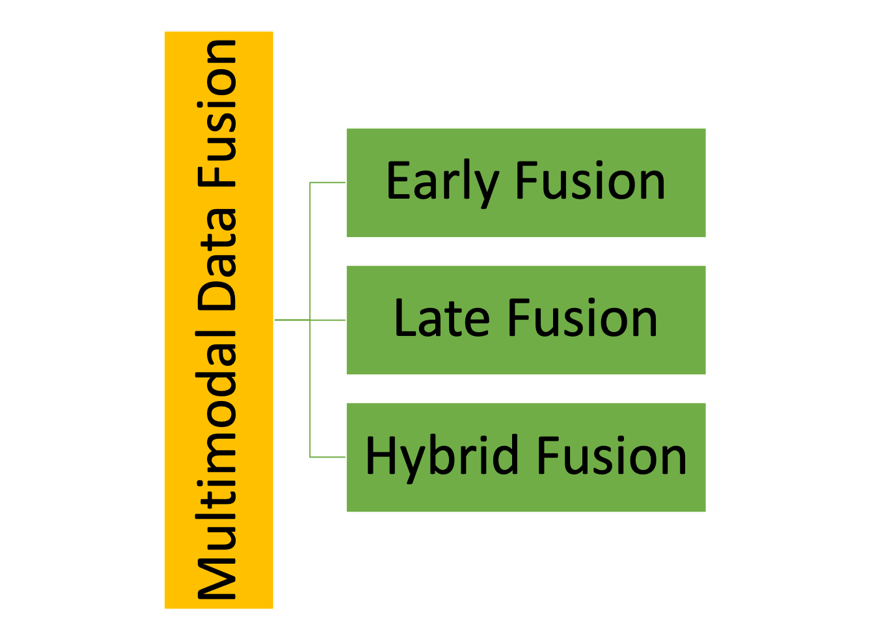


Figure 13. Multimodal data fusion approaches

Underneath the timing lies a **representation** problem. If one believes that physiology and behaviour share a small *semantic* core (arousal, valence, control), then it makes sense to learn **modality-invariant** embeddings while preserving **modality-specific** factors for cases where channels disagree (Hazarika et al., 2020). Recent multimodal transformers operationalise this by creating cross-attention pathways where audio can attend to face, face to text, and so on, letting the model soft-align **what** changes with **when** it changes (Tsai, Bai, Liang, Kolter, & Morency, 2019; Liang et al., 2024). Self-supervised pretraining has pushed this further: models trained to align raw video, audio, and text without labels learn representations that later adapt to emotion tasks with remarkably few annotations (Akbari et al., 2021; Alayrac et al., 2022). For embodied emotion, such pretraining tames data scarcity and reduces the pressure to over-label, which is ethically important when labels touch mental states (Can et al., 2023).

Physiological fusion presents its subtleties. Autonomic signals **diverge on purpose** under regulation: a person can keep a neutral face while EDA spikes, or slow their breathing and recover HRV while the voice still carries strain (Gross, 2015; Can et al., 2023). Good models treat divergence not as noise to be averaged out but as **evidence of strategy**. Intermediate fusion methods from neuroimaging (e.g., multi-set CCA and joint ICA) were designed to extract covarying brain–behaviour components and are increasingly used to bind physiology and behaviour as well (Sui et al., 2020). In our domain, mCCA–jICA has been used to link grey–white matter patterns with acceptance ability (high vs. low), highlighting how **shared** and **specific** structure can be separated and interpreted in human terms (Grecucci et al., 2023). The moral is straightforward: choose a fusion family that makes *disagreement interpretable* rather than invisible.

The recent wave of **multimodal sentiment/emotion** work offers a toolkit we can adapt rather than reinvent. Tensor-fusion and low-rank fusion decompositions keep interaction terms without exploding parameters (Zadeh, Chen, Poria, Cambria, & Morency, 2017; Liu et al., 2018). Modality-invariant/specific splits (MISA) regularise the latent space so that what is shared does not memorise a single channel’s quirks (Hazarika et al., 2020). Multimodal transformers (MulT) learn directional cross-attention (e.g., how voice changes guide the reading of face) (Tsai et al., 2019). More recently, masked-prediction objectives over raw video–audio–text (VATT) reduce reliance on manual features, which is helpful in the wild where feature extraction can fail silently (Akbari et al., 2021). Survey work since 2020 converges on two lessons: pretrain broadly, then **fine-tune gently** with regularisation and calibration; and treat missingness as a first-class citizen rather than a preprocessing error (Poria et al., 2018; Mai et al., 2019; Ramaswamy & Palaniswamy, 2024; Can et al., 2023). Comparable machine-learning frameworks have also been applied to affect-linked constructs beyond core emotions, such as love addiction, where model features and explanations clarify predictive factors (Farahani et al., 2025).

There is also a **causal** question we must not duck. Predicting labels is not the same as understanding processes. When we deploy systems that will nudge learners or trigger clinical check-ins, stakeholders ask counterfactual questions such as *why?* and *what if?*. While deep fusion models are excellent pattern recognisers, their explanations can drift toward after-the-fact narratives. Two design moves help. First, anchor part of your representation in **human-meaningful features** (SCR rate, RMSSD, F0 variability, AU12 intensity) so the model can *speak* physiology when asked to justify a prediction (Mauss, Levenson, McCarter, Wilhelm, & Gross, 2005; Benedek & Kaernbach, 2010; Latif et al., 2021). Second, use **calibrated** outputs with risk bands and decision rules validated on held-out days and subgroups; a calibrated 0.8 must mean *8 out of 10 similar cases were like this*, or your deployment will quietly erode trust (Guo, Pleiss, Sun, & Weinberger, 2017; Can et al., 2023). This is less glamorous than a state-of-the-art leaderboard result, but for artificial psychology, it is the difference between a demo and a system.

A persistent difficulty is **domain shift**: the model learns in quiet rooms and then meets backpacks, video calls, and dialects. We can do better than hope. Lightweight **personalisation** (subject-specific normalisation, a few minutes of fine-tuning, or a small adapter module) often restores performance cheaply (Ringeval et al., 2019; Gideon, Khorram, Aldeneh, M. McInnis, & Provost, 2019). **Semi-supervised** and **self-supervised** updates that continue to learn on-device without labels (e.g., predicting the next segment or aligning modalities that co-occur) reduce drift while keeping raw data private (Fang et al., 2022; Akbari et al., 2021). Newer privacy-preserving techniques, such as on-device representation learning with differentially private updates, are entering the literature and are especially relevant for mental-health contexts (Mireshghallah, Shokri, & Gjoneski, 2022).

What about **text** and **context**, given that Chapter 02 places language alongside the body? In real deployments, textual traces (journals, messages, brief EMA notes) act as a *semantics stabiliser*. When physiology is noisy and face is masked, a sentence fragment (I’m overwhelmed, but okay) can disambiguate arousal’s **valence**. Modern language models fine-tuned for affect provide context embeddings that, fused with physiology, improve robustness, particularly when the world changes (Haq & Jackson, 2011; Poria et al., 2018). Care is needed to prevent leakage of sensitive attributes; model cards should report subgroup performance and any known shortcuts the system might exploit (Mitchell et al., 2019).

**Evaluation** for fusion systems must go beyond accuracy. We must report concordance or F1, but also **calibration**, **temporal stability**, **subgroup parity**, and **ablation** results showing that each modality contributes as intended. Held-out days, held-out devices, and held-out raters form the minimum stress tests for an embodied emotion model claiming to work in the wild (Can et al., 2023; Ringeval et al., 2019). If the model cannot answer, *what happens when the watch dies?* or *what if microphones are muted?*, it is not ready.

In short: choose when to fuse based on synchronisation reality; learn representations that separate what modalities share from what they keep to themselves; pretrain broadly, calibrate carefully, and personalise lightly; treat disagreement between channels as a **psychological** fact, not a modelling nuisance; and evaluate like you mean to deploy. This is how we turn a tangle of signals into a system that earns its keep in classrooms, clinics, and daily life.

**References:**

Acheampong, F. A., Wenyu, C., & Nunoo‐Mensah, H. (2020). Text‐based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, *2*(7), e12189.

Aguilera, A., Mellado, D., & Rojas, F. (2023). An assessment of in-the-wild datasets for multimodal emotion recognition. *Sensors*, *23*(11), 5184.

Akbari, H., Yuan, L., Qian, R., Chuang, W. H., Chang, S. F., Cui, Y., & Gong, B. (2021). Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in neural information processing systems*, *34*, 24206-24221.

Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, *116*, 56-76.

Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, *35*, 23716-23736.

Armstrong, T., & Olatunji, B. O. (2012). Eye tracking of attention in the affective disorders: A meta-analytic review and synthesis. *Clinical psychology review*, *32*(8), 704-723.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.

Barbieri, F., Ronzano, F., & Saggion, H. (2016). What does this emoji mean? a vector space skip-gram model for twitter emojis.

Barker, D., Tippireddy, M. K. R., Farhan, A., & Ahmed, B. (2025). Ethical Considerations in Emotion Recognition Research. *Psychology International*, *7*(2), 43.

Barrett, L. F. (2017). *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt.

Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., & Grundmann, M. (2020). Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*.

Benedek, M., & Kaernbach, C. (2010). Decomposition of skin conductance data by means of nonnegative deconvolution. *psychophysiology*, *47*(4), 647-658.

Botes, M. (2025). Regulatory challenges of digital health: the case of mental health applications and personal data in South Africa. *Frontiers in Pharmacology*, *16*, 1498600.

Boucsein, W. (2012). *Electrodermal Activity.* Springer.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, *10*, 1-47.

Buechel, S., & Hahn, U. (2016, August). Emotion analysis as a regression problem—dimensional models and their implications on emotion representation and metrical evaluation. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence* (pp. 1114-1122).

Calvo, R. A., & D’Mello, S. K. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1), 18–37.

Can, Y. S., Mahesh, B., & André, E. (2023). Approaches, applications, and challenges in physiological emotion recognition—a tutorial overview. *Proceedings of the IEEE*, *111*(10), 1287-1313.

Carvalho, N., Noiret, N., Vandel, P., Monnin, J., Chopard, G., & Laurent, E. (2014). Saccadic eye movements in depressed elderly patients. *Plos one*, *9*(8), e105355.

Chamishka, S., Madhavi, I., Nawaratne, R., Alahakoon, D., De Silva, D., Chilamkurti, N., & Nanayakkara, V. (2022). A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling. *Multimedia Tools and Applications*, *81*(24), 35173-35194.

Champion, Pierre. "Anonymizing speech: Evaluating and designing speaker anonymization techniques." *arXiv preprint arXiv:2308.04455* (2023).

Coan, J. A., & Allen, J. J. (2004). Frontal EEG asymmetry as a moderator and mediator of emotion. *Biological psychology*, *67*(1-2), 7-50.

Cowen, A. S., Elfenbein, H. A., Laukka, P., & Keltner, D. (2019). Mapping 24 emotions conveyed by brief human vocalization. *American psychologist*, *74*(6), 698.

Dael, N., Mortillaro, M., & Scherer, K. R. (2012). Emotion expression in body action and posture. *Emotion*, *12*(5), 1085.

Das, R., & Singh, T. D. (2023). Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*, *55*(13s), 1-38.

Davey, S., Halberstadt, J., & Bell, E. (2021). Where is emotional feeling felt in the body? An integrative review. *PloS one*, *16*(12), e0261685.

Davidson, R. J. (2004). What does the prefrontal cortex “do” in affect: perspectives on frontal EEG asymmetry research. *Biological psychology*, *67*(1-2), 219-234.

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.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5203-5212).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Ekman, P., & Rosenberg, E. L. (Eds.). (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS) (2nd edition)*. Oxford University Press, USA.

Eyben, F., & Schuller, B. (2015). openSMILE:) The Munich open-source large-scale multimedia feature extractor. *ACM SIGMultimedia Records*, *6*(4), 4-13.

Fang, P. F., Li, X., Yan, Y., Zhang, S., Kang, Q. Y., Li, X. F., & Lan, Z. Z. (2022). Connecting the dots in self-supervised learning: a brief survey for beginners. *Journal of Computer Science and Technology*, *37*(3), 507-526.

Farahani, H., Watson, P., Bezan, T., Kovač, N., Winter, L. C., Blagojević, M., ... & Jiménez, P. (2024). PsAIchology: An Intelligent Direction in Psychological Sciences. In *10th International Scientific Conference Technics, Informatics and Education-TIE 2024*. Faculty of Technical Sciences Čačak, University of Kragujevac.

Farahani, H., Watson, P., Kovač, N., Sheykhangafshe, F. B., Azadfallah, P., Allahyari, A., ... & Chesli, R. R. (2025). Unraveling the Complexity of Love Addiction Using Machine Learning Algorithms: The Influence of Positive and Negative Affect, Interpersonal Needs, and Self-Hate. In *International Handbook of Love: Transcultural and Transdisciplinary Perspectives* (pp. 1-22). Cham: Springer Nature Switzerland.

Gendron, M., & Barrett, L. F. (2009). Reconstructing the past: A century of ideas about emotion in psychology. *Emotion Review*, 1(4), 316–339.

Gideon, J., McInnis, M. G., & Provost, E. M. (2019). Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG). *IEEE Transactions on Affective Computing*, *12*(4), 1055-1068.

Gobl, C., & Chasaide, A. N. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech communication*, *40*(1-2), 189-212.

Grecucci, A., Dadomo, H., Salvato, G., Lapomarda, G., Sorella, S., & Messina, I. (2023). Abnormal brain circuits characterize borderline personality and mediate the relationship between childhood traumas and symptoms: a mCCA+ jICA and random forest approach. *Sensors*, *23*(5), 2862.

Gross, J. J. (2015). Emotion regulation: Current status and future prospects. *Psychological inquiry*, *26*(1), 1-26.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, July). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321-1330). PMLR.

Hanisch, S., Arias-Cabarcos, P., Parra-Arnau, J., & Strufe, T. (2021). Privacy-protecting techniques for behavioral biometric data: A survey. *arXiv preprint arXiv:2109.04120*.

Haq, S., & Jackson, P. J. (2011). Multimodal emotion recognition. In *Machine audition: principles, algorithms and systems* (pp. 398-423). IGI Global Scientific Publishing.

Hashem, A., Arif, M., & Alghamdi, M. (2023). Speech emotion recognition approaches: A systematic review. *Speech Communication*, *154*, 102974.

Hazarika, D., Zimmermann, R., & Poria, S. (2020, October). Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 1122-1131).

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2015). *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press.

Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, *29*, 3451-3460.

Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial Intelligence: The global landscape of ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.

Joshi, S., & Gold, J. I. (2020). Pupil size as a window on neural substrates of cognition. *Trends in cognitive sciences*, *24*(6), 466-480.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and trends® in machine learning*, *14*(1–2), 1-210.

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE access*, *7*, 117327-117345.

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449-460.

Klein, M., Witthöft, M., & Jungmann, S. M. (2025). Interoception in Individuals with Autism Spectrum Disorders-A Systematic Literature Review and Meta-Analysis. *Frontiers in Psychiatry*, *16*, 1573263.

Kollias, D. (2022). Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2328-2336).

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2020). Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 2880-2894.

Kossaifi, J., Tzimiropoulos, G., Todorovic, S., & Pantic, M. (2017). AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, *65*, 23-36.

Kovač, N., Ratković, K., Farahani, H., & Watson, P. (2024). A practical applications guide to machine learning regression models in psychology with Python. *Methods in Psychology*, *11*, 100156.

Kovač, N., Ratković, K., Farahani, H., & Watson, P. (2025b). Machine learning regression models for internal shame. Acta Psychologica, 260, 105721.

Kovač, N., Ratković, K., Watson, P., Farahani, H., & Bagheri Sheykhangafshe, F. (2025). Machine learning classification models for predicting chronic pain. *Current Psychology*, 1-14.

Kret, M. E., & Sjak-Shie, E. E. (2019). Preprocessing pupil size data: Guidelines and code. *Behavior research methods*, *51*(3), 1336-1342.

Kröger, J. L., Lutz, O. H. M., & Müller, F. (2020). What does your gaze reveal about you? On the privacy implications of eye tracking. In *IFIP International Summer School on Privacy and Identity Management* (pp. 226-241). Springer, Cham.

Kuppens, P., & Verduyn, P. (2017). Emotion dynamics. *Current Opinion in Psychology*, *17*, 22-26.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious?. *Perspectives on psychological science*, *7*(1), 18-27.

Lahat, D., Adali, T., & Jutten, C. (2015). Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, *103*(9), 1449-1477.

Latif, S., Qadir, J., Qayyum, A., Usama, M., & Younis, S. (2020). Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, *14*, 342-356.

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. (2021). Survey of deep representation learning for speech emotion recognition. *IEEE Transactions on Affective Computing*, *14*(2), 1634-1654.

Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, *13*(3), 1195-1215.

Liang, P. P., Zadeh, A., & Morency, L. P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, *56*(10), 1-42.

Lin, W., & Zhang, Y. (2022, July). Review of researches on the emotion recognition and affective computing based on HCI. In *2022 7th International Conference on Image, Vision and Computing (ICIVC)* (pp. 573-578). IEEE.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L. P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

MacIntyre, A. D. (2022). *The analysis of breathing and rhythm in speech*. University of London, University College London (United Kingdom).

Mai, S., Hu, H., & Xing, S. (2019, July). Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 481-492).

Mathôt, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of cognition*, *1*(1), 16.

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, *5*(2), 175.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., ... & Lucic, M. (2021). Revisiting the calibration of modern neural networks. *Advances in neural information processing systems*, *34*, 15682-15694.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

Mohammad, S. M. (2022). Best practices in the creation and use of emotion lexicons. *arXiv preprint arXiv:2210.07206*.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, *29*(3), 436-465.

Novak, P. K., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. *PloS one*, *10*(12), e0144296.

Orquin, J. L., & Holmqvist, K. (2018). Threats to the validity of eye-movement research in psychology. *Behavior research methods*, *50*(4), 1645-1656.

Pang, H. T., Zhou, X., & Chu, M. (2024). Cross-cultural differences in using nonverbal behaviors to identify indirect replies. *Journal of Nonverbal Behavior*, *48*(2), 323-344.

Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, *72*, 101317.

Pennebaker, J. W., & Chung, C. K. (2011). Expressive writing and its links to mental and physical health. *Oxford handbook of health psychology*, 417-437.

Pereira, R., Mendes, C., Ribeiro, J., Ribeiro, R., Miragaia, R., Rodrigues, N., ... & Pereira, A. (2024). Systematic review of emotion detection with computer vision and deep learning. *Sensors*, *24*(11), 3484.

Pinna, T., & Edwards, D. J. (2020). A systematic review of associations between interoception, vagal tone, and emotional regulation: Potential applications for mental health, wellbeing, psychological flexibility, and chronic conditions. *Frontiers in psychology*, *11*, 1792.

Poria, S., Majumder, N., Hazarika, D., Cambria, E., Gelbukh, A., & Hussain, A. (2018). Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, *33*(6), 17-25.

Poria, S., Majumder, N., Hazarika, D., Ghosal, D., Bhardwaj, R., Jian, S. Y. B., ... & Mihalcea, R. (2021). Recognizing emotion cause in conversations. *Cognitive Computation*, *13*(5), 1317-1332.

Ramaswamy, M. P. A., & Palaniswamy, S. (2024). Multimodal emotion recognition: A comprehensive review, trends, and challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *14*(6), e1563.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., ... & Pantic, M. (2019, October). AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop* (pp. 3-12).

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., ... & Pantic, M. (2017, October). Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge* (pp. 3-9).

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and emotion*, *23*(7), 1307-1351.

Schuller, B., & Batliner, A. (2013). *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons.

Senju, A., Vernetti, A., Kikuchi, Y., Akechi, H., Hasegawa, T., & Johnson, M. H. (2013). Cultural background modulates how we look at other persons’ gaze. *International Journal of Behavioral Development*, *37*(2), 131-136.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. F., & Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, *65*, 3-14.

Stappen, L., Baird, A., Christ, L., Schumann, L., Sertolli, B., Messner, E. M., ... & Schuller, B. W. (2021). The MuSe 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress. In *proceedings of the 2nd on multimodal sentiment analysis challenge* (pp. 5-14).

Steil, J., Koelle, M., Heuten, W., Boll, S., & Bulling, A. (2019, June). Privaceye: privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features. In *Proceedings of the 11th ACM symposium on eye tracking research & applications* (pp. 1-10).

Tan, X., & Zhang, X. L. (2021, June). Speech enhancement aided end-to-end multi-task learning for voice activity detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6823-6827). IEEE.

Thayer, J. F., Åhs, F., Fredrikson, M., Sollers III, J. J., & Wager, T. D. (2012). A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews*, *36*(2), 747-756.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016, March). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5200-5204). IEEE.

Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019, July). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting* (Vol. 2019, p. 6558).

Uono, S., & Hietanen, J. K. (2015). Eye contact perception in the west and east: A cross-cultural study. *Plos one*, *10*(2), e0118094.

VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media.

Wen, M., Dong, Z., Zhang, L., Li, B., Zhang, Y., & Li, K. (2022). Depression and cognitive impairment: current understanding of its neurobiology and diagnosis. *Neuropsychiatric Disease and Treatment*, 2783-2794.

World Health Organization. (2024). *Ethics and governance of artificial intelligence for health: large multi-modal models. WHO guidance*. World Health Organization.

Wu, Y., Mi, Q., & Gao, T. (2025). A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions. *Biomimetics*, *10*(7), 418.

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018, July). Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2236-2246).

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.