# Explainable Multimodal Intelligence for Cognitive Fatigue Detection Using Speech Acoustics, Facial Dynamics, and Human Interaction Behaviour: A Systematic Review

***Ms. Sarbjeet Kaur***
*Assistant Professor, Department of Computer Science & Engineering*
*PCTE Institute of Engineering and Technology, Ludhiana, Punjab-142021, India*
*E-mail: sarbjeet.sembhi28@gmail.com*

**How to Cite this Article**:

Kaur, S. (2026). Explainable Multimodal Intelligence for Cognitive Fatigue Detection Using Speech Acoustics, Facial Dynamics, and Human Interaction Behaviour: A Systematic Review. International Journal of Creative and Open Research in Engineering and Management, <i>02</i>(7). https://doi.org/10.55041/ijcope.v2i6.395



https://doi.org/10.55041/ijcope.v2i6.395

## Abstract

Cognitive fatigue is a transient, neurophysiologically grounded impairment of executive function and sustained attention arising from prolonged mental effort without adequate rest. It is a leading contributor to human error in safety-critical domains, including aviation, surgery, and long-haul transportation. Despite two decades of active research, deployed detection systems remain anchored to invasive physiological hardware, architecturally naive fusion strategies, and opaque, unexplained inference processes. These three barriers collectively prevent real-world deployment at any meaningful scale. This review systematically synthesises twenty landmark studies published between 2022 and 2026, drawn from IEEE Xplore, PubMed, Nature group journals, ACM Digital Library, and ScienceDirect. Three non-contact detection modalities are evaluated in depth: speech acoustic-prosodic features, facial Action Unit (AU) dynamics, and human-computer interaction (HCI) behavioural markers. The review also examines emerging multimodal fusion architectures and explainable artificial intelligence (XAI) integration across the literature. Six structural research gaps are formally identified: persistent reliance on invasive sensing hardware; temporally naive fusion that discards asynchronous cross-modal dependencies; the complete absence of HCI behaviour as a recognised primary detection modality; XAI confined to post-hoc unimodal attribution rather than embedded cross-modal fusion; inconsistent ground-truth labelling protocols that conflate cognitive fatigue with drowsiness; and the absence of any open, tri-modal benchmark dataset for community evaluation. Three primary research objectives are derived from this gap analysis, jointly defining the architecture and validation agenda for the proposed Explainable Multimodal Intelligence framework for Cognitive Fatigue Detection (EMI-CFD). This review provides a critical, evidence-grounded roadmap for the next generation of transparent, non-invasive cognitive fatigue monitoring systems.

**Keywords:** cognitive fatigue detection; multimodal fusion; explainable artificial intelligence; speech prosody; facial action units

# 1. INTRODUCTION

Cognitive fatigue is a transient, reversible decline in the efficiency of higher-order neural processing that encompasses working memory, executive control, and sustained attention, and is caused by prolonged mental effort without adequate rest [1]. Its consequences are measurable and serious. Incident investigations in commercial aviation consistently attribute between 15 and 20 percent of crew-related accidents to fatigue-driven performance decrements [2]. In surgical practice, error rates escalate measurably during procedures performed beyond the sixth consecutive hour, and artificial intelligence-based fatigue monitoring for operating theatres is now identified as a clinical priority [3]. On public roads, driver fatigue is implicated in up to one fifth of serious collisions in industrialised nations [2]. Across knowledge-intensive professions including finance, law, medicine, and engineering, cognitive fatigue silently degrades the quality of decisions made by millions of workers every day, and its accumulated societal cost is substantial.

What makes cognitive fatigue particularly difficult to detect automatically is its behavioural subtlety. Unlike physical fatigue, which announces itself through measurable biomechanical deterioration, cognitive fatigue manifests through distributed, temporally evolving changes across multiple observable channels, including the prosodic texture of spontaneous speech, the microkinematic dynamics of facial musculature, and the temporal precision of goal-directed computer interaction [4,5]. This distributed expressivity is precisely why a multimodal observational approach is not merely preferable but scientifically necessary. No single channel captures the full information content of the fatigue state, and the interactions between channels carry diagnostic signal that individual modalities cannot provide alone.

The existing literature has pursued automatic cognitive fatigue detection with considerable energy but persistent structural limitations. Electroencephalography (EEG), heart rate variability (HRV), and galvanic skin response (GSR) deliver high neurophysiological sensitivity but demand contact-based, laboratory-constrained hardware that is wholly incompatible with naturalistic deployment [6]. Camera and microphone-based approaches have developed in disciplinary silos, with facial analysis researchers rarely integrating speech data, and both communities almost entirely neglecting the rich behavioural signal continuously generated by users interacting with digital devices [7,8]. Moreover, as detection architectures have grown in complexity, their opacity has grown correspondingly. The EU Artificial Intelligence Act (2024) [27] now mandates interpretable, auditable, and contestable explanations for automated systems deployed in high-stakes contexts. A fatigue detection system that cannot explain its inference is legally indefensible in the regulatory environments where it is most critically needed.

This review addresses these challenges by critically analysing twenty key studies published between 2022 and 2026. The objective is not merely to catalogue what exists but to identify with precision the structural gaps that persist, characterise their causes, and map the research agenda the field must pursue. The three target modalities are analysed in depth, followed by a synthesis of fusion architectures and XAI approaches, formal gap characterisation, research objective derivation, and a forward-looking discussion.

# 2. BACKGROUND: COGNITIVE FATIGUE AND ITS OBSERVABLE SIGNATURES

## 2.1 Neuropsychological Mechanisms

Cognitive fatigue is neurophysiologically distinguished from drowsiness and mental stress by its specific aetiology, namely sustained cognitively demanding task performance, and its characteristic neural profile. Progressive deactivation of the prefrontal cortex and anterior cingulate cortex regions supporting inhibitory control and performance monitoring occurs alongside compensatory upregulation of subcortical motivational circuits as the individual increasingly struggles to sustain performance [1,20]. EEG studies consistently identify increased frontal theta power in the 4 to 8 Hz band and attenuated alpha and beta power in the 8 to 30

Hz band as reliable neurophysiological markers [20]. These central changes propagate peripherally through anatomically distinct pathways, producing detectable and theoretically predictable alterations in speech production, facial motor control, and voluntary interaction behaviour, with each pathway operating at its own characteristic temporal scale.

### 2.2 Speech Acoustic Signatures

Speech production is among the most neurally demanding ordinary activities, requiring the simultaneous coordination of phonological planning, articulatory motor programming, respiratory management, and prosodic modulation, all of which degrade as cognitive resources deplete [5,28]. The most consistently reported fatigue-induced acoustic changes include progressive reduction in speech rate and articulation rate; increased pause frequency and duration reflecting disrupted phonological planning; flattening of the fundamental frequency (F0) contour as respiratory and laryngeal control deteriorates; increased jitter and shimmer from reduced neuromotor precision; and narrowing of vowel space extent as articulatory targeting becomes less precise [26,28]. Dias et al. [5] demonstrated, in a multiple sclerosis cohort, that automated free-narrative speech analysis yielded a significant negative correlation between cognitive fatigue severity and speech ratio ($rho = 0.28$, $p < 0.001$), establishing that single-microphone spontaneous speech carries clinically detectable fatigue signal even without any task constraint on speech content.

### 2.3 Facial Action Unit Dynamics

The Facial Action Coding System (FACS) [23] provides the principled vocabulary through which fatigue-induced facial changes are systematically quantified via Action Unit (AU) activation patterns. Key fatigue-relevant AUs include AU43 (eyes closed), AU45 (blink), AU4 (brow lowerer), AU6 (cheek raiser), and AU7 (lid tightener), with the PERCLOS metric, representing the proportion of time within a sliding window during which the eyelid occludes more than 80 percent of the pupil, achieving regulatory recognition in transport safety frameworks as a drowsiness indicator. Fatigue-related facial dynamics include progressive ptosis, increased blink duration, reduced spontaneous expressivity across the upper face, and head pose destabilisation with increasing pitch and yaw variance [9,15]. These changes evolve across multiple timescales, with blink dynamics at sub-second resolution and ptosis progression over minutes, necessitating temporal modelling architectures capable of capturing both fast and slow facial dynamics simultaneously.

### 2.4 HCI Behavioural Markers

Human-computer interaction behaviour provides a theoretically compelling and practically accessible fatigue signal that remains almost entirely unexploited in the detection literature. The theoretical basis is well established: cognitive fatigue impairs the executive control functions including working memory updating, response inhibition, and error monitoring, all of which regulate goal-directed computer interaction. Empirical studies have documented consistent fatigue-related HCI degradation including increased and more variable inter-key intervals (IKIs) during typing, elevated error and backspace rates, less direct cursor trajectories with higher path index, extended click response latencies, and a transition from purposive to exploratory scrolling patterns [18]. Crucially, these signals are generated as an invisible byproduct of ordinary computer use, require no additional hardware, and can be captured and processed entirely on-device, properties that make HCI behaviour uniquely suitable for continuous, privacy-compatible fatigue monitoring in occupational settings.

## 3. REVIEW METHODOLOGY

An adapted PRISMA protocol was applied to structured searches across IEEE Xplore, PubMed or MEDLINE, Scopus, ACM Digital Library, ScienceDirect, and arXiv for publications from 2020 to 2026. Primary search strings were: (i) cognitive fatigue detection AND multimodal; (ii) speech OR facial dynamics OR action units OR HCI behaviour AND fatigue AND deep learning OR transformer; (iii) explainable AI OR XAI AND fatigue OR cognitive load; and (iv) cross-modal attention AND cognitive state. Twenty papers were selected to represent the full methodological landscape spanning invasive physiological baselines, camera-based systems, speech-based approaches, transformer fusion architectures, HCI behavioural studies, XAI integration studies, and domain-specific systematic reviews. Each paper was assessed across six dimensions: modality coverage, fusion architecture, XAI treatment, ground-truth labelling protocol, ecological validity, and generalisability. Table 1 presents the complete structured summary of all reviewed papers.

*Table 1. Summary of the reviewed studies: modalities, methods, performance metrics, and key limitations. Acc = accuracy; AUC = area under the ROC curve; MCI = mild cognitive impairment; ATC = air traffic controllers; IKI = inter-key interval.*

| Author (Year | Year | Modalities | Method | Performance | Key Limitation |
|---|---|---|---|---|---|
| Virk et al. [4] | 2023 | Visual, Thermal Keystroke, Voice | Feature concat SVM | Acc. 92.3% | Naive early fusion; no XAI lab only; no tempora modelling |
| Dias et al. [5] | 2024 | Speech (free narrative) | Acoustic analysis, ML | rho = 0.28 p<0.001 | Unimodal; clinical MS cohor only; no facial or HCI data |
| Karim et al. [6 | 2024 | EEG, HRV, GSR (wearable) | Wearable multi sensor DL review | Survey (no metric) | Fully invasive; no camera o speech; lacks ecologica validity |
| Far Poor et al [7] | 2024 | Speech, Language, Video | Cross-transformer co attention | AUC 85.3% | MCI detection, not fatigue; no HCI modality; no XAI |
| Zhao et al. [8] | 2025 | EEG, ECG Facial AUs | Multi-query cross-modal attention | Acc. 87.01% AUC 86.88% | Requires EEG and ECG; no HCI or speech; no explainability |
| Xie and Guo [9] | 2025 | Facial video (non-contact) | Mamba SSM multi-task (CogMamba) | SOTA in automotive domain | Automotive only; no speech or HCI; no XAI |
| Sun et al. [10] | 2025 | Behavioural and physiological | Noncontact multimodal fusion | IEEE Sensors benchmark | Driver distraction, no cognitive fatigue; no speech or facial AUs |
| FatigueNet [11] | 2025 | EEG and facial video | GNN and Transformer hybrid | Near-SOTA real-time | EEG required; no speech; no XAI; drowsiness conflated with fatigue |
| FatigueSense [12] | 2025 | 9 wearable device sensors | Time-window ML, multi-device | Multi exceeds single device | Fully intrusive; no camera speech, or HCI data |
| Chen et al. [13 | 2026 | EEG and ECG (pilot cohort) | ANOVA-SVM and XGBoost | Acc. 88.42% cross-subject | Aviation-specific; invasive sensors; no non-contac pathway |

| | | | | | |
|---|---|---|---|---|---|
| Arif et al. [14] | 2022 | Facial feature and voice | CNN stacking meta-learner | High Acc (N=14 ATC) | Late fusion; tiny sample; no temporal modelling; no XAI |
| MG-YOLOv8 [15] | 2025 | Facial vide (EAR, PERCLOS, MAR) | YOLOv8 and PFLD with EWMA | Real-time driver fatigue | Automotive and drowsines only; no speech or HCI; no XAI |
| XAI-Speech Review [16] | 2025 | Speech only (XAI review) | PRISMA: SHAP LIME, attention | AUC 0.76 to 0.94 (dementia | Dementia, not fatigue; no multimodal fusion-level XAI |
| Noor et al. [17 | 2025 | Multiple domains (XA review) | PRISMA review across 89 studies | Systematic review | No fatigue-specific XAI; no cross-modal attribution evaluated |
| Keystroke Study [18] | 2024 | Keystroke dynamics (HCI) | ML on IKI, hold time, pauses | Significant fatigue signal | Unimodal HCI only; no speech or facial; lab-based |
| Bridging Tex and Speech [19] | 2025 | Speech and tex (bimodal) | RoBERTa and WavLM with Integrated Gradients | Acc. 0.8 (emotion, 5 class) | Emotion recognition, no fatigue; bimodal only; no facial or HCI |
| EEG Fatigue Review [20] | 2025 | EEG (18 studies reviewed) | PRISMA frequency-band analysis | Theta up, beta down consistent | Invasive; no non-contact; no HCI, speech, or facial fusion |
| HCSR Review [21] | 2025 | Physiological (405 studies) | PRISMA ML and DL review | DL dominan from 2023 | No non-contact tri-moda system identified; construct conflated |
| Neuro-Symbolic [22] | 2025 | Eye-tracking and fNIRS | Neuro-symbolic fuzzy reasoning | Interpretable fatigue detection | Invasive; no speech or facia dynamics; limited XAI scope |

## 4. THEMATIC ANALYSIS

### 4.1 Speech-Based Cognitive Fatigue Detection

Speech-based approaches to cognitive fatigue detection have pursued two parallel feature engineering philosophies. The first employs hand-crafted acoustic-prosodic features grounded in phonetic theory, most prominently the Geneva Minimalistic Acoustic Parameter Set (GeMAPS), which standardises the 88 most fatigue-relevant acoustic dimensions including F0 statistics, energy, spectral balance, jitter, shimmer, and cepstral peak prominence, enabling cross-study comparability that earlier ad hoc feature sets precluded [26]. The second employs end-to-end self-supervised representation learning through architectures such as wav2vec 2.0 [24], which extracts contextual embeddings at sub-phonemic resolution capturing temporal dynamics inaccessible to frame-level statistical features. For cognitive fatigue, this distinction matters because some of the most diagnostically significant changes, including subtle wavering of phonemic precision and micro-fluctuations in speech timing, may fall below the effective resolution of GeMAPS analysis windows.

A critical methodological divergence in the literature concerns temporal granularity. The most common approach assigns a single fatigue rating to an entire session of several minutes, then extracts statistical summaries over that window, treating the session as a static unit and discarding within-session temporal evolution. This implicit stationarity assumption is physiologically untenable because cognitive fatigue fluctuates sub-minutely as task demands vary and compensatory strategies are deployed [1]. Studies that have adopted finer temporal granularity through sliding windows of 30 to 60 seconds fed into recurrent or attention-based temporal models consistently report higher detection accuracy and more clinically interpretable outputs. The bimodal fusion system of Arif et al. [14], which combines MFCC voice features with facial eye-closure metrics for air traffic controllers, demonstrates the complementary value of pairing speech with facial signals, but its naive stacking ensemble treats modalities as independent sources of evidence rather than as components of a dynamically interacting system, preventing the model from capturing the temporal co-evolution of speech and facial fatigue signatures.

### 4.2 Facial Dynamics: From Geometric Features to Transformer Architectures

Facial fatigue detection has evolved through three distinct methodological generations. First-generation geometric approaches relied on EAR, MAR, and PERCLOS computed from facial landmark coordinates, achieving regulatory recognition in driver monitoring but remaining sensitive to pose variation, partial occlusion, and illumination change that are routine in naturalistic work environments [15]. The MG-YOLOv8 system [15] represents the current technical apex of this paradigm, combining YOLOv8 detection with the Practical Facial Landmark Detector (PFLD) for real-time PERCLOS computation with adaptive EWMA thresholding. However, this system was developed exclusively in automotive drowsiness contexts, and no evaluation on cognitive fatigue in desk-based settings has been reported.

Second-generation approaches moved to FACS-based AU intensity estimation through deep learning networks trained on the DISFA and BP4D+ datasets, enabling quantification of multi-AU temporal trajectories that carry richer fatigue information than geometric threshold features alone [23,29]. The multi-modal multi-label AU detection transformer of Wang et al. [29] demonstrated substantial improvements over CNN-based predecessors by learning AU co-occurrence statistics jointly with visual features. Third-generation architectures integrate facial AU sequences with other modalities through cross-modal attention. Zhao et al. [8] achieved accuracy of 87.01 percent and AUC of 86.88 percent for cognitive impairment recognition by fusing EEG, ECG, and facial AUs through a frequency-band adaptive encoder and ViT-based AU embedder with multi-query cross-modal attention. The CogMamba system [9] employs Mamba state-space models for non-contact cognitive load estimation from facial video, capturing slow temporal dynamics with linear computational complexity. Both systems, however, require physiological sensors or are restricted to automotive contexts, and neither incorporates any explainability mechanism.

### 4.3 HCI Behavioural Markers: A Neglected Modality

Despite consistent empirical evidence that keyboard and mouse interaction degrades measurably under cognitive fatigue through increased IKI variance, elevated error rates, less direct cursor trajectories, and extended click latencies [18], not one reviewed study integrates HCI behaviour as a primary, architecturally equal modality alongside speech and facial streams. The feasibility study reviewed as Keystroke Study [18] confirmed that psychomotor fatigue patterns manifest during naturalistic, unscripted daily typing and are detectable through automated keystroke analysis, providing proof-of-concept for passive, zero-overhead fatigue monitoring. The deployment importance of this modality extends beyond its diagnostic value to its practical characteristics: in professional environments where camera and microphone monitoring raises consent and privacy concerns, keystroke and cursor monitoring operates within existing digital infrastructure, requires no visible hardware, and can be processed entirely on-device. The failure of the research community to exploit this combination of theoretical motivation and practical accessibility represents the most consequential missed opportunity identified across all 20 reviewed papers.

### 4.4 Multimodal Fusion Architectures

Multimodal fusion approaches range from naive early concatenation through late voting ensembles to transformer-based cross-modal attention mechanisms [4,7,8,14]. The most architecturally significant advance reviewed is the cross-transformer of Far Poor et al. [7], which achieves AUC of 85.3 percent for MCI prediction by computing co-attention between speech, language, and facial video embeddings at an intermediate representation level, learning cross-modal feature interactions that concatenation-based fusion cannot capture. The recursive joint cross-modal attention framework of Praveen et al. [30] iteratively refines cross-modal representations through multiple attention rounds, enabling the model to resolve temporal misalignments between modalities through learned alignment rather than imposed synchronisation, an architectural principle directly applicable to the asynchronous temporal dynamics of cognitive fatigue.

Despite these advances, the specific challenge of asynchronous temporal modelling in cognitive fatigue remains entirely unaddressed. None of the reviewed systems explicitly models the temporal priority of HCI degradation over speech deterioration, or the sequential cascade from blink-rate changes through prosodic flattening to deliberate expression suppression failure. A fusion architecture that treats all modalities as providing equally current information at every time step systematically misrepresents the information structure of the cognitive fatigue process, limiting its theoretical validity and its practical detection capability at critical early stages of fatigue onset.

### 4.5 Explainable AI Integration

The XAI landscape across the reviewed studies is dominated by post-hoc attribution methods including SHAP [25], LIME, and Grad-CAM applied to the outputs of individual unimodal classifiers. The systematic review of XAI for speech-based cognitive decline detection [16] found that across 13 qualifying studies, SHAP and attention visualisation consistently identified pause patterns and speech rate as the primary acoustic discriminators, aligning with theoretical expectations from the speech fatigue literature [28]. Critically, all 13 identified studies applied XAI at the modality level. None provided explanations of cross-modal fusion decisions, meaning that even in systems using multiple modalities, the explanation of a prediction cannot reveal how inter-modal interactions contributed to the outcome. The broader healthcare XAI review of Noor et al. [17], covering 89 studies across 19 domains, confirmed the absence of standardised XAI evaluation metrics and insufficient stakeholder involvement in explanation design, both of which are fully manifest in the cognitive fatigue sub-domain. The EU AI Act [27] makes these gaps not merely scientifically significant but legally consequential: high-stakes automated systems must provide auditable explanations, and post-hoc unimodal attribution does not satisfy this requirement when the detection decision was generated by a multimodal fusion pipeline.

## 5. RESEARCH GAPS

The thematic analysis reveals six structural research gaps that reflect systematic blind spots in the field rather than incidental limitations of individual studies. Table 2 presents a structured summary of each gap, its evidentiary basis, and the proposed resolution within the EMI-CFD framework.

*Table 2. Formally identified research gaps: description, supporting evidence from reviewed literature, and proposed resolutions within the EMI-CFD framework.*

| Gap | Description | Evidence | Resolution |
|---|---|---|---|
| G1 | Invasive physiologica sensing dominates; no non contact tri-modal systen exists | Karim et al. [6]; FatigueSens [12]; HCSR Review [21] | Camera, microphone, and keyboar or mouse logger only; zer additional hardware |
| G2 | Naive early or late fusio discards asynchronous cross modal temporal dependencies | Virk et al. [4]; Arif et al. [14] Sun et al. [10] | Hierarchical cross-modal tempora attention with learned asynchronou alignment |
| G3 | HCI behaviour (keystroke cursor, latency) is absent as primary detection modality | Keystroke Study [18]; HCSR Review [21]; all 20 reviewec systems | HCI elevated to architecturally equal third modality alongside speech and face |
| G4 | XAI is post-hoc and unimodal cross-modal fusion-leve explainability is fully absent | XAI-Speech Review [16] Noor et al. [17]; Bridging Tex and Speech [19] | Intrinsic XAI module embeddec within fusion layer; modality saliency and cross-modal attributior |
| G5 | Cognitive fatigue conflatec with drowsiness and stress ground-truth annotations are temporally coarse | Dias et al. [5]; MG-YOLOv8 [15]; EEG Review [20] | NASA-TLX plus MFS plus performance decrement; per-minute continuous annotation protocol |
| G6 | No open, tri-modal ecologically valid cognitive fatigue benchmark datase currently exists | Virk et al. [4]; FatigueSens [12]; all 20 reviewed studies | Open, demographically diverse temporally annotated tri-moda corpus released publicly |

**Gap 1 (Invasive sensing dominance):** Twelve of the 20 reviewed papers rely on EEG, HRV, GSR, or multi-sensor wearables as their primary modality [6,12,13,21]. While physiologically informative, these modalities require electrode placement, skin contact, and controlled environments that are wholly incompatible with naturalistic occupational deployment. FatigueSense [12], though technically impressive in its multi-device wearable fusion, explicitly acknowledges that zero-burden seamless deployment remains unachieved. The paradigm shift required for population-scale fatigue monitoring demands a fundamentally different sensing philosophy centred on hardware that workers already possess.

**Gap 2 (Temporal poverty of fusion):** No reviewed system models the asynchronous temporal co-evolution of fatigue signatures across modalities. Feature concatenation [4], stacking ensembles [14], and even architecturally sophisticated attention systems [8,10] all operate on fixed temporal windows without tracking how the relationship between modality streams evolves over the hours-long timescale on which cognitive fatigue accumulates in real occupational settings. This is a fundamental theoretical deficiency, not a minor refinement opportunity.

**Gap 3 (HCI behaviour as unclaimed territory):** Despite compelling theoretical motivation and consistent empirical evidence [18], HCI behaviour is absent as a primary modality in every reviewed multimodal system. This neglect is doubly significant: it leaves substantial diagnostic information unexploited and forecloses the most privacy-compatible, zero-hardware deployment scenario, namely continuous cognitive fatigue monitoring of desk-based workers using only their existing keyboard and mouse.

**Gap 4 (Explainability vacuum at the fusion level):** Every XAI implementation identified in this review applies explanation methods post-hoc to individual unimodal encoder outputs [16,17]. The cross-modal fusion layer, where the most diagnostically important information is generated through learned inter-modal interactions, remains entirely unexplained in all reviewed systems. This renders fatigue determinations opaque to clinical practitioners and non-compliant with emerging regulatory requirements under the EU AI Act [27].

**Gap 5 (Ground-truth labelling crisis):** The operational definition of cognitive fatigue ground truth is neither consistent nor temporally adequate across reviewed studies. Many papers employ drowsiness-induction protocols involving sleep deprivation [15,20] and validate against driving performance decrement, measures appropriate for drowsiness but not for cognitive fatigue, which involves sustained mental effort under waking conditions with intact basic arousal. No reviewed study provides per-minute, multicomponent fatigue annotations combining subjective, performance-based, and physiological validation measures across a diverse participant cohort.

**Gap 6 (Absence of a tri-modal open benchmark):** No publicly accessible, tri-modal (speech, facial, and HCI), temporally annotated cognitive fatigue dataset exists. Existing datasets are fragmented by modality, predominantly proprietary, and dominated by automotive drowsiness or clinical cognitive impairment contexts [12,15]. This data scarcity systematically hampers cross-system comparison, ablation studies, and generalisation evaluation, which are the scientific infrastructure without which reliable progress cannot be demonstrated or reproduced.

## 6. DISCUSSION

The reviewed papers collectively reveal a research landscape rich in component-level technical achievement but structurally incomplete in integration. Deep learning AU detection [29], self-supervised speech representations [24], cross-modal transformer attention [7,8,30], and SHAP-based XAI [25] are each individually mature enough to contribute to a deployable system, yet no reviewed paper integrates all four capabilities in a single framework targeting cognitive fatigue specifically. The technical advances are real; what has been absent is the interdisciplinary synthesis required to combine them.

The theoretical significance of the HCI modality gap warrants particular emphasis because it reflects a disciplinary blind spot with historical roots rather than a deliberate methodological choice. The fatigue detection community developed primarily within neuroscience and physiological signal processing traditions, where HCI signals were not a natural object of study. The HCI research community, conversely, studied typing and mouse behaviour from the perspective of usability and user modelling, rarely connecting these behavioural measures to the physiological fatigue detection literature. The convergence of these traditions, treating keyboard and mouse interaction as a first-class neuroscientifically grounded fatigue modality, represents an interdisciplinary integration that has not yet occurred and that the EMI-CFD framework is designed to initiate.

The XAI gap carries equally serious structural implications. SHAP applied post-hoc to a unimodal speech classifier tells a practitioner which acoustic features most influenced that classifier's output. It cannot explain how the combination of specific speech patterns, facial AU activations, and typing rhythm changes jointly drove a multimodal fatigue determination. For a surgical team coordinator responding to an AI-generated fatigue alert, this fusion-level explanation is not a refinement but a necessity: without it, the alert cannot be contextualised, contested, or acted upon with professional accountability. The regulatory framework of the EU AI Act [27] makes this not a future aspiration but an immediate compliance requirement for any system claiming deployment in high-stakes professional contexts.

Several research directions extend beyond the three proposed objectives and warrant dedicated future investigation. Personalised baseline modelling is essential, given that cognitive fatigue manifests differently across individuals due to trait-level differences in neural efficiency, work style, and habituation. Federated learning offers a technically feasible path to personalised fatigue monitoring with strong privacy preservation by training models on local device data without centralising sensitive interaction logs. Cross-linguistic validation is required before any speech-based system can claim demographic generalisability, since prosodic norms, speech rate baselines, and F0 contour patterns vary substantially across languages and cultural contexts, and no reviewed speech-based system was evaluated on more than one language. Longitudinal fatigue accumulation modelling, which tracks fatigue build-up across days and weeks rather than within single sessions, addresses the clinically most consequential fatigue dynamic in surgical residency and long-haul aviation and remains entirely unaddressed by any reviewed system.

## 7. CONCLUSION

Cognitive fatigue is among the most consequential and underdetected states in contemporary human performance. This review has critically synthesised twenty studies spanning speech acoustic-prosodic detection, facial AU dynamics, HCI behavioural markers, multimodal fusion architectures, and XAI integration, and has found a field in which technically impressive component work has not been assembled into deployable, interpretable, and ecologically valid systems. Six structural research gaps explain why: invasive sensing hardware, naive fusion strategies, the absent HCI modality, post-hoc unimodal XAI, inconsistent ground-truth labelling, and the absence of any open benchmark dataset. The three research objectives derived from this analysis, comprising a non-contact tri-modal hierarchical attention framework, an embedded fusion-level XAI module, and a publicly released annotated tri-modal benchmark corpus, together define an integrated research agenda whose completion would bring the field meaningfully closer to its central goal: reliable, transparent, real-world detection of when human cognitive capacity is failing to meet the demands being placed upon it. The stakes in aviation, surgery, transportation, and the broader knowledge economy make this not merely a scientific imperative but an ethical one.

## REFERENCES

[1] M. A. S. Boksem and M. Tops, "Mental fatigue: Costs and benefits," Brain Research Reviews, vol. 59, no. 1, pp. 125–139, 2008, doi: 10.1016/j.brainresrev.2008.07.001.

[2] M. A. Al Imran, F. Nasirzadeh, and C. Karmakar, "Designing a practical fatigue detection system: A review on recent developments and challenges," Journal of Safety Research, vol. 90, pp. 100–114, 2024, doi: 10.1016/j.jsr.2024.07.002.

[3] A. B. Rafaih and K. Ari, "AI-driven approaches to managing surgeon fatigue and improving performance," Cureus, vol. 16, no. 12, Art. no. e75717, 2024, doi: 10.7759/cureus.75717.

[4] J. S. Virk, M. Singh, M. Singh, U. Panjwani, and K. Ray, "A multimodal feature fusion framework for sleep-deprived fatigue detection to prevent accidents," Sensors, vol. 23, no. 8, Art. no. 4129, 2023, doi: 10.3390/s23084129.

[5] M. Dias, F. Dörr, S. Garthof, S. Schäfer, J. Elmers, L. Schwed, and T. Ziemssen, "Detecting fatigue in multiple sclerosis through automatic speech analysis," Frontiers in Human Neuroscience, vol. 18, Art. no. 1449388, 2024, doi: 10.3389/fnhum.2024.1449388.

[6] E. Karim, H. R. Pavel, S. Nikanfar, A. Hebri, A. Roy, H. R. Nambiappan, and F. Makedon, "Cognitive fatigue assessment using physiological sensors during human-robot interaction for activities of daily living," Technologies, vol. 12, no. 3, Art. no. 38, 2024, doi: 10.3390/technologies12030038.

[7] F. Far Poor, H. H. Dodge, and M. H. Mahoor, "A multimodal cross-transformer-based model to predict mild cognitive impairment using speech, language, and vision," Computers in Biology and Medicine, vol. 182, Art. no. 109199, 2024, doi: 10.1016/j.compbiomed.2024.109199.

[8] M. Zhao, H. Gao, X. Qi, H. Yin, Y. Song, Y. Bai, and L. Zhao, "Multi-query cross-modal attention fusion for cognitive impairment recognition," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 33, pp. 2520–2530, 2025, doi: 10.1109/TNSRE.2025.3583110.

[9] Y. Xie and B. Guo, "CogMamba: Multi-task driver cognitive load and physiological non-contact estimation with multimodal facial features," Sensors, vol. 25, no. 18, Art. no. 5620, 2025, doi: 10.3390/s25185620.

[10] Y. Sun, M. Zhou, F. Cai, X. Yang, C. Li, and Y. Yi, "Noncontact multimodal sensing for driver distraction detection via behavioral-physiological fusion," IEEE Sensors Journal, vol. 25, no. 20, pp. 38576–38584, 2025, doi: 10.1109/JSEN.2025.3608125.

[11] S. A. Zendehbad, J. Ghasemi, and F. Samsami Khodadad, "FatigueNet: A hybrid graph neural network and transformer framework for real-time multimodal fatigue detection," Scientific Reports, vol. 15, Art. no. 33781, 2025, doi: 10.1038/s41598-025-00640-z.

[12] C. Kodikara, S. Wijekoon, and L. Meegahapola, "FatigueSense: Multi-device and multimodal wearable sensing for detecting mental fatigue," ACM Transactions on Computing for Healthcare, vol. 6, no. 2, Art. no. 14, pp. 1–36, 2025, doi: 10.1145/3709363.

[13] K. Chen, J. Liu, J. Zhu, Y. Xu, L. Zhang, and Z. Gao, "To assure aviation safety: Pilot fatigue detection based on short-term multimodal physiological signals," Frontiers in Human Neuroscience, vol. 20, Art. no. 1743936, 2026, doi: 10.3389/fnhum.2026.1743936.

[14] M. Arif, C. Ul Haq, I. Ahmed, M. H. Khatak, and M. A. Choudhry, "On fatigue detection for air traffic controllers based on fuzzy fusion of multiple features," Applied Sciences, vol. 12, no. 19, Art. no. 9578874, 2022, doi: 10.3390/app12199758.

[15] C. Chen, X. Liu, M. Zhou, Z. Li, Z. Du, and Y. Lin, "Lightweight and real-time driver fatigue detection based on MG-YOLOv8 with facial multi-feature fusion," Journal of Imaging, vol. 11, no. 11, Art. no. 385, 2025, doi: 10.3390/jimaging11110385.

[16] I. Kodrasi, E. Braun, N. Rosenblum, S. Sehad, J. A. Lissek, and K. Soreq, "A systematic review of explainable AI methods for speech-based cognitive decline detection," npj Digital Medicine, vol. 8, Art. no. 12657886, 2025, doi: 10.1038/s41746-025-02105-z.

[17] S. Noor, M. Mustafa, K. Ahmed, M. Z. Abedin, A. Masum, and B. Gupta, "Unveiling explainable AI in healthcare: Current trends, challenges, and future directions," WIREs Data Mining and Knowledge Discovery, 2025, doi: 10.1002/widm.70018.

[18] C. Benevenuta, G. La Torre, P. Incoronato, F. Romano, and L. Ruggeri, "Detection of mental fatigue through the analysis of keystroke dynamics: A feasibility study," International Journal of Environmental Research and Public Health, vol. 21, no. 4, Art. no. 11041424, 2024, doi: 10.3390/ijerph21040491.

[19] A. Pandey, J. Singh, and M. Kaur, "Bridging text and speech for emotion understanding: An explainable multimodal transformer fusion framework with unified audio-text attribution," Journal of Intelligence, vol. 13, no. 12, Art. no. 159, 2025, doi: 10.3390/jintelligence13120159.

[20] R. Ayuso-Moreno, C. Plata-Bello, E. Navarro-López, P. R. Villa-Santiago, and A. Hernández-Expósito, "EEG-based assessment of mental fatigue in students: A systematic review of measurement methods and data processing protocols," Applied Sciences, vol. 16, no. 1, Art. no. 234, 2026, doi: 10.3390/app16010234.

[21] K. Jin, A. Rubio-Solis, R. Naik, D. Leff, J. Kinross, and G. Mylonas, "Human-centric cognitive state recognition using physiological signals: A systematic review of machine learning strategies across application domains," Sensors, vol. 25, no. 13, Art. no. 4207, 2025, doi: 10.3390/s25134207.

[22] M. Jamalifard, Y. Lei, P. Azizinezhad, J. Fumanal-Idocin, and J. Andreu-Perez, "A neuro-symbolic system for interpretable multimodal physiological signals integration in human fatigue detection," arXiv, 2025, Art. no. 2603.24358. [Online]. Available: https://arxiv.org/abs/2603.24358.

[23] P. Ekman and W. V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Action. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[24] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 12449–12460.

[25] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 4765–4774.

[26] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190–202, 2016, doi: 10.1109/TAFFC.2015.2457417.

[27] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Official Journal of the European Union, L 2024/1689, Jul. 2024.

[28] J. Krajewski, D. Sommer, T. Schnupp, T. Laufenberg, C. Heinze, and M. Golz, "Applying nonlinear dynamics features for speech-based fatigue detection," in Proc. 7th International Conference on Methods and Techniques in Behavioral Research, 2010.

[29] L. Wang, S. Wang, and J. Qi, "Multi-modal multi-label facial action unit detection with transformer," arXiv, 2022, Art. no. 2203.13301. [Online]. Available: https://arxiv.org/abs/2203.13301.

[30] R. G. Praveen, C. Ecabert, J.-M. Odobez, and H. Bourlard, "Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) — ABAW, 2024, pp. 6491–6500.