8 Integrating Multimodal Data for Comprehensive Understanding of Embodied Emotions

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:** The chapter delves into the profound insights that can be extracted from medical imaging, with a specific focus on CT and MRI data for understanding and identifying emotions. It commences by discussing the potential emotional insights that can be gleaned from these imaging modalities, underlining their significance in the domain of embodied emotion. Subsequently, the chapter explores the intricate process of neuroimaging data preprocessing, an essential step in preparing CT and MRI data for subsequent emotion analysis. Moving forward, it uncovers the emotional patterns hidden within CT and MRI scans, shedding light on how these patterns can be interpreted. Furthermore, the chapter delves into the development of automated emotion detection algorithms tailored for neuroimaging data, substantiating these concepts with compelling case studies and experiments that underscore the practicality and efficacy of this approach.

**Keywords:** Embodied Emotion Across Modalities; Fusing Sensors for Holistic Embodied Emotion Analysis; Machine Learning; Integrating Multimodal Emotion Data

8.1 Introduction

Emotions evolve through interconnected changes in voice, facial expression, physiology, and cognition. Emotions rarely unfold in isolation within a single channel of human experience; rather, they dynamically evolve through coordinated changes in voice, facial expression, autonomic physiology, and cognitive appraisals that are continuously shaped by temporal, contextual, and cultural factors. Each modality, be it vocal prosody, micro facial movements, or cardiac variability, offers a distinct yet complementary vantage point into the underlying affective process (Scherer, 2009). The meaning of emotion thus emerges from the integration of these expressive and physiological modalities, where cross-channel synchrony provides more diagnostic value than any single signal alone (Kragel & LaBar, 2016). Through this multimodal lens, emotion becomes a temporally extended, embodied event rather than a transient, discrete reaction, unfolding across milliseconds in neural and physiological shifts and across minutes or hours in expressive and cognitive alignment.

Recent advancements in affective neuroscience and computational modelling have deepened understanding of this multimodal nature by linking dynamic emotion patterns to both neural circuitry and behavioural expression. For instance, synchronised analyses of facial electromyography, vocal tone, and autonomic signals reveal latent structures of emotional complexity not captured by self-report or single modality observation (Harrison et al., 2013; Celeghin et al., 2017). These integrative approaches enable the discovery of emotion signatures that distinguish adaptive from maladaptive affect regulation, bridging laboratory measures with ecologically valid human experience (Mauss & Robinson, 2009). Importantly, multimodal frameworks recognise that emotional expression and perception are not universal constants but culturally encoded patterns shaped by language, display rules, and social context (Mesquita, 2017). Understanding emotion through these converging modalities provides psychologists and neuroscientists with a more holistic and temporally precise depiction of how affect arises, transforms, and influences behaviour.

As computational and physiological recording tools continue to expand, multimodal data analysis offers a transformative paradigm for both basic and applied psychology. By synthesising cross-temporal records of voice, facial dynamics, neurophysiology, and cognition, researchers can construct multidimensional trajectories that trace the unfolding of emotion in real time. This integration is not simply technical; it represents a conceptual shift toward embracing emotional life as an emergent phenomenon grounded in the complex interaction of brain, body, and social environment. Within clinical and therapeutic settings, such multimodal perspectives promise to reveal early markers of emotional dysregulation, support patient-specific formulations, and extend emotion research from laboratories into the rich temporal texture of everyday human experience.

This interplay is precisely why a multimodal approach has become essential in the empirical study of embodied emotion. Audio and visual streams capture expressive behaviour, while biosignals such as electroencephalography (EEG) and electromyography (EMG) record concurrent changes in neural and peripheral activity. Together, these channels offer complementary perspectives on the same underlying episode, and their integration has consistently been shown to enhance recognition performance compared to single streams, both in laboratory settings and in real-world data. Recent reviews summarise these improvements and attribute them to cross-modal complementarity and error compensation between asynchronous cues, especially when speech prosody and facial actions are combined with physiological signals. This convergence is now extensively documented across benchmark corpora and application contexts (Koromilas & Giannakopoulos, 2021; Pillalamarri & Shanmugam, 2025).

As research has expanded beyond acted dialogues to include spontaneous and web-based sources, four common recording regimes have emerged: covert recordings of spontaneous interactions, acted performances, elicitation through carefully designed tasks, and annotation of public media. Each regime balances control with ecological validity, resulting in different distributional shifts in language, scene dynamics, and emotional intensity. A widely referenced taxonomy maps this landscape and advocates for using diverse datasets to enhance generalisation beyond laboratory environments (Koromilas & Giannakopoulos, 2021).

Within this framework, corpora vary significantly in what they reveal about embodiment. For example, CMU-MOSEI (Zadeh et al., 2018) offers broad-scale diversity and naturalistic speech, making it an ideal testbed for cross-speaker modelling. Conversely, IEMOCAP (Busso et al., 2008) provides carefully recorded audio and motion imagery with strong ground truth from acted dyads. These datasets exemplify the tension between ecological validity and annotation control that underpins multimodal affect research, influencing which models perform best in specific studies (Koromilas & Giannakopoulos, 2021). Methodologically, the field has shifted from rule-based pipelines to representation learning. Deep models can learn the coupled structure across audio, text, and vision, and focus attention on the most informative stream amid noisy data. Recent reviews emphasise that advances in attention mechanisms and representation learning have been crucial for practical performance, though they also increase the demands for data and computation (Hosseini et al., 2024; Ramaswamy & Palaniswamy, 2024; Chandraumakantham et al., 2024).

Fusion remains a key design choice when integrating modalities. Classical taxonomies distinguish early or feature-level fusion, decision-level fusion, and intermediate approaches that combine representations learned from each stream. Comparative reviews show that no single strategy is superior. Feature-level fusion performs well when signals are well aligned and low-dimensional; decision fusion is more robust when channels are missing; and intermediate fusion often outperforms when cross-modal interactions are complex or alignment is weak. Guidance from EEG-focused multimodal reviews suggests choosing feature fusion when timing is critical and cues are complementary, and weighting features by their relative informativeness. Decision fusion is preferable when sensors may fail or drift at different times, and exploring cross-modal attention is advised when interactions are inherently predictive (Pillalamarri & Shanmugam, 2025).

Alignment itself presents a primary challenge. Speech prosody develops over hundreds of milliseconds, facial actions occur at video rates, and neural signatures can lag behind behaviour. Recent work tackles these mismatches with temporal alignment architectures and graph-based fusion that reason jointly about interactions across time and between streams. These techniques lessen reliance on strict synchronisation and have achieved state-of-the-art results on unaligned benchmarks using far fewer parameters than earlier transformers (Koromilas & Giannakopoulos, 2021).

The embodied perspective also underscores the role and limitations of physiology. Physiological signals often provide genuinely complementary information about arousal and regulatory effort, yet they are nonstationary and person-specific. Variability both between and within subjects remains a key obstacle for cross-subject generalisation and has driven a wave of domain adaptation and generalisation techniques. The latest systematic reviews catalogue sources of uncertainty, from measurement to instrumentation and environment, and recommend explicit cross-subject and cross-session evaluation designs (Pillalamarri & Shanmugam, 2025). Evaluation practices in multimodal emotion research reflect this complexity. Class-based metrics such as macro F1 measure balance across categories, while aggregated metrics like accuracy and weighted accuracy track overall performance. Many studies now include agreement-oriented metrics when continuous affect is predicted. The broader recommendation is to adopt speaker-independent splits and report both per-class and overall metrics to avoid overly optimistic estimates due to overfitting on speakers or contexts (Koromilas & Giannakopoulos, 2021; Hosseini et al., 2024).

A final theme in contemporary literature is the shift from benchmarking to deployment. Multimodal systems are entering health education and customer service, where continuous monitoring, adaptation, and empathy are valued. Surveys emphasise that the benefits of multimodal integration come with obligations, including the need for larger and more diverse datasets, careful handling of bias and missing data, and transparent modelling to ensure users and clinicians can trust predictions about sensitive internal states (Chandraumakantham et al., 2024; Hosseini et al., 2024; Pillalamarri & Shanmugam, 2025).

This chapter begins with these themes. The goal is to show how integrating heterogeneous signals produces a more comprehensive view of embodied emotion while being honest about the scientific and engineering trade-offs involved. The subsequent discussion treats fusion as a modelling choice inherently linked to data realities such as alignment, missing data, and subject variability. It considers evaluation as a design decision that must account for both category balance and person independence. Most importantly, it views embodiment not merely as a slogan but as a practical premise. If emotion is spread across voice, face, and body, a science of emotion should be distributed accordingly, employing methods that learn from the strengths and blind spots of each stream together (Ramaswamy & Palaniswamy, 2024; Koromilas & Giannakopoulos, 2021; Pillalamarri & Shanmugam, 2025). This integration aligns with the ‘Artificial Psychology’ (PsAIchology) framing that connects AI methods with psychological science (Farahani et al., 2024).

8.2 Understanding Embodied Emotion Across Modalities

Emotion is not confined to one organ or one output channel. It unfolds as coordinated change across brain, body, and behaviour that can be sensed through movement of the face and body, modulations of the voice, patterns of words, and cascades of neural and peripheral physiology. A modern view treats these concurrent changes as complementary windows into an underlying embodied process rather than competitors for a single correct signal. Evidence for this view spans psychology and affective neuroscience, where vocal prosody, facial action, posture, autonomic physiology, and central electrophysiology are all shown to covary with shifts in affect, and where interoceptive mechanisms link bodily state to subjective feeling. Reviews that survey these response systems emphasise the value of reading emotion through multiple channels in parallel and summarise how text, audio, visual behaviour, and physiology each contribute distinct but interacting information about the same episode of experience (Ramaswamy & Palaniswamy, 2024; Cai, Li, & Li, 2023). Interoception provides a mechanistic bridge by which cardiac and respiratory rhythms, electrodermal activity, and muscle activation inform perception and action, and contemporary syntheses highlight manipulations of interoception as a route to changing affective state and its readouts across modalities (Weng et al., 2021).

In practice, each modality samples a different projection of this latent process. Visual analysis resolves the configuration and dynamics of the face and body, where systems based on action units, landmarks, optical flow, and spatiotemporal deep encoders capture both static morphology and rapid micromovements. Large surveys document steady gains of such models in the wild and their sensitivity to context, lighting, and occlusion. They also underline that body gestures and head pose complement facial cues when expression is subtle or constrained, which motivates joint visual encodings that respond to whole-person behaviour rather than isolated faces (Leong et al., 2023; Du et al., 2019; Canal et al., 2022). Audio carries spectral and temporal correlates of arousal and valence through pitch, intensity, spectral tilt, and rhythmicity. Meta-analyses show that prosody discriminates emotions above chance across languages and that deep models operating on spectrograms exceed classical prosodic baselines, but they also show sensitivity to speaker identity and recording conditions, which motivates domain-aware or self-supervised pretraining before fusion (Juslin, Laukka, & Bänziger, 2018; Kakuba, Poulose, & Han, 2023).

Physiology contributes a complementary slice of the state space. Electrodermal activity indexes sympathetic drive, electrocardiography and photoplethysmography report chronotropic and inotropic control, respiration patterns reflect metabolic and regulatory shifts, and facial electromyography records covert action in zygomatic and corrugator systems that often escape the camera. Electroencephalography resolves oscillatory and event-related dynamics associated with appraisal and attention. Reviews across physiology summarise characteristic signatures at the level of features and their typical reliability, while also warning that single channels are often ambiguous because many non-emotional factors can elicit similar changes. This motivates simultaneous sampling of several peripheral and central channels and model families that can learn the joint structure of their trajectories (Egger, Ley, & Hanke, 2019; Shu et al., 2018; Ramaswamy & Palaniswamy, 2024). A recent critical review of multimodal datasets catalogues how these physiological channels are increasingly paired with audio and video so that emotional episodes can be studied through both behaviour and bodily state on the same time base (Al-Azani & El-Alfy, 2025).

This multimodal orientation is not only theoretical. It is now the practical default in affective computing because no single signal captures the breadth of an episode and because different channels fail in different ways. Large surveys that span more than two hundred papers converge on the same point. Vision excels at specific expression categories but degrades under pose changes. Audio generalises better to distance and partial occlusion but is brittle to noise and dialect. Physiology can be recorded without visible behaviour and is difficult to fake, yet it requires instrumentation and careful artifact control. Text reveals appraisal and goals when available, but is sparse in many interactive settings. The advantage of fusing channels is strongest when the signals bear complementary error patterns and when training data refreshes each channel under the same elicitation and label scheme (Cai et al., 2023; Ramaswamy & Palaniswamy, 2024; Pan et al., 2023; Wu et al., 2025).

Understanding how these channels align in time is essential. Emotional responses recruit systems with different latencies. Facial muscle activity can precede visible expression. Electrodermal responses build more slowly than changes in fundamental frequency. EEG dynamics can move on the order of tens of milliseconds. Dataset builders, therefore, stress synchronized acquisition and careful alignment across devices, since timebase mismatch can inflate apparent discordance among channels and confuse fusion architectures. Public corpora such as MAHNOB-HCI (Soleymani et al., 2011) and RECOLA (Ringeval et al., 2013) have become reference points because they provide audio, video, and physiology recorded in lockstep with consistent annotation scales, and they have catalysed a generation of multimodal benchmarks (Cai et al., 2023). Recent surveys underline that collecting such corpora is costly and that many teams still rely on bespoke data, which limits comparability and external validity (Al-Azani & El-Alfy, 2025).

A second axis of understanding concerns representation. Classical pipelines derived hand-crafted features such as mel-frequency cepstral coefficients for audio, action units for video, power in canonical EEG bands, or time-domain statistics for heart rate variability. These features still serve as strong baselines and, when carefully engineered, encode domain insight that helps when data are scarce or labels are noisy. The field has nevertheless shifted toward deep encoders that learn hierarchical representations directly from raw or lightly processed streams. On the visual side, this includes three-dimensional convolutional encoders and attention models that summarise facial dynamics. On the audio side, convolutional transformers operating on spectrograms or learnable filterbanks provide state-of-the-art performance. For physiology, temporal convolutional networks and transformers learn patterns in multichannel signals with fewer assumptions about stationarity. Comparative studies show that deep encoders, when trained or adapted with self-supervised objectives, generally outperform classical features and reduce the need for manual signal processing, although they may be more brittle under distribution shift (Rouast, Adam, & Chiong, 2021; García-Hernández et al., 2024; Montero Quispe et al., 2022; Zhang et al., 2024).

These representational advances matter most when channels are combined. Fusion can occur at the feature level, the decision level, or through joint sequence models that pass attention across streams. Surveys and benchmarks chronicle steady gains from moving beyond naive concatenation toward architectures that explicitly model cross-modal influence. Examples include cross-modal attention, where audio cues steer the model toward high-motion parts of the face, and conditioning, where physiological signals adjust the visual encoder to highlight features linked to arousal. Graph-based and transformer frameworks now dominate, and head-to-head comparisons report that hybrid designs that mix early and late fusion often offer the best tradeoff between flexibility and robustness to missing modalities (Jiang et al., 2020; Liu et al., 2021; Wu et al., 2025; Cai et al., 2023). The most recent overviews echo the same conclusion and locate the next frontier in self-supervised pretraining, weak supervision from large but noisy web corpora, and robust handling of missing or asynchronous streams (Ramaswamy & Palaniswamy, 2024; Zhang et al., 2024; Pan et al., 2023).

Understanding embodied emotion also requires attention to the structure of labels. Discrete schemes such as the seven basic categories are widely used for interpretability and have driven progress in video- and audio-based recognition. Dimensional schemes in arousal and valence align more naturally to physiology and allow continuous annotation. Contemporary surveys recommend treating labels as complementary and, where possible, learning shared representations that support both category and dimension outputs to encourage generalisation across contexts and datasets (Ramaswamy & Palaniswamy, 2024; Cai et al., 2023). The choice of label space also shapes evaluation. Macro-averaged metrics are preferred when class priors are balanced, while the area under the Receiver Operating Characteristic (ROC) curve and average precision reveal discrimination beyond a single threshold and are essential when the task is imbalanced or when calibrated probabilities are required for decision support. Reviews of benchmarking practice call for reporting both accuracy and probabilistic quality measures, such as the Brier score, to capture overconfidence and for auditing subgroup performance to surface cultural and demographic sensitivity (Ahmed, Aghbari, & Girija, 2023; Al-Azani & El-Alfy, 2025).

Reviews identify three recurring ethical risks for emotion recognition. Katirai argues that systems can yield inequitable outcomes when built on questionable assumptions, that they process sensitive emotional data with attendant privacy risks, and that deployment in domains such as employment or healthcare can produce harm (Katirai, 2024). A recent analysis of multimodal emotion datasets echoes these points, noting how consent and licensing restrictions limit dataset distribution and reproducibility and underscoring the importance of explicit informed consent for multimodal recordings (Al-Azani & El-Alfy, 2025). These concerns are not an afterthought. They shape data collection protocols and motivate the multimodal route because it allows selective emphasis on less identifying channels, such as peripheral physiology, when visual or audio capture is inappropriate.

Taken together, these strands outline a coherent picture of embodied emotion across modalities. Visual, vocal, textual, and physiological views each provide a partial but meaningful slice of the same latent process. Their temporal dynamics and error profiles differ in ways that make them natural partners. Public datasets and modern deep sequence models have made it practical to align and learn from them jointly. As the surveys repeatedly note, progress now depends on synchronized and diverse corpora, on architectures that model cross-modal influence while tolerating missing data, and on evaluation that values calibrated probabilities and fairness alongside headline accuracy (Cai et al., 2023; Ramaswamy & Palaniswamy, 2024). This integrative perspective frames the rest of the chapter, where methods for sensing, representing, and fusing these signals are developed with the goal of a more faithful account of affect as it is enacted by the whole person.

8.3 Fusing Sensors for Holistic Embodied Emotion Analysis

The promise of multimodal emotion science rests on a simple observation. No single sensor captures the breadth of how bodies register and express affect. Speech carries prosodic contours that mark arousal and valence. Faces and posture reveal rapid motoric adjustments that accompany appraisal. Brain and peripheral physiology index covert regulation and autonomic change. When these streams are analysed in concert, the result is often more accurate and more robust than any one channel alone. Reviews across affective computing consistently document this complementarity and show that fusion is the organising principle that turns heterogeneous signals into a coherent account of embodied emotion (Ramaswamy & Palaniswamy, 2024). Related machine-learning frameworks have also been applied to affect-linked constructs beyond core emotions, such as love addiction, where features and explanations clarify predictive factors (Farahani et al., 2025).

Fusing sensors begins with a design choice about representation. Early fusion concatenates features from each modality to form a single joint vector, which is then fed to a classifier. This strategy gives the learner maximal access to cross-modal interactions, but it is sensitive to scale, missing values, and the curse of dimensionality effects. Late fusion trains a specialist per modality and combines their outputs by averaging, stacking, or learned gating. This strategy is robust to sensor dropouts and allows per-modality calibration at the expense of modelling rich cross-terms. Hybrid or intermediate fusion blends both by learning modality-specific encoders and then integrating them through attention, tensor factorisation, or graph message passing in a mid-level space. Contemporary surveys in affective and multimodal sentiment analysis trace these patterns and catalogue their strengths and limits with numerous exemplars across audio, vision, language, and physiology (Lian et al., 2023a; Gandhi et al., 2023; Zhao et al., 2025).

The theoretical rationale for fusion follows two complementary arguments: (1) the information argument holds that modalities carry partly non-overlapping cues. Acoustic energy in higher frequency bands is sensitive to breathy or tense phonation, while corrugator activity and frontalis activity in facial electromyography track subtle negative and surprise-related dynamics, even when faces are partially occluded. When these cues are pooled, the joint representation increases mutual information with the latent emotion; (2) the reliability argument holds that when noise or occlusion affects one channel, others can stabilise inference. Evidence theory and decision level combination frameworks formalise this idea through rules that weight sources by credibility and consistency, and these ideas have informed multimodal inference for decades and continue to appear in modern deep fusion systems (Lian et al., 2023a).

In practice, embodied emotion fusion is shaped by temporal structure. Emotion unfolds over hundreds of milliseconds to many seconds, and different sensors sample at different rates. For a concrete illustration, one public multimodal corpus records audio at 44.1 kHz while video is captured at ten frames per second, alongside high-frequency biosignals, which immediately creates heterogeneous time bases that a model must reconcile (Chen et al., 2022b). Surface electromyography in the same corpus is sampled at ten kilohertz with a fifty to one thousand hertz bandpass, and electroencephalography is sampled at five kilohertz and then downsampled to one kilohertz for analysis, which underscores the need to align fast neural or muscular activity with much slower facial and vocal streams (Chen et al., 2022b). Temporal encoders, therefore, need to summarise and align information across scales either before fusion or inside the network. Surveyed architectures caution that forcing strict word-level or frame-level synchrony can be brittle and instead advocate learning correspondences directly from data, a view that motivates cross-modal attention over loosely aligned sequences and avoids spurious matches introduced by naive resampling (Zhao et al., 2025). In practice, modality-specific temporal embeddings are learned with one-dimensional convolutions for local dynamics, recurrent units for longer dependencies, and self-attention for global context, after which cross-modal attention layers exchange information on a shared time base (Zhao et al., 2025). Contemporary reviews of multimodal emotion recognition further document transformer-based fusion that learns inter- and intra-modal relations without explicit hard alignment, consolidating evidence that careful temporal modelling is central to reliable fusion in affect analysis (Ramaswamy & Palaniswamy, 2024).

Attention-based fusion has become a dominant pattern because it allows the model to learn when to listen to which sensor. Cross-modal attention can project audio keys and queries into a space where visual values are retrieved or the inverse, which yields localised correspondences between prosodic bursts and micro facial actions. Hierarchical attention can first weight features within each modality, then integrate across modalities. Recent transformer architectures extend this idea with shared or alternating attention blocks that pass information among streams while preserving modality identity. Reviews of affective modelling and multimodal sentiment analysis document strong performance gains from such attention mechanisms across laboratory and in the wild settings (Ramaswamy & Palaniswamy, 2024; Gandhi et al., 2023; Zhao et al., 2025).

The choice among early, late, and hybrid fusion depends on data scale, annotation quality, and application constraints. Early fusion shines when there is adequate data to learn cross-terms and when modalities are reliably present for most trials. Late fusion is preferred when sensors are unevenly available, when calibration and interpretability at the decision level matter, or when one wishes to swap modalities without retraining the entire stack. Hybrid fusion becomes attractive when we expect sparse but important interactions, such as moments when a short vocal burst aligns with a brief brow raise. Comprehensive surveys recommend matching the fusion strategy to the error structure of the modalities and to the target deployment setting rather than adopting a single recipe (Lian et al., 2023a; Le et al., 2023).

Handling missing or asynchronous inputs is central in embodied emotion projects. In real environments, a camera may be occluded or a microphone may clip while physiological sensors drift or disconnect. Robust pipelines, therefore, include imputation at the feature level, a mixture of experts with a learned gating network that downweights missing experts, or architectures explicitly trained with random modality dropout so that the network learns to operate with any subset of inputs. Reviews report that such techniques reduce failure rates and allow graceful degradation when sensors are unreliable, which is essential in mobile or clinical contexts (Le et al., 2023; Ramaswamy & Palaniswamy, 2024).

A second practical axis concerns the feature space. Handcrafted features remain competitive for physiological channels where domain knowledge encodes meaningful spectral bands and summary statistics. For example, discrete cosine or band power features in electroencephalography and amplitude dispersion in facial electromyography are standard and are often concatenated with learned embeddings from audio or vision encoders. On the learned side, consistency, specificity objectives encourage networks to retain a unique modality structure while aligning shared emotion-relevant factors. Information bottleneck formulations likewise force intermediate representations to preserve cross-modal signals that predict affect while discarding nuisance variation. Recent work in multimodal fusion for affect leverages these principles to improve generalisation and reduce overfitting to superficial cues such as illumination, identity, or background noise (Zhao et al., 2025; Ramaswamy & Palaniswamy, 2024).

Embodied emotion often requires integration of physiology with behaviour. Classic affective corpora such as DEAP (Koelstra et al., 2011), DECAF (Donahue et al., 2014), and AMIGOS (Miranda-Correa et al., 2018) illustrate how electroencephalography and peripheral signals complement vision and audio, and they continue to motivate fusion strategies that respect both slow autonomic trends and fast motoric expressions. Surveys of datasets underscore that these collections differ in sampling rates, annotation granularity, and synchronisation fidelity, which again shapes fusion designs. Models that first regularise within modality, then combine decisions or mid-level embeddings have proven to be reliable choices across such varied conditions (Cai et al., 2023).

Interpretability is not optional in health, education, or safety applications. Decision-level fusion supports modality attribution, which helps analysts see whether a classification leaned primarily on voice prosody, facial motion, or physiological changes. Mid-level fusion can be probed with attention maps and sensitivity analysis to reveal cross-modal interactions, for example, a case where the model attends jointly to a sharp rise in pitch and a transient frontalis activation when labelling surprise (Ramaswamy & Palaniswamy, 2024). Recent reviews recommend reporting per-modality contributions and explanation diagnostics alongside accuracy to catch spurious correlations and demographic shortcuts and to promote trust in multimodal affect systems (Udahemuka et al., 2024).

Evaluation practices must reflect both accuracy and reliability. When fusion is successful, macro-averaged F1 and area under precision-recall curves typically improve because minority classes benefit from signals that are salient in only some channels. Yet probability, quality and calibration remain important because practitioners act on model confidence as well as labels. Many fusion pipelines, therefore, calibrate per modality posteriors before combination or calibrate the fused output after stacking. Reviews recommend reporting discrimination and calibration together to give a complete picture of utility in downstream decisions that affect people (Ramaswamy & Palaniswamy, 2024).

Finally, embodied emotion work benefits from a disciplined engineering perspective. Data preprocessing should harmonise sampling rates and apply sensor-specific denoising before any joint modelling. Feature spaces should be standardised so that early fusion is not dominated by a single high-variance stream. When late or hybrid fusion is used, per-modality validation ensures that the integrated system does not hide a failing sensor behind a strong counterpart. Across recent literature, one sees a convergence toward modular pipelines in which each stream has its own encoder and diagnostics, and a fusion block learns when and how to combine them, a view that aligns with the natural division of embodied signals into articulatory, visual motor, and physiological families (Gandhi et al., 2023; Le et al., 2023).

Figure 1 depicts the behaviour of a two-expert late-fusion scheme across mixing weights and shows a clear maximum near equal weighting, a pattern consistent with complementary signals whose errors are not fully shared. When the weight moves from either extreme toward an even split, performance first rises and then levels, peaking near the middle. If the two experts make mistakes independently often enough, averaging their probabilities suppresses idiosyncratic noise and reduces variance, so the fused decision achieves a higher macro F1 than either expert alone. If the experts are made more similar or their errors become tightly coupled, the curve flattens and the gain disappears. That contrast is the essence of useful multimodal fusion. The benefit does not come from stacking models for its own sake. It comes from combining signals that carry complementary evidence about the same underlying emotional state.

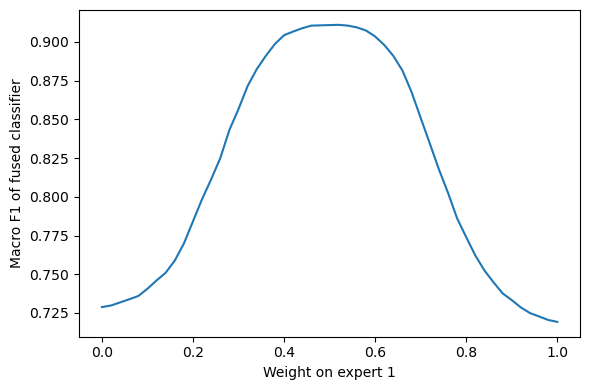


Figure 1. Weighted late fusion of two calibrated experts

This intuition links directly to the practice guidelines that close this section. Embodied emotion work benefits from a disciplined engineering perspective. Data preprocessing should harmonise sampling rates and apply sensor-specific denoising before any joint modelling. Feature spaces should be standardised so that early fusion is not dominated by a single high-variance stream. When late or hybrid fusion is used, per-modality validation ensures that the integrated system does not hide a failing sensor behind a strong counterpart. Across recent literature there is a convergence toward modular pipelines in which each stream is processed by its own encoder with per-modality diagnostics and a dedicated fusion block learns how and when to combine these representations (Zhao et al., 2024; Udahemuka et al., 2024), a view that aligns with the standard grouping of embodied signals into behavioral articulatory and visual-motor channels such as facial and vocal cues and physiological channels both central and peripheral such as EEG electrodermal and cardiovascular measures (Ramaswamy & Palaniswamy, 2024). Taken together, these principles define fusion as the central act in multimodal emotion inference. When we respect the temporal character of each sensor, choose fusion mechanisms that match the data regime, plan for missing inputs, and audit interpretation and calibration, integrated systems yield a richer and more trustworthy view of embodied affect than any single stream.

Taken together, these principles define fusion as the central act in multimodal emotion inference. When we respect the temporal character of each sensor, choose fusion mechanisms that match the data regime, plan for missing inputs, and audit interpretation and calibration, integrated systems yield a richer and more trustworthy view of embodied affect than any single stream. This is not simply a performance claim. It reflects the fact that emotions are embodied in distributed and partially independent physiological and behavioural processes. Fusion aligns our analytical tools with that biological reality and is therefore the cornerstone of a comprehensive science of embodied emotions (Ramaswamy & Palaniswamy, 2024; Lian et al., 2023a).

8.4 Machine Learning Approaches for Integrating Multimodal Emotion Data

A learning system that integrates audio, vision, language, and physiology must cope with signals that differ in sampling rate, statistical structure, and noise profile, yet still produce a single coherent estimate of an underlying affective state. This integration problem is not only about adding more features. It is a modelling task that addresses alignment across time, representation within each sensor, and principled fusion under uncertainty. Contemporary research converges on three building blocks: (1) modality-specific encoders transform raw streams such as spectrograms, facial landmarks, or band power from EEG into compact embeddings; (2) a fusion mechanism exchanges information across embeddings while preserving modality-specific cues; (3) the output head produces calibrated class probabilities or continuous affect scores and is evaluated with discrimination and reliability metrics. Surveys across affective computing consistently show that performance gains emerge when these elements are designed together rather than in isolation, particularly when physiological and behavioural channels are combined in a single learning objective (Ramaswamy & Palaniswamy, 2024; García-Hernández et al., 2024). For psychology-focused, Python-based pipeline patterns and regression modelling, see (Kovač et al., 2024).

A first layer of design concerns representation learning inside each modality. In speech, mel frequency cepstral coefficients, prosodic contours, and log-mel filterbank features remain strong inputs for classical learners and are also used as front ends for convolutional and transformer encoders that model long-range temporal patterns through self-attention (Lian et al., 2023b; García-Hernández et al., 2024). In the visual stream, modern pipelines start from face detection and landmark tracking, then pass frames or short clips through spatiotemporal networks that capture action units and micro-movements predictive of emotion, with attention replacing hand-crafted temporal pooling in many recent systems (Lian et al., 2023b). Physiological channels require different preprocessing. For EEG, short windows are transformed into spectral power over canonical bands delta, theta, alpha, beta, and gamma or into entropy and asymmetry indices, then aggregated across regions before classification. Reviews in sensing and biomedical engineering emphasise that band power and entropy families are the most reproducible correlates for affect classification, especially when combined with robust normalisation and artefact rejection (García-Hernández et al., 2024).

Fusion is the second design layer. Early fusion concatenates features across modalities before learning and is effective when streams are well aligned. Late fusion trains a separate model per modality and combines their outputs with a meta-learner or with probability rules. Hybrid fusion inserts cross-modal interactions between modality-specific encoders and the final decision, enabling both local and global integration. Standard taxonomies define these three families and document their tradeoffs in alignment sensitivity, parameter efficiency, and robustness to missing data, with hybrid strategies often achieving the best balance on modern corpora (Lian et al., 2023b; Zhao et al., 2025; Al-Azani & El-Alfy, 2025).

Classical learning remains a strong baseline for integrated emotion analysis. Comparable supervised pipelines are also used across psychological health domains, for example the classification of chronic pain outcomes (Kovač et al., 2025). Support vector machines, logistic regression, random forests, and gradient boosting perform well when paired with careful feature scaling and cross-validation. In a late fusion setting, a stacked generalizer trained on out-of-fold predictions from modality-specific base models can exploit complementary strengths without demanding time alignment at the feature level. Stacking has a long history as an error-reducing meta-learning device and fits naturally when modalities differ in dimensionality and sampling rate (Wolpert, 1992). For probabilistic outputs, Platt scaling and isotonic regression convert raw scores to calibrated probabilities so that a fused system can weight modalities on a common scale and reason about confidence (Platt, 1999; Zadrozny & Elkan, 2002; Guo et al., 2017).

Deep learning has shifted the frontier by learning both representation and fusion end-to-end. Tensor-based fusion layers preserve unimodal signals while modelling multiplicative interactions, as in the Tensor Fusion Network (Zadeh et al., 2017), which factorizes cross-modal dynamics without manual feature design and has influenced subsequent affect models (Zadeh et al., 2017). Cross-modal transformers learn to align asynchronous streams by attending from one modality to another, replacing explicit alignment with data-driven correspondences. In multimodal sequence learning, attention blocks handle different sampling rates and variable lag, which is essential in emotion, where physiological responses trail facial or vocal cues. This approach is now widely cited for unaligned language-vision-audio sequences and has been adapted to affective computing tasks (Tsai et al., 2019; Lian et al., 2023b). Recent reviews document a rising use of attention and transformer families in emotion recognition, reflecting their ability to capture long-range dependencies and cross-modal context (García-Hernández et al., 2024).

Representation learning increasingly benefits from pretraining and self-supervision. Contrastive objectives align audio embeddings with visuo-facial or textual embeddings drawn from the same event, while pushing apart mismatched pairs. The goal is a modality-invariant yet content-sensitive space that supports transfer to downstream emotion labels with few annotations. Surveyed evidence shows that such objectives improve generalisation and robustness, particularly when paired with augmentations that mimic real recording conditions, for example, reverberation in speech or occlusion in faces (Gandhi et al., 2023; Ramaswamy & Palaniswamy, 2024).

A practical challenge in deployed systems is missing or degraded modalities. Cameras occlude, microphones saturate, electrodes drift. Late fusion can simply drop a base model at test time, yet hybrid and end-to-end systems must be trained to handle absence. Two strategies recur in the literature. The first is modality dropout, which randomly removes streams during training so the network learns to rely on whichever channels are present (Li, 2025). The second is teacher-student distillation (Touvron et al., 2021), where a full model teaches a reduced model through soft targets so that the student can operate with fewer inputs at inference. Reviews emphasise that robustness to missing channels is not automatic and should be validated explicitly with controlled ablations (Yoon & Kim, 2025; Lian et al., 2023b).

Learning targets and evaluation deserve equal attention. Categorical labels such as anger or joy can be complemented with dimensional targets such as valence and arousal, and multitask heads enable a model to share information across these views of affect. Performance should include macro-averaged discrimination metrics that respect class balance and explicit measures of probability quality. In emotion applications where decisions guide feedback to users, calibrated probabilities matter as much as raw accuracy. The Brier score (Brier, 1950) summarises overall probabilistic accuracy, while reliability diagrams and expected calibration error quantify systematic bias in confidence. Modern neural networks are often miscalibrated out of the box, yet simple post-hoc calibration improves decision quality without changing classification, which aligns with best practice for safety-critical or user-facing affect systems (Guo et al., 2017). Comparable psychology-focused pipelines using calibrated regression have been applied to predict internal shame, providing an example of explainable, probability-aware modelling in an affect domain (Kovač et al., 2025a).

The effect of post-hoc calibration on late-fusion reliability is illustrated in Figure 2. The dashed diagonal marks perfect calibration. Figure 2 (A) plots predicted confidence against observed accuracy for the uncalibrated classifier. The curve sits above the diagonal at mid to high confidence, which indicates overconfidence and yields an Expected Calibration Error (ECE) of 0.099 and a Maximum Calibration Error (MCE) of 0.230. Figure 2 (B) shows the same model after isotonic calibration. Points now align with the diagonal across the confidence range, and summary errors fall to ECE of 0.009 and MCE of 0.019 while discrimination is unchanged. This is the intended outcome of post-hoc calibration, where probabilities are adjusted without altering class decisions or macro F1, consistent with best practice in probabilistic evaluation for classification (Brier, 1950; Guo, Pleiss, Sun, & Weinberger, 2017).

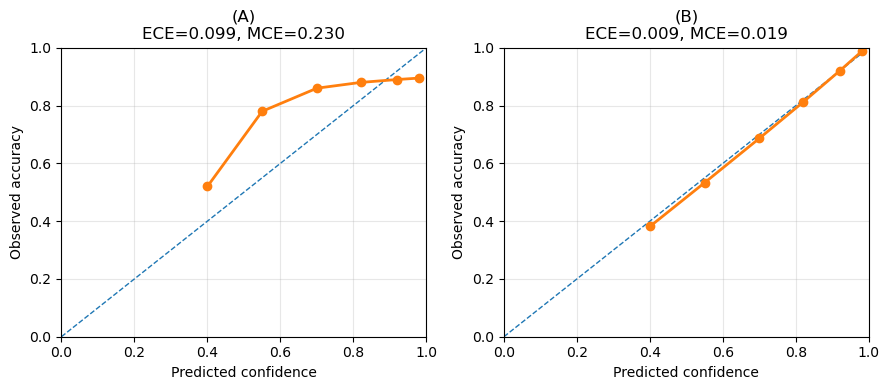


Figure 2. Reliability before and after calibration: (A) uncalibrated late-fusion reliability curve; (B) the same classifier after isotonic calibration

Interpretability closes the loop between sensors and models. Permutation importance and SHapley Additive exPlanations (SHAP) values expose which variables drive decisions and whether these align with intended psychophysiological theory. Attention maps localise the temporal segments or facial regions that contribute most to predictions. Surveys in multimodal sentiment and emotion analysis encourage reporting these diagnostics alongside accuracy to ensure that gains from fusion do not arise from artefacts, for example, speaker identity or background music, and to support fair comparisons across datasets and populations (Lian et al., 2023b; García-Hernández et al., 2024).

Finally, there is growing interest in responsible machine learning for affect (Baltrušaitis et al., 2018). This includes bias assessment across gender, age, and culture, uncertainty reporting, and clear provenance of training data. Reviews stress that multimodal systems amplify the risk of shortcut learning because multiple channels can entangle spurious cues. Robust pipelines therefore include stratified evaluation, cross-subject validation when appropriate, and transparent release of code and preprocessing steps to support replication (García-Hernández et al., 2024; Ramaswamy & Palaniswamy, 2024).

8.5 Case Studies in Multimodal Embodied Emotion

Here, we translate the chapter’s ideas into practice by examining a specific case study based on a carefully selected multimodal dataset. The dataset used is PME4 (Chen et al., 2022a), collected to investigate how discrete emotions are expressed through coordinated streams of behaviour and physiological responses. Eleven adult participants with acting experience produced seven labelled expressions, following Ekman’s basic set plus neutral. The labels include anger, fear, disgust, sadness, happiness, surprise, and neutral. Each trial lasted five seconds, with brief pauses in between, and the sessions were divided into five blocks to ensure each participant contributed multiple repetitions per emotion. Data collection is genuinely multimodal. When a participant spoke the sentence, *The sky is green*, four synchronised sensors recorded data simultaneously. A Universal Serial Bus (USB) webcam captured facial video at ten frames per second with 960 by 720 resolution. The laptop microphone recorded audio at 44.1 kHz. Eight scalp EEG electrodes at F3, Fz, F4, Cz, P3, Pz, P4, and O2 measured brain activity, bandpass filtered within the standard low-frequency range for cognitive research. Six facial EMG channels captured muscle activity from muscles involved in visible facial actions related to expression. After basic quality checks, the public dataset contains 3829 valid trials (Chen et al., 2022b). This design provides closely aligned audio, visual, neural, and muscular evidence for the same emotional state, making it ideal for exploring how different channels complement each other and testing whether combining them enhances recognition beyond any single stream.

The PME4 collection was developed to fill a gap in existing benchmarks. Many emotion datasets depend on spontaneous reactions to films or music and then rely on subjective ratings. PME4 uses posed expressions given by trained actors, which reduces ambiguity in labels and results in better alignment between what the subject intends to express and what the sensors detect. The protocol also maintains a useful structure within each trial. There is a short interval before speech begins, then the utterance itself, followed by a brief tail, allowing researchers to analyse pre-speech, during-speech, and post-speech dynamics if desired. The hardware choices keep the setup straightforward enough to be replicated while still capturing the essential data. Audio and video display the overt performance that people tend to interpret naturally. EEG and EMG record internal timing and muscle activation patterns that can be subtle but informative, especially when faces are partially obscured by electrodes. Collectively, these streams enable us to explore not only which modality performs best on its own but also how they interact.

In this section, the PME4 dataset acts as a testbed for an end-to-end multimodal analysis of embodied emotion. We create a unified trial-level table that combines features from audio, video, EEG, and EMG, along with the emotion label and the subject identifier. From this foundation, we develop interpretable baselines for each modality and a late fusion model that merges the streams. Generalisation is tested using a conventional train and test split, as well as a leave-one-subject-out (LOSO) protocol to assess robustness across individuals. Reliability is evaluated alongside accuracy through probability calibration, confidence distributions, and class-specific behaviour, making the quality of predicted probabilities transparent. The case study aims to show, within a single reproducible pipeline, how modelling emotion as a joint phenomenon across coordinated channels can lead to measurable improvements in predictive power and trustworthiness when tested on a public and well-documented resource.

We analyse the PME4 corpus of synchronised audio, video, EEG and EMG trials, reshaped into a single trial-level table with subject identifiers and one of seven emotion labels (anger, disgust, fear, happy, neutral, sad, surprise). Figure 3 A summarises the label distribution. The seven bars are nearly identical in height, with counts between about 545 and 548 per class, which confirms that the dataset is effectively balanced and that macro-averaged metrics are appropriate. Figure 3 B reports the headline result. The best single stream in this run is audio with a macro F1 of 0.439, whereas the late-fusion model reaches 0.466. The absolute gain is Δ ≈ 0.026, a modest but meaningful improvement that supports the claim that combining modalities corrects a portion of the errors each stream makes on its own.

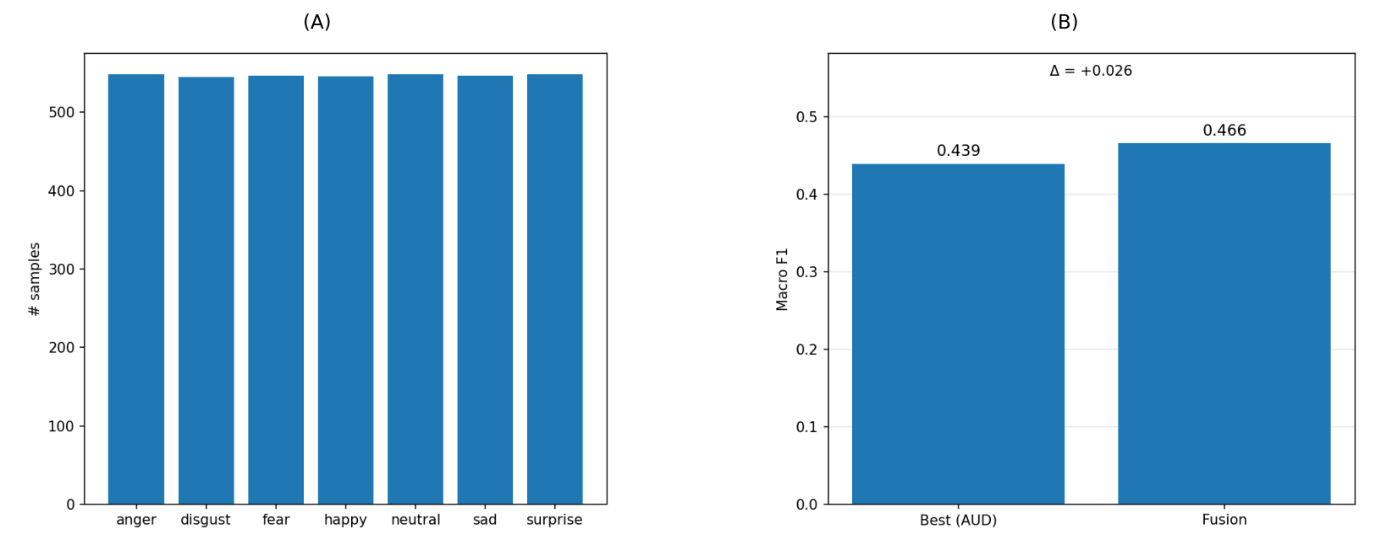


Figure 3. Dataset overview and headline fusion result: (A) class counts in the trial-level table; (B) macro F1 for the best single modality and the late-fusion model

For tables, we use abbreviated column names: *Acc* is accuracy, *F1m* is macro F1, *Prec\_m* and *Rec\_m* are macro precision and recall, *AUC\_m* is the macro one-versus-rest ROC Area Under the Curve (AUC), *AP\_m* is macro average precision, *Brier* is the mean squared error of the probability assigned to the true class, and *LogLoss* is the cross-entropy on the full probability vector. The quantitative headline is summarised in Table 1. The table reports all the mentioned metrics. Consistent with Figure 3 B, the late-fusion model edges the best single modality on macro F1 by about +0.026 and shows comparable or better probabilistic quality in the Brier and log-loss measures, where lower is better. This gives a compact view of both classification accuracy and calibration for the two headline systems.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Acc** | **F1\_m** | **Prec\_m** | **Rec\_m** | **AUC\_ovr\_m** | **AP\_m** | **LogLoss** | **Brier (true-class)** | **Brier (multiclass)** |
| **LateFusion\_Stacking** | 0.466 | 0.466 | 0.467 | 0.466 | 0.819 | 0.488 | 1.446 | 0.508 | 0.096 |
| **AllFeatures\_HGB** | 0.615 | 0.615 | 0.618 | 0.615 | 0.902 | 0.69 | 1.445 | 0.331 | 0.082 |
| **LateFusion\_Stacking+Cal** | 0.465 | 0.464 | 0.466 | 0.465 | 0.814 | 0.486 | 1.574 | 0.531 | 0.097 |
| **AllFeatures\_HGB+Cal** | 0.621 | 0.621 | 0.627 | 0.621 | 0.897 | 0.687 | 1.131 | 0.4 | 0.076 |

Table 1. Headline test metrics for the best single modality and the late-fusion model

Taken together, the balanced label distribution in Figure 3 A and the fusion advantage in Figure 3 B motivate the rest of the case study. The data offer a fair test of multimodal learning, and the initial comparison shows that integrating audio, video, EEG, and EMG pays off at evaluation time.

Figure 4 examines the structure of mistakes and shows when combining streams pays off. Figure 4 (A) presents the row-normalised confusion matrix for the late-fusion model. The diagonal cells sit in the mid-forties to low fifties for most classes, which is expected in a seven-way task with subtle boundaries. The off-diagonal pattern is not random. Fear and sadness are frequently confused in both directions, and neutral often drifts toward sadness. Surprise leaks into fear and happiness, while anger and disgust are mixed more than other pairs. These confusions cluster along affective similarity, which suggests that unimodal cues are sometimes insufficient to separate neighbours in valence or arousal.

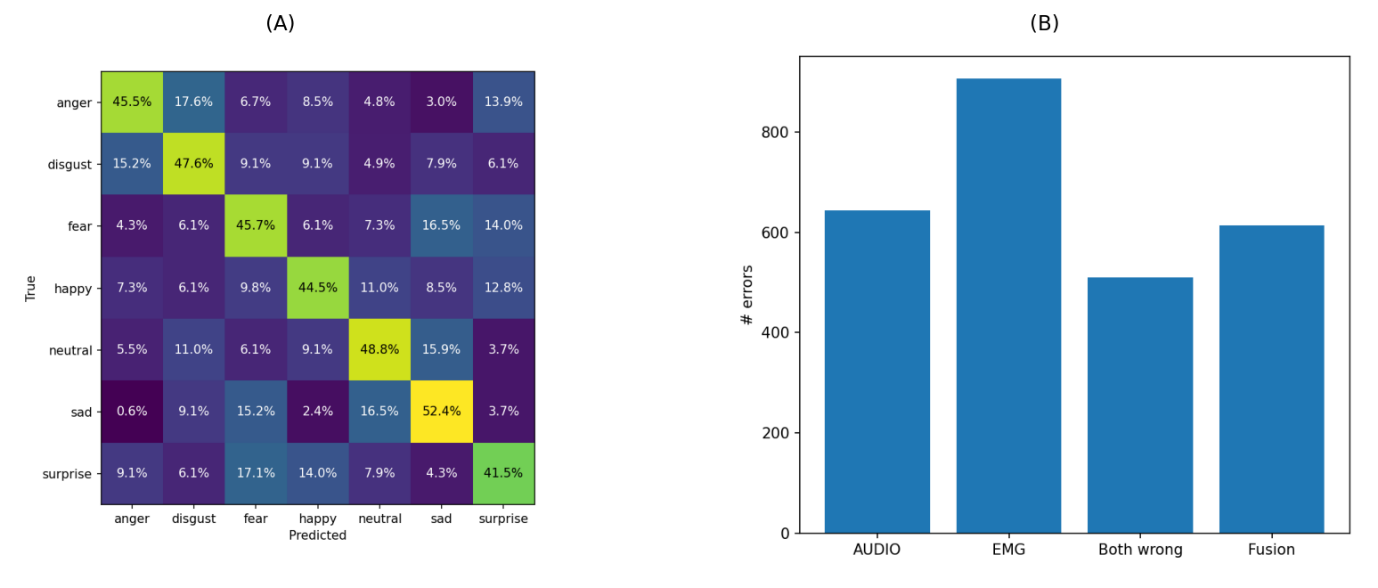


Figure 4. Error structure and fusion gains: (A) confusion matrix for the late-fusion model with rows normalised to percent per true class; (B) error counts for audio and EMG single-stream models, the overlap where both are wrong, and the late-fusion model.

Figure 4 (B) compares error counts for the two strongest single streams and for fusion using the same test split. Audio makes 643 mistakes and EMG makes 906. The overlap where both are wrong is 510, which marks the hardcore of cases that neither stream can resolve on its own. The fusion model makes 614 mistakes, fewer than EMG by about thirty-two per cent and slightly fewer than audio by 29 errors. This indicates that many errors are complementary. When audio is uncertain but EMG carries discriminative muscle activity, the fusion learner can follow the reliable cue, and vice versa. The residual 510 errors align with the densest off-diagonal bands in Figure 4 (A), such as fear versus sadness and neutral versus sadness, where both streams often point the wrong way. Those pairs are candidates for additional information from EEG or facial video or for more tailored features.

To understand how well the late-fusion model separates each emotion from the rest across decision thresholds, we examine receiver–operator behaviour and the associated precision–recall trade-offs on the held-out test set. Figure 5 (A) plots one-vs-rest ROC curves for all seven categories. The traces arch well above the diagonal reference, which indicates that the fused signal carries discriminative structure for every class. The spacing between curves is informative. Emotions with more stereotyped vocal or facial patterns tend to trace higher true-positive rates at the same false-positive rate, reflecting clearer separation in the fused feature space. Classes with broader acoustic and physiological expression display shallower segments of the curve, which signals that additional context or features may be needed to push the operating point further toward the top left.

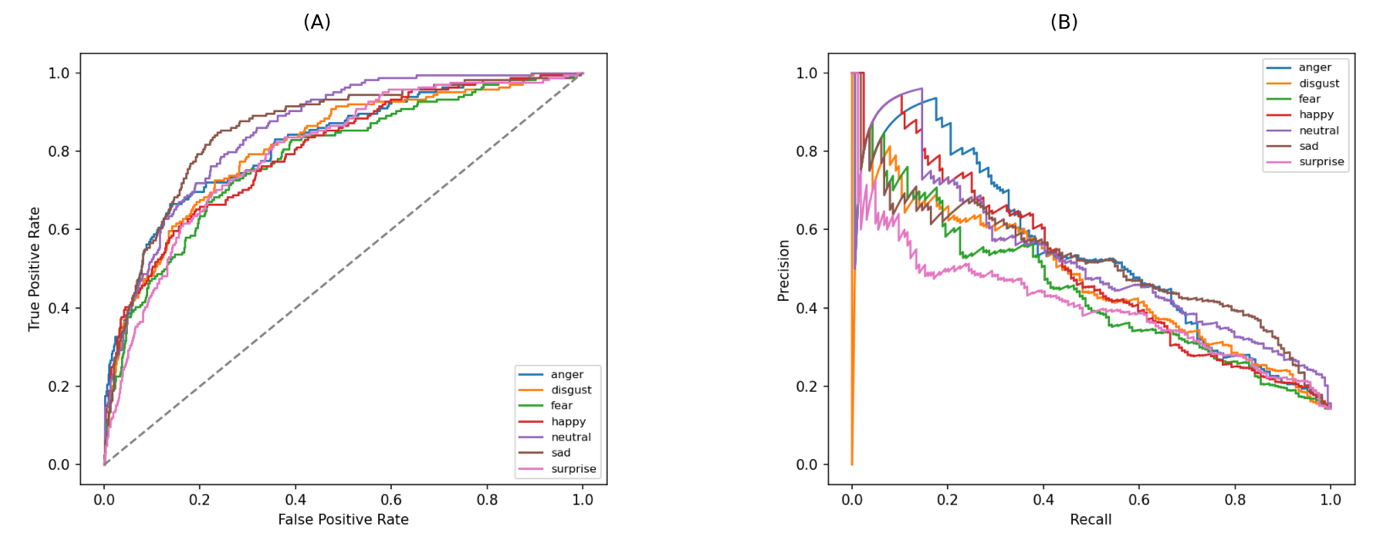


Figure 5. Discrimination of the late fusion model: (A) one-vs-rest ROC by class; (B) one-vs-rest precision–recall by class

Figure 5 (B) shifts the lens to precision–recall. Because each curve treats one class as *positive* against all others, precision falls as recall rises, revealing how quickly false alarms accumulate when we try to recover more positives. Several classes maintain relatively high precision over a wide recall range, which is consistent with the ROC view that their positive instances are well clustered. Others show a steeper decline, telling us that raising recall will require accepting more ambiguous cases. Together, these panels give a threshold-free picture of separability. They also guide the choice of operating points for applications that favour precision or recall, and they complement the macro F1 summaries in Table 1 by revealing where the score is earned along the full trade-off curve rather than at a single threshold.

The first question is how confident the fusion model is when it makes a prediction. Figure 6 (A) places the maximum predicted probability for each test instance into a histogram. Most cases cluster between roughly 0.30 and 0.60 with a taper toward very high confidence. This shape tells us the model rarely declares extreme certainty and spends much of its probability mass in mid-range values, which is typical when classes overlap and sensors provide partially redundant evidence. Most predictions sit in the mid-range, with relatively few at very high confidence. That distribution already hints that the model is cautious rather than over-certain.

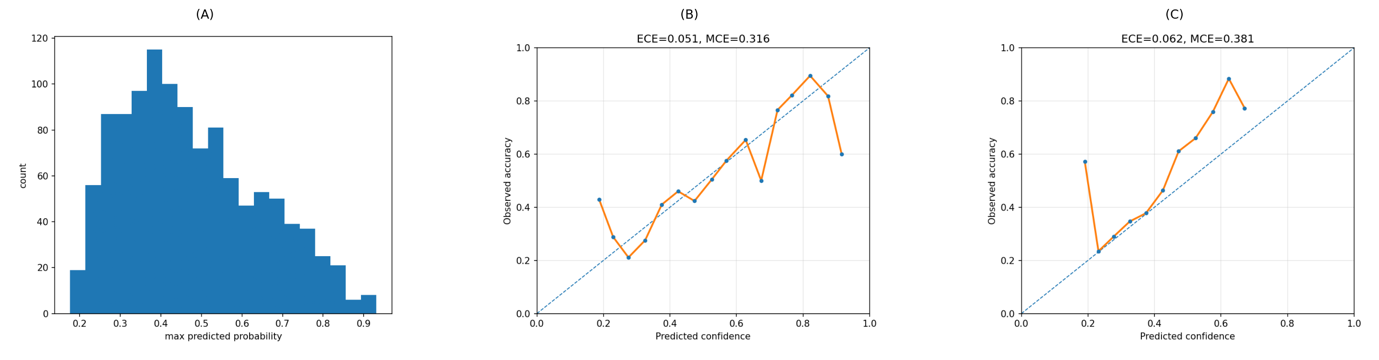


Figure 6. Confidence and reliability of the fusion model: (A) distribution of the maximum predicted probability on the test set; (B) reliability curve before calibration with ECE and MCE; (C) reliability curve after isotonic calibration with ECE and MCE.

The second question is whether those reported probabilities are trustworthy. Figure 6 (B) plots a reliability curve for the uncalibrated fusion model by binning predictions and comparing the bin’s average confidence to the bin’s empirical accuracy. The curve tracks the diagonal reasonably well through the middle of the range, then deviates near the extremes. The accompanying summary on the plot reports an expected calibration error close to 0.051 and a maximum calibration error around 0.316, indicating small but noticeable miscalibration in some bins.

Figure 6 (C) repeats the reliability analysis after isotonic calibration. The curve straightens in the mid-probability regime while the extremes remain data-limited, so the maximum error can fluctuate. In our run, the expected calibration error is about 0.062, and the maximum calibration error is about 0.381. Together with Figure 6 (A), these plots suggest that calibrating probabilities can refine trust in the score scale without materially altering class ranking or the overall discrimination reported earlier. In practical terms, the fusion system’s probabilities are already usable, and post-hoc calibration mainly refines the middle of the scale without altering which examples rank above which.

Table 2 places these findings next to those from the all-features gradient-boosting model. Across both models, the calibrated variants keep accuracy and macro-F1 essentially unchanged while adjusting the probability quality measures. The fusion model maintains very low expected miscalibration, suggesting its scores can be interpreted directly as chances of being correct for most of the range. The larger maximum error reminds us to be cautious with the most confident outputs, where bins can be small and noisy. The gradient-boosting model follows the same pattern, reinforcing the view that probability calibration is a complementary step that improves interpretability rather than raw discrimination.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Calibrated** | **ECE** | **MCE** | **Bins** |
| **Fusion** | FALSE | 0.051 | 0.316 | 16 |
| **HGB** | FALSE | 0.212 | 0.314 | 15 |
| **Fusion** | TRUE | 0.062 | 0.381 | 11 |
| **HGB** | TRUE | 0.127 | 0.223 | 15 |

Table 2. Calibration summary for late-fusion stacking and all-features gradient boosting, before and after isotonic calibration

Figure 7 (A) traces how the late-fusion model improves as we give it more training trials. The cross-validation curve climbs steadily from roughly the low twenties to the mid-forties in accuracy as the sample count grows, while the training curve sits higher and slowly declines toward sixty percent. This pattern is typical for a model whose capacity is fixed while data volume increases. The widening data supply reduces variance and raises out-of-fold performance, yet there remains a stable gap between train and validation that signals headroom either for more expressive fusion or for stronger regularisation tuned to the larger regime. The key message is that additional multimodal data continues to improve accuracy and that fusion benefits predictably from scale.

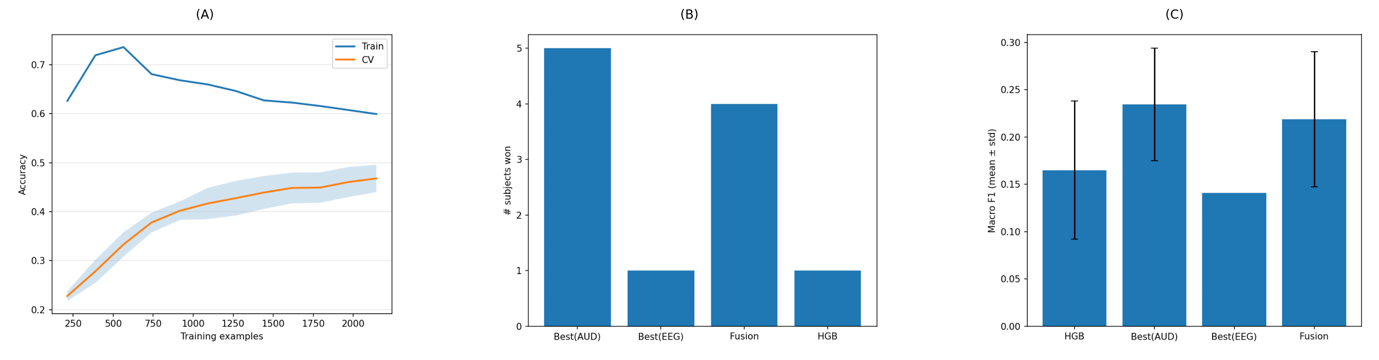


Figure 7. Learning and cross-subject generalisation: (A) learning curve for the late-fusion model with accuracy on the training set and in cross-validation as the number of training trials increases; (B) LOSO winners by subject counted across approaches; (C) LOSO macro F1 averaged over subjects with standard deviation bars for each approach

Figure 7 (B) turns to person-level robustness by counting, subject by subject, which approach achieves the best macro F1 when each individual is held out during training. No single strategy dominates. A simple audio baseline often tops the chart, and fusion is a close second with several wins of its own. EEG alone rarely wins, and the all-features gradient boosting baseline only occasionally does. This distribution hints at real heterogeneity across participants. For some speakers, voice carries most of the discriminative signal. For others, the combination of streams gives the model complementary cues that a single modality misses.

Figure 7 (C) summarises those LOSO results more systematically by averaging macro F1 across people and adding standard deviation bars. Audio has the highest mean with modest spread, which is consistent with its many wins in Figure 7 (B) and suggests a relatively stable unimodal signal. Fusion comes next with a slightly lower mean but a wider error bar. Those wider spread matches the intuition from Figure 7 (B). Fusion helps some subjects a lot and others less, which is what we expect when different individuals lean on different channels. The gradient boosting baseline lags and shows high variability, which indicates that stacking calibrated per-modality models is the more reliable way to integrate streams in this dataset.

Taken together, these learning and LOSO views convey two practical lessons: (1) multimodal fusion keeps improving with more data, so future collections that expand the number of trials should translate directly into better accuracy; (2) across-person generalisation is uneven, which argues for techniques that adapt the fusion weights or calibrate features at the subject level when deployment allows it.

Figure 8 summarises how the all-features gradient-boosting model distributes weight across streams and variables. Figure 8 (A) aggregates permutation importance by modality. Almost all positive contributions come from audio, with a total importance of about 0.114. EEG sits near zero, and EMG is slightly negative at roughly −0.150. Negative values can arise when a feature block is weak or redundant, so that randomising it occasionally helps the model on the test set. The takeaway is that in this tabular representation, the classifier leans primarily on acoustic cues while the physiological channels add little to this particular model.

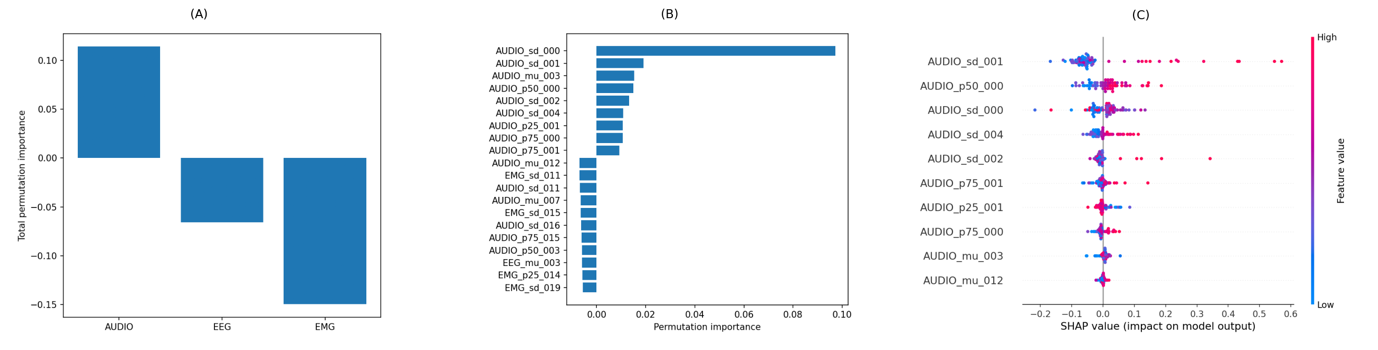


Figure 8. What drives decisions: (A) modality-level permutation importance for the all-features gradient-boosting model; (B) top twenty variables ranked by permutation importance; (C) SHAP summary for the top audio variables showing the direction and spread of their impact on one class score

Figure 8 (B) ranks individual variables. The list is dominated by summary statistics from the audio stream. The single most influential variable is *AUDIO\_sd\_000* with importance near 0.097, which is far ahead of the next features, such as *AUDIO\_sd\_001* around 0.019, and the audio median *AUDIO\_p50\_000* near 0.015. A few EMG dispersion terms and one EEG mean appear, but their effects are an order of magnitude smaller. Because these variables are time-aggregate descriptors of the MFCC trajectories, the pattern suggests that temporal variability in a small set of spectral bands is the primary driver of separation among the seven emotions in PME4.

Figure 8 (C) uses SHAP to check directionality on the top audio predictors for one representative class. For several variables, the red points have high feature values clustered on one side of the vertical zero line, while the blue points have low values clustered on the other. That consistent separation indicates monotonic influence rather than incidental correlation. In practical terms, higher short-term spectral variability for the leading MFCCs tends to push the score toward the class, and lower variability tends to pull it away, while the different percentiles, medians and quartiles modulate the effect.

These results align with earlier performance findings. Audio alone is a strong unimodal baseline, and much of the predictive mass in the all-features model resides there. At the same time, late fusion still outperforms the best single stream on average, which implies that other modalities provide complementary information once they are modelled separately and combined at the score level rather than mixed in one tabular learner. This contrast points to a concrete avenue for future work on PME4. Richer feature learning for EEG and EMG or models that capture cross-modal interactions, could raise the non-audio contributions without sacrificing the strong acoustic signal.

This case study turned the chapter’s ideas into a complete multimodal workflow on PME4, moving from a unified trial table to unimodal baselines, late fusion, discrimination analyses, calibration, and cross-subject tests. A good reading path starts broad and then narrows. Figure 3 establishes that the labels are balanced and shows a small but consistent improvement when scores from different streams are fused. Figure 4 explains where that gain comes from by contrasting the confusion structure with the overlap of errors across strong single streams. Figure 5 then looks at class-wise separability across thresholds, so the reader is not tied to one decision point. Figure 6 focuses on probability quality. Confidence is compared with observed accuracy, and the calibrated variant nudges probabilities into better agreement without changing headline accuracy. Figure 7 addresses generalisation across people and shows that more data are likely to help. Figure 8 ends with model evidence about what drives decisions, with acoustic features contributing most in this tabular representation and other streams adding complementary signals that fusion can exploit.

Three conclusions follow: (1) embodied emotion benefits from combining sensors. Even when one stream is strongest, complementary errors mean that a careful fusion does better than any single source; (2) the quality of probabilities matters for scientific use and for downstream systems. Reporting AUC and average precision together with Brier score, log loss, and calibration errors gives a truer picture than accuracy alone; (3) across-person robustness is the key challenge. LOSO remains an essential test and should guide future data collection and modelling choices.

The pipeline shown here is intentionally simple so that readers who are not specialists can reproduce it. Start from a clean trial table, compare fair baselines for each stream, fuse only at the score level, and always check both what improves and why. This provides a practical template for building trustworthy multimodal models of embodied emotion.

**References:**

Ahmed, N., Al Aghbari, Z., & Girija, S. (2023). A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, *17*, 200171.

Al-Azani, S., & El-Alfy, E. S. M. (2025). A review and critical analysis of multimodal datasets for emotional AI. *Artificial Intelligence Review*, *58*(10), 334.

Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, *41*(2), 423-443.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1-3.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*(4), 335-359.

Cai, Y., Li, X., & Li, J. (2023). Emotion recognition using different sensors, emotion models, methods and datasets: A comprehensive review. *Sensors*, *23*(5), 2455.

Canal, F. Z., Müller, T. R., Matias, J. C., Scotton, G. G., de Sa Junior, A. R., Pozzebon, E., & Sobieranski, A. C. (2022). A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, *582*, 593-617.

Celeghin, A., Diano, M., Bagnis, A., Viola, M., & Tamietto, M. (2017). Basic emotions in human neuroscience: neuroimaging and beyond. *Frontiers in psychology*, *8*, 1432.

Chandraumakantham, O., Gowtham, N., Zakariah, M., & Almazyad, A. (2024). Multimodal emotion recognition using feature fusion: an LLM-based approach. *IEEE Access*, *12*, 108052-108071.

Chen, J., Ro, T., & Zhu, Z. (2022b). Emotion recognition with audio, video, EEG, and EMG: A dataset and baseline approaches. IEEE Access, 10, 13229–13242.

Chen, J., Zhu, Z., & Ro, T. (2022a). PME4: Emotion recognition with audio, video, EEG, and EMG (Version 1) [Data set]. figshare.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014, January). Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning* (pp. 647-655). PMLR.

Du, Z., Wu, S., Huang, D., Li, W., & Wang, Y. (2019). Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition. *IEEE Transactions on Affective Computing*, *12*(3), 565-578.

Egger, M., Ley, M., & Hanke, S. (2019). Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science*, *343*, 35-55.

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.

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, *91*, 424-444.

García-Hernández, R. A., Luna-García, H., Celaya-Padilla, J. M., García-Hernández, A., Reveles-Gómez, L. C., Flores-Chaires, L. A., ... & Villalba-Condori, K. O. (2024). A systematic literature review of modalities, trends, and limitations in emotion recognition, affective computing, and sentiment analysis. *Applied Sciences*, *14*(16), 7165.

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.

Harrison, N. A., Kreibig, S. D., & Critchley, H. D. (2013). A two-way road. *The Cambridge handbook of human affective neuroscience*, 82-106.

Hosseini, S., Yamaghani, M. R., & Poorzaker Arabani, S. (2024). A review of the methods of recognition multimodal emotions in sound, image and text. *International Journal of Applied Operational Research-An Open Access Journal*, *12*(1), 29-41.

Jiang, Y., Li, W., Hossain, M. S., Chen, M., Alelaiwi, A., & Al-Hammadi, M. (2020). A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Information Fusion*, *53*, 209-221.

Juslin, P. N., Laukka, P., & Bänziger, T. (2018). The mirror to our soul? Comparisons of spontaneous and posed vocal expression of emotion. *Journal of nonverbal behavior*, *42*(1), 1-40.

Kakuba, S., Poulose, A., & Han, D. S. (2023). Deep learning approaches for bimodal speech emotion recognition: Advancements, challenges, and a multi-learning model. *IEEE Access*, *11*, 113769-113789.

Katirai, A. (2024). Ethical considerations in emotion recognition technologies: a review of the literature. *AI and Ethics*, *4*(4), 927-948.

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

Koromilas, P., & Giannakopoulos, T. (2021). Deep multimodal emotion recognition on human speech: A review. *Applied Sciences*, *11*(17), 7962.

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. (2025a). 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.

Kragel, P. A., & LaBar, K. S. (2016). Decoding the nature of emotion in the brain. *Trends in cognitive sciences*, *20*(6), 444-455.

Le, H. D., Lee, G. S., Kim, S. H., Kim, S., & Yang, H. J. (2023). Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. *Ieee Access*, *11*, 14742-14751.

Leong, S. C., Tang, Y. M., Lai, C. H., & Lee, C. K. M. (2023). Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing. *Computer science review*, *48*, 100545.

Li, X. (2025, March). Enhancing Video Based Emotion Recognition with Multi-Head Attention and Modality Dropout. In *CONF-MLA 2024: Proceedings of the 2nd International Conference on Machine Learning and Automation, CONF-MLA 2024, November 21, 2024, Adana, Turkey* (p. 133). European Alliance for Innovation.

Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023a). A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, *25*(10), 1440.

Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023b). A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, *25*(10), 1440.

Liu, W., Qiu, J. L., Zheng, W. L., & Lu, B. L. (2021). Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(2), 715-729.

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and emotion*, *23*(2), 209-237.

Mesquita, B., Boiger, M., & De Leersnyder, J. (2017). Doing emotions: The role of culture in everyday emotions. *European Review of Social Psychology*, *28*(1), 95-133.

Miranda-Correa, J. A., Abadi, M. K., Sebe, N., & Patras, I. (2018). Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE transactions on affective computing*, *12*(2), 479-493.

Montero Quispe, K. G., Utyiama, D. M., Dos Santos, E. M., Oliveira, H. A., & Souto, E. J. (2022). Applying self-supervised representation learning for emotion recognition using physiological signals. *Sensors*, *22*(23), 9102.

Pan, B., Hirota, K., Jia, Z., & Dai, Y. (2023). A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing*, *561*, 126866.

Pillalamarri, R., & Shanmugam, U. (2025). A review on EEG-based multimodal learning for emotion recognition. *Artificial Intelligence Review*, *58*(5), 131.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, *10*(3), 61-74.

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

Rouast, P. V., Adam, M. T., & Chiong, R. (2019). Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing*, *12*(2), 524-543.

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

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

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

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning* (pp. 10347-10357). PMLR.

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).

Udahemuka, G., Djouani, K., & Kurien, A. M. (2024). Multimodal Emotion Recognition using visual, vocal and Physiological Signals: a review. *Applied Sciences*, *14*(17), 8071.

Weng, H. Y., Feldman, J. L., Leggio, L., Napadow, V., Park, J., & Price, C. J. (2021). Interventions and manipulations of interoception. *Trends in neurosciences*, *44*(1), 52-62.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, *5*(2), 241-259.

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

Yoon, S., & Kim, B. (2025). Multi-Scale Temporal Fusion Network for Real-Time Multimodal Emotion Recognition in IoT Environments. *Sensors*, *25*(16), 5066.

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*.

Zadrozny, B., & Elkan, C. (2002, July). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 694-699).

Zhang, S., Yang, Y., Chen, C., Zhang, X., Leng, Q., & Zhao, X. (2024). Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications*, *237*, 121692.

Zhao, F., Zhang, C., & Geng, B. (2024). Deep multimodal data fusion. *ACM computing surveys*, *56*(9), 1-36.

Zhao, K., Zheng, M., Li, Q., & Liu, J. (2025). Multimodal Sentiment Analysis-A Comprehensive Survey From a Fusion Methods Perspective. *IEEE Access*.