# Title: Multimodal Cognitive State Recognition during Problem-Solving in Mixed Reality OSHA Training Using Physiological Signals

# Physiological Signal-Based Classification of Cognitive and Emotional States in Problem Solving (2020–2025)

Recent studies have advanced the use of EEG and peripheral physiology to detect mental states during problem-solving. Multimodal approaches (combining EEG, eye metrics, heart rate, and skin conductance) have achieved very high accuracy in controlled settings.  For example, one review notes that combining EEG with other signals has produced classification accuracies exceeding 90% .  Key findings include that pupil dilation predicts insight (“Aha!”) about 500 ms before solution , while deep “flow” engagement shows elevated EEG alpha/theta power and distinctive heart-rate variability .  A large wearable-sensor dataset (EEG, accelerometer, EDA, PPG) recorded office-style tasks (mental arithmetic, Stroop, n-back, etc.) to enable real-time workload/stress estimation .  These works demonstrate robust links between physiology and cognitive/emotional load, but also highlight gaps: most models are offline and lab-based, labeling schemes vary, and real-world generalization remains limited .

* **Cognitive and emotional states studied.**  Studies have focused on workload, attention, and affective states like stress or engagement.  For example, working-memory load (Sternberg task) correlates with larger pupil dilation and heart-rate changes , and distracted or low-attention states exhibit increased EEG alpha (8–13 Hz) power .  Insight (“Aha!”) solutions are marked by a sudden pupil dilation peak .  The “flow” state (optimal engagement) vs boredom has been probed using EEG, ECG, GSR and motion: flow yields dominant alpha/theta EEG and U-shaped HRV patterns , whereas boredom shows higher physical activity.  GSR/EDA reliably increases with cognitive load or stress .  **Table 1** summarizes representative studies correlating signals with states. (EEG frequency bands, ocular metrics, HRV, etc., reliably index attention, load and emotion .)
* **Gaps and emergent states.**  Some relevant states have been rarely targeted.  Classic problem-solving phases like *impasse* (feeling stuck) or *confusion* are seldom directly measured by sensors.  Similarly, emotions beyond arousal (e.g. frustration, enjoyment) are underexplored in physiological studies.  A recent framework attempted to label “systemic” states such as cognitive load, mind wandering, distraction and sense of urgency from multimodal data , but achieved only modest accuracy, highlighting that such states are still open research challenges.  In general, states like boredom and mind-wandering get little coverage (boredom was noted only as “anti-flow” activity ), suggesting opportunities for future work.

## Correlation of Signals with Cognitive/Emotional States

* **EEG features.**  EEG band-power and connectivity have been extensively linked to cognitive states.  For example, increased **theta (4–8 Hz)** and **delta (0.5–4 Hz)** power often accompany rising working memory load, while higher **alpha (8–13 Hz)** generally indicates relaxed or idling states .  Frontal alpha asymmetry has been used to infer affect (more left vs right activity for positive vs negative moods) .  EEG coherence/connectivity metrics (e.g. phase-locking value, coherence graphs) can capture functional network changes under load or emotion .  In one study, graph-based EEG connectivity (phase-locking value) was used to classify workload; the best connectivity feature (PLV) yielded about 65% accuracy cross-task .
* **Eye-tracking & pupillometry.**  Oculomotor measures reflect attention and load.  Fixation duration and saccade amplitude change with task difficulty .  Pupil diameter is a well-established load indicator: it dilates with higher task difficulty or memory load .  Notably, pupil size surges ~500 ms before “Aha!” insight solutions .  Advanced pupil features like the Low-High Index of Pupillary Activity (LHIPA) quantify oscillatory power in low vs high-frequency pupil dynamics and correlate with cognitive effort .
* **Cardiovascular and EDA (Empatica) metrics.**  Wearable sensors like Empatica E4 provide accelerometry (ACC), blood-volume pulse (BVP/PPG), skin conductance (EDA/GSR) and skin temperature.  **Heart rate and HRV (from PPG)** reliably reflect workload: higher heart rate and lower HRV accompany higher load .  For instance, combining pupillometry and heart-rate features distinguished low vs high difficulty .  **Electrodermal activity (EDA/GSR)** rises with arousal/stress; studies find that GSR features (phasic SCR peaks, tonic SCL level) can predict workload changes with ~75% accuracy .  A wrist-worn study using Empatica E4 extracted GSR and PPG features to classify stress vs rest, achieving ~76.5% accuracy .  **Motion (ACC)** can signal gross user activity or fidgeting; for example, lower motion distinguished flow from boredom .  Accelerometer data also help remove movement artifacts from other signals.  **Skin temperature (TEMP)** tends to decrease under stress (vasoconstriction) though it is less commonly used alone for emotion.

## Data Processing by Signal Type

* **EEG preprocessing:** Raw EEG is typically **band-pass filtered** (e.g. 0.5–50 Hz) and notch-filtered at 50/60 Hz to remove drift and line noise.  Artifacts (blinks, muscle noise) are often removed using **Independent Component Analysis (ICA)** or regression of EOG channels.  Data may be re-referenced and epoched around task events.  (No single citation here, as this is standard practice.)
* **Pupillometry:** Raw pupil size data contain many **invalid samples** (blinks, reflection errors).  A typical pipeline (Kret 2019) first identifies invalid points (dropouts during blinks) and removes them .  The remaining valid samples are interpolated (spline or linear) and **upsampled** to regularize the data .  The interpolated pupil signal is then **low-pass filtered** to remove high-frequency noise, and optionally baseline-corrected against a pre-stimulus period .  Throughout, careful interpolation and smoothing is crucial to avoid contaminating subtle pupil dynamics .
* **GSR/EDA:** EDA signals are often **decomposed** into tonic and phasic components via methods like Continuous Decomposition Analysis (Ledalab/CDA) .  CDA (or related nonnegative deconvolution) separates the slow **Skin Conductance Level (SCL)** from rapid **Skin Conductance Responses (SCR)**.  From this, one extracts features such as SCR peak amplitudes (reflecting phasic arousal) and mean tonic level .  In practice, EDA is also low-pass filtered (<5 Hz) to reduce high-frequency noise before decomposition.
* **Photoplethysmogram (BVP):** PPG (blood volume pulse) is bandpass-filtered (e.g. 0.5–4 Hz) to capture the cardiac pulse waveform.  Then **peak detection** yields inter-beat intervals (IBI), from which heart rate (HR) and heart-rate variability (HRV) features (RMSSD, LF/HF ratio) are computed.  PPG may be artifact-rejected based on motion or signal quality.
* **Accelerometry (ACC):** ACC is often converted to a **resultant magnitude** (√(x²+y²+z²)) and low-pass filtered to quantify movement energy.  High-frequency components indicate abrupt motions, while trends can index posture or restfulness.  ACC data also serve to **derivative-correct** other signals (e.g. remove PPG motion artifacts).
* **Skin temperature (TEMP):** TEMP is typically smoothed or low-pass filtered, and relative changes over time are noted. Rapid drops in TEMP can signal acute stress but are subtle over short tasks.

## Feature Extraction for Time-Series Data

Typical feature sets from these signals include:

* **Time-domain statistics:** e.g. mean, median, standard deviation, skewness, kurtosis of the signal (or z-scored signal) .  For example, one EDA study lists mean, median, std and percentiles of z-scored EDA .  For EEG or ACC, one might use overall mean power or movement energy.  Event-related features include pupil **peak dilation** and **latency**, SCR **peak counts**/amplitudes, or heart-rate peaks.
* **Frequency-domain features:** For EEG, band-power in delta/theta/alpha/beta/gamma is fundamental .  For HRV, power in low-frequency (LF) and high-frequency (HF) bands are used.  ACC or gyroscope data may be analyzed for spectral peaks.  Pupil/EDA generally yield better features in time-domain than spectral.
* **Time-frequency / wavelets:** Wavelet transforms extract transient features.  Notably, the Low-High Index of Pupillary Activity (LHIPA) uses wavelet decomposition: it computes the ratio of low-frequency to high-frequency components of the pupil signal .  LHIPA has been shown to inversely correlate with task difficulty .  Similarly, EEG can be decomposed via wavelets to capture nonstationary dynamics.
* **Connectivity and spatial features:** EEG **functional connectivity** (e.g. phase-locking value, coherence) between channel pairs is often used.  Graph-based features (e.g. node degrees of EEG coherence networks) have been exploited .  Frontal **asymmetry** (difference in power between left vs right prefrontal electrodes) is another common EEG feature for emotion .
* **Oculomotor features:** In addition to pupillary metrics, eye movement features are used: **fixation durations**, **saccade amplitudes**, and **blink rate** can all reflect cognitive load.  One study found fixation duration and saccade amplitude had the highest LDA coefficients for discriminating task load .
* **Derived features (HR/HRV, EDA):**  From HR time series one can extract RMSSD (time-domain HRV), LF/HF ratio (frequency-domain HRV), or nonlinear measures (entropy of beat-to-beat intervals).  From EDA, extracted phasic features include SCR **amplitude sum** and SCL mean .

These features are typically computed for each task window or trial and then fed into classifiers.

## Classification Models

Modern approaches span classical machine learning to deep neural nets:

* **Classical ML:** Early works used Support Vector Machines, Random Forests, and LDA on handcrafted features.  For example, a Random Forest on Empatica E4 features (GSR & PPG) classified stress vs rest at ~76% accuracy .  Similarly, an LDA on oculometric and biometric features achieved ~65% accuracy across four task conditions .  These models rely on explicit feature engineering and work well for smaller datasets.
* **Deep learning (CNN, RNN, hybrids):** Deep networks can learn features directly from raw or minimally processed signals.  Convolutional Neural Networks (CNNs) have been applied to EEG time-series or spectrograms.  One study used a CNN on gamma-band EEG signals to classify “high” vs “low” cognitive states, achieving ~91% accuracy, outperforming SVM and RF baselines .  Recurrent networks (LSTM/GRU) capture temporal dynamics; hybrid CNN–LSTM architectures have also been used.  For instance, a CNN–LSTM (with ResNet backbone) applied to EEG for emotion classification (PTSD context) reported ~98% accuracy , significantly higher than earlier methods.  These deep models typically require more data and computation but can extract complex patterns.
* **Transformer-based models:** Recently, Transformer architectures (with self-attention) have been introduced for physiological time series.  They excel at modeling long-range dependencies.  A hybrid CNN-Transformer model for EEG-based mental workload classification achieved ~85–90% accuracy (85.5% average on 5-fold CV) , outperforming previous CNN-only models.  Reviews note rapid growth in transformer-based EEG studies (e.g. time-series transformers, vision transformers on EEG) .  Graph Attention Networks (GAT) have also been applied to EEG, forming models like Stacked Graph Attention CNNs, though their reported accuracy (e.g. ~65% cross-subject) is still modest .
* **Hybrid and ensemble models:** Many studies combine multiple models or modalities.  Ensembles of CNNs with attention, or feature fusion from EEG, GSR, and eye data, can boost performance.  For example, one cognitive load study fused multi-band EEG features via a 3D CNN with weighting factors to improve CNN performance .  However, interpretability of deep ensembles remains a challenge.
* **Large Language Models (LLMs):** To date, LLMs (e.g. GPT, BERT) are designed for text and have not been directly applied to raw biosignals.  There is no standard method for feeding multi-channel time-series into an LLM.  Instead, research focuses on signal-adapted architectures (CNN/RNN/Transformer) rather than text-based LLMs.  Some vision transformers have been used by treating EEG maps as images, but pure language models are not used for physiological data.

## Model Performance Comparison

Reported accuracies vary widely depending on task and data:

* Pupillometry classifiers have achieved high performance for simple tasks.  For instance, a random forest using just pupil features classified easy vs hard levels in an educational game with **87.5% accuracy** .  (By contrast, using only game-logging metrics gave 75%.)
* CNNs on EEG often exceed 90% in constrained settings.  As noted, a CNN on gamma-band EEG reached ~91% , and a complex CNN–LSTM architecture reported ~98% .
* In multimodal scenarios, performance can be even higher.  The hybrid CNN-Transformer model achieved up to 90.9% best-case accuracy on a public EEG workload dataset .  Reviews of EEG-emotion classification similarly report >90% accuracy when fusing EEG with other signals .
* However, cross-task or cross-subject generalization is harder.  Graph-based models on multiple tasks only reached ~65% .  Simple models (e.g. LDA on eye/HRV) yielded ~65%–70% when discriminating multiple classes.
* Overall, **deep models consistently outperform shallow ones** on large datasets .  The trade-off is computational cost and need for data.  Table 2 (below) summarizes select studies and their reported accuracies.

**Table 1.** *Examples of physiological features correlated with cognitive/emotional states.* (Summary of key signal-state findings.)

| **Signal** | **Example Features** | **Related Cognitive/Emotional State** | **References** |
| --- | --- | --- | --- |
| EEG | Band power (θ, α, β, etc.), connectivity (PLV/coherence), asymmetry | Working memory load (↑θ,↑δ), decreased attention (↑α), emotional valence (frontal asymmetry) |  |
| Pupil (dilation) | Mean diameter change, peak dilation, LHIPA (low/high-frequency ratio) | Cognitive load/difficulty (larger dilation for harder tasks), insight (“Aha!”) |  |
| Fixations/Saccades | Average fixation duration, saccade amplitude | Task engagement and difficulty (longer fixations for high load) |  |
| Heart (PPG → HR/HRV) | Mean HR, RMSSD, LF/HF ratio | Task difficulty (↑HR, ↓HRV with higher load); stress level |  |
| GSR (EDA) | Phasic SCR peak count, amplitude; tonic SCL level | Arousal/stress (more SCR peaks, higher SCL under stress) |  |
| ACC (Motion) | Movement magnitude, variance | Cognitive state context (e.g. ↓motion in focused states, ↑ in boredom) |  |

**Table 2.** *Example classification results from recent studies (2020–2025).*

| **Study (Year)** | **Modalities** | **Task/States** | **Model** | **Accuracy** |
| --- | --- | --- | --- | --- |
| Campanella *et al.* (2023) | Empatica (EDA, PPG) | Stress vs. relaxation | Random Forest (χ² feature selection) | 76.5% |
| Mitre-Hernández *et al.* (2021) | Eye-tracking (pupil) | Game difficulty (Easy/Hard) | Random Forest | 87.5% |
| Avital *et al.* (2024) | EEG (gamma band) | High vs. low cognitive state | 1D-CNN (Gamma band input) | 91.4% |
| Wei *et al.* (2025) | EEG connectivity (PLV graph) | Workload (cross-task) | Stacked Graph Attention-CNN | 65.1% |
| Parveen & Bhavsar (2025) | EEG | Mental workload (low/med/high) | Hybrid CNN–Transformer | 90.9% (best, 5-fold CV) |
| *Table notes:* Some studies report best-case accuracy; direct comparisons are difficult due to different tasks and data splits. |  |  |  |  |

## Discussion: Gaps and Future Directions

Despite the above progress, several gaps remain.  **Understudied states:**  Many nuanced cognitive/emotional states in problem solving lack physiological classifiers.  For instance, the impasse state (being stuck on a problem) has no clear sensor signature identified; similarly, complex emotions like frustration or curiosity during problem solving are rarely directly measured.  One recent effort did include *mind wandering*, *distraction*, and *sense of urgency* as labels , but accuracy was only modest, suggesting these states are hard to isolate.  *Flow vs. boredom* was explicitly examined only in one study .  Thus, states such as confusion, frustration, impasse, or even positive affect (interest) are ripe for future work.

**Real-world robustness and personalization:**  Most models assume controlled tasks and well-preprocessed data.  Real environments introduce noise (motion, varying lighting) and individual differences.  The literature notes a need for adaptive, real-time algorithms and standardized protocols .  For example, **labeling variability** (how subjects report emotion or “Aha!”) is a challenge .  Multi-task and transfer learning approaches may help models generalize across tasks.  Additional contextual cues (task type, performance metrics) might be combined with physiology to improve inference .

**Advanced models and multimodality:**  Future work may further integrate modalities.  Multimodal fusion (EEG + eye + heart + motion) already boosts accuracy .  Deep learning models (CNNs, transformers) continue to outperform classical classifiers , but interpretability and data requirements are issues.  Graph and attention-based networks are emerging tools for EEG, as are temporal transformers .  However, no research has leveraged large language models directly on sensor data; such models may be adapted in multi-modal systems but this remains unexplored.

In summary, physiological signal analysis can effectively classify many cognitive and affective states during problem solving . Key results show reliable EEG, ocular, and peripheral correlates of load, attention, insight, and stress.  Yet challenges persist in standardizing methods, expanding target state labels, and improving robustness.  Addressing underexplored states (e.g. impasse, frustration) and refining data pipelines are promising avenues.  Overall, the literature suggests rich potential for continuous, real-time assessment of mental states via wearable sensors, provided that models evolve to handle variability and complexity in real problem-solving contexts.