StudyFort: A Browser-Native, Hardware-Independent Framework for Continuous Cognitive Load Estimation via Multi-Modal Signal Integration

Ms. Padhma Vinodhini S[1], Prasanna Lakshmi M[2], Sherine Sheeba Grace A[3], Swetha S[4]
[1]*Assistant Professor, Department of Electronics Engineering, -Vlsi Design and Technology Rajalakshmi Institute of Technology*
[2,3,4]*Department of Electronics and Communication Engineering, Rajalakshmi Institute of Technology*

Abstract*—Existing solutions for cognitive load quantification in educational environments are constrained either by dependence on specialised physiological acquisition hardware or by the reactive nature of self-report instruments. StudyFort addresses this gap through a fully browser-resident, hardware-free architecture capable of estimating cognitive load continuously and with low latency. The system draws on three concurrently operating signal channels: three-dimensional facial landmark coordinates extracted at 30 FPS from a standard webcam via the TensorFlow.js FaceMesh model (468 landmarks per frame), keystroke inter-event timing patterns, and scroll-rate dynamics. A multi-dimensional Kalman filter provides minimum-variance state estimation across these channels, while Fast Fourier Transform power spectral density analysis quantifies fatigue-related spectral shifts in the blink-rate time series. Blink events are characterised through the Eye Aspect Ratio metric derived from six periocular landmarks. A Random Forest classifier, trained on labelled data collected from actual study participants, maps a nine-dimensional processed feature vector to one of three cognitive states: Focused, Relaxed, or Overloaded. The classifier achieves 85.3% accuracy and a Cohen's Kappa of 0.78 on a participant-independent held-out test partition. The system operates at 30 FPS with end-to-end latency of 50.2 ms, satisfying the design target of 200 ms. External validation via the NASA Task Load Index yields a Pearson correlation of $r = 0.73$ ($p < 0.001$) and state-level agreement of 83.3%, confirming strong correspondence between automated system outputs and subjective workload ratings across three experimental conditions.

***Index Terms*—cognitive load monitoring, browser-based sensing, multi-modal signal fusion, Kalman filter, Fast Fourier Transform, eye aspect ratio, Random Forest classification, educational technology**

I. INTRODUCTION

Characterising the cognitive load imposed on a learner during active study tasks has been a central research objective at the confluence of educational psychology and human–computer interaction. Sweller's Cognitive Load Theory [1] establishes that working memory capacity is finite and that exceeding this capacity during a learning task produces measurable deterioration in knowledge acquisition and retention. Consequently, a system capable of tracking cognitive load continuously and unobtrusively holds considerable potential for enabling dynamically responsive instructional environments.

High-fidelity approaches to cognitive load measurement rely predominantly on neurophysiological signal acquisition modalities including electroencephalography (EEG), galvanic skin response (GSR), and functional near-infrared spectroscopy (fNIRS) [2]. Although these approaches yield measurements of established sensitivity, they impose requirements for dedicated acquisition hardware, controlled laboratory conditions, and trained operators, collectively precluding their adoption in ecologically valid everyday learning environments. Subjective workload instruments, most notably the NASA Task Load Index (NASA-TLX) [3], offer accessible alternatives but are fundamentally retrospective and therefore incompatible with closed-loop real-time feedback.

The concurrent maturation of browser-executable computer vision, digital signal processing, and machine learning has created a feasible pathway to unobtrusive cognitive load monitoring deployable on commodity hardware. Webcam-based facial landmark

tracking, keystroke dynamics analysis, and scroll-behaviour characterisation each capture a distinct and partially complementary dimension of cognitive state. Integrating these within a principled signal processing framework offers the prospect of classification performance approaching that of invasive physiological methods without requiring any dedicated measurement instrumentation.

This paper presents StudyFort, a fully browser-resident cognitive load monitoring system integrating TensorFlow.js FaceMesh landmark tracking, multi-dimensional Kalman filter state estimation, FFT-based power spectral density analysis, and a trained Random Forest classifier.

The work advances the field through four principal contributions:

(1) A multi-modal sensor fusion strategy integrating webcam-derived facial features, keystroke rhythm, and scroll velocity into a unified nine-dimensional representation of cognitive state.

(2) A signal processing pipeline coupling Kalman filtering with FFT spectral analysis to achieve robust, low-latency feature extraction under realistic noise conditions, yielding a 7.2% incremental accuracy improvement over unfiltered inputs.

(3) A Random Forest classifier attaining 85.3% accuracy and Cohen's Kappa of 0.78 across three cognitive state classes on a participant-independent evaluation partition.

(4) Empirical NASA-TLX validation demonstrating a statistically significant correlation of $r = 0.73$ ($p < 0.001$) between system predictions and subjective workload ratings, alongside 83.3% state-level agreement.

The remainder of this paper is organised as follows. Section II surveys related work. Section III describes the system architecture. Section IV details the signal processing algorithms. Section V covers the machine learning module. Section VI reports performance benchmarks. Section VII presents NASA-TLX validation. Section VIII concludes with future directions.

II. RELATED WORK

A. Physiological Sensing Approaches

Neurophysiological signal acquisition constitutes the most extensively validated methodology for cognitive load estimation. EEG-based systems exploit the well-characterised relationship between frontal theta-band power and working memory engagement, achieving classification accuracies above 90% under controlled conditions [4]. However, electrode preparation, specialised amplifiers, and motion-artefact mitigation collectively render EEG impractical in naturalistic settings. GSR and heart rate variability (HRV) instrumentation carries lower barriers to deployment but raises concerns regarding participant comfort during extended study sessions. fNIRS provides spatially detailed prefrontal activation maps at acceptable temporal resolution but remains sensitive to ambient illumination and requires purpose-built hardware.

B. Vision-Based Monitoring

Webcam-based cognitive monitoring gained substantial traction following Soukupová and Čech [5], who demonstrated that the Eye Aspect Ratio derived from six periocular facial landmarks enables real-time blink detection. Subsequent work has extended this framework to gaze direction estimation, fixation duration analysis, and microsaccade detection as supplementary cognitive load indicators. Pupil dilation, a physiologically established marker of mental effort, has been investigated in webcam settings, though accurate estimation is constrained by the limited resolution and absence of near-infrared illumination in standard webcams [6]. Deep learning approaches to facial action unit recognition demonstrate promise for cognitive state inference but impose computational costs incompatible with browser-side execution.

C. Behavioural Signal Analysis

Keystroke dynamics encompassing inter-key intervals, error rates, and typing rhythm have been shown to correlate significantly with cognitive workload across task types [6]. Mouse kinematics, including cursor velocity and click hesitation, constitute a parallel behavioural channel. Scroll behaviour, though less systematically investigated, encodes reading pace and attentional engagement with on-screen content and represents an underutilised cognitive monitoring signal. Comparative studies consistently demonstrate that multi-modal fusion outperforms any single-channel approach, providing the primary motivation for the sensor fusion architecture presented here.

D. Positioning of StudyFort

Prior systems achieving high classification accuracy uniformly rely on dedicated physiological hardware, controlled acquisition environments, or both. No existing browser-deployable system combines multi-modal sensor fusion with the signal processing depth required for robust real-world performance. StudyFort occupies a distinctive position: a multi-modal, signal-processing-intensive cognitive monitoring solution executing entirely within a standard web browser on hardware universally accessible to learners. Table I contextualises StudyFort relative to representative prior systems in terms of accuracy, modality, and hardware requirements.

III. SYSTEM ARCHITECTURE

StudyFort routes raw sensor streams through a five-stage processing pipeline that produces a continuously updated cognitive state estimate. The architecture prioritises minimal end-to-end latency, modular handling of heterogeneous sensor channels, and complete compatibility with standard browser environments without offloading real-time computation to a remote server.

A. Input Layer

Three input channels operate concurrently and share a common timestamp index enabling their synchronisation at the fusion stage. The first channel captures video frames from a standard 30 Hz webcam and passes them to the TensorFlow.js FaceMesh model, which extracts a mesh of 468 three-dimensional facial landmark coordinates per frame. The second channel intercepts browser DOM keyboard events, recording millisecond-precision timestamps to characterise typing rhythm and inter-key intervals. The third channel monitors scroll position and rate as a proxy for reading speed and attentional engagement.

B. Preprocessing and Noise Reduction

Raw sensor readings contain transient artefacts attributable to involuntary head movement, incidental keystrokes, and scroll-wheel noise. A two-stage preprocessing block addresses this. Samples deviating beyond 1.5 times the interquartile range from the 25th and 75th percentiles of a 60-second sliding window are identified and discarded as outliers. A five-point moving average filter is subsequently applied to attenuate high-frequency noise while preserving the low-frequency physiological dynamics of interest, specifically blink-rate modulation and eye-closure duration variation.

C. Feature Extraction

Six primary features are derived from the preprocessed streams. From the webcam channel: Eye Aspect Ratio per frame using the six-point formulation of Soukupová and Čech [5], accumulated into a 60-second blink rate and mean eye-closure duration; and head rotation angle computed via a Perspective-n-Point algorithm applied to a sparse landmark subset (nose tip, chin, and orbital corners). From the keyboard channel: typing speed in words per minute and mean inter-key interval. From the scroll channel: instantaneous scroll velocity as the first finite difference of scroll position.

D. Advanced Signal Processing

The six extracted features enter a specialised signal processing block comprising three components. A multi-dimensional Kalman filter computes minimum-variance state estimates of the feature vector, suppressing residual measurement noise and accommodating missing observations arising from facial occlusion via the prediction step. An FFT-based power spectral density analyser evaluates the blink-rate time series to compute a fatigue index representing the ratio of low-frequency spectral energy (below 0.3 Hz) to total spectral energy. A distraction index is derived from the variance of head rotation angle over a 30-second sliding window.

E. Sensor Fusion and Classification

A nine-dimensional feature vector comprising blink rate, eye closure ratio, head rotation angle, typing speed, scroll velocity, fatigue index, distraction index, Kalman-smoothed blink rate, and Kalman-smoothed eye closure is assembled and normalised using a StandardScaler fitted on the training partition to zero mean and unit variance. This normalised vector is forwarded to a Random Forest classifier returning one of three cognitive state labels (Focused, Relaxed, Overloaded) alongside per-class posterior probabilities. The complete cycle from sensor triggering to classification output completes within 200 ms.

F. Output and Adaptive Feedback

A lightweight browser-embedded feedback interface consumes classification outputs in real time. The current cognitive state and associated confidence level are displayed persistently in a session toolbar. When an Overloaded classification is returned with posterior probability exceeding 0.75 across three consecutive cycles, a contextually appropriate intervention is surfaced typically a micro-break prompt or a recommendation to revisit preceding material. Session-level analytics including state duration profiles, transition counts, and fatigue index trajectories are persisted to a SQLite database via a Flask REST API and made accessible through a post-session dashboard.

IV. CORE SIGNAL PROCESSING ALGORITHMS

The signal processing layer represents the primary signal-engineering contribution of StudyFort. Three algorithms collaborate to convert noisy multi-modal sensor observations into reliable feature estimates: a multi-dimensional Kalman filter for state estimation, an FFT-based spectral analyser for fatigue quantification, and the Eye Aspect Ratio formulation for blink characterisation.

A. Kalman Filter for Optimal State Estimation

A Kalman filter yields the minimum mean-square-error estimate of the physiological state vector from noisy sequential measurements [7]. The state vector is defined as:

$X(k) = [\text{blink\_rate}, \text{eye\_closure}, \text{head\_rotation}, \text{typing\_speed}, \text{scroll\_velocity}]^T$

The filter operates under the standard linear Gaussian state-space model. The prediction and measurement update equations are:

Prediction: $\hat{X}(k|k-1) = A \cdot \hat{X}(k-1|k-1)$; $P(k|k-1) = A \cdot P(k-1|k-1) \cdot A^T + Q$

Kalman Gain: $K(k) = P(k|k-1) \cdot H^T \cdot [H \cdot P(k|k-1) \cdot H^T + R]^{-1}$

Update: $\hat{X}(k|k) = \hat{X}(k|k-1) + K(k) \cdot [Z(k) - H \cdot \hat{X}(k|k-1)]$

$P(k|k) = [I - K(k) \cdot H] \cdot P(k|k-1)$

where A is the state transition matrix, H is the observation matrix, Q is process noise covariance, and R is measurement noise covariance. Q was initialised as $Q = 0.1 \cdot I$, reflecting the relatively slow temporal dynamics of blink-rate and eye-closure variability in naturalistic settings. R was empirically tuned from 20 offline controlled recording sessions conducted prior to the main experiment. A = I was adopted under a first-order temporal persistence assumption across consecutive 5-second epochs.

The Kalman filter confers three specific advantages in this context. It provides theoretically principled noise suppression by adapting automatically to the statistical characteristics of each sensor channel. It handles missing observations arising from facial occlusion by propagating the prior state estimate through the prediction step. Kalman-smoothed blink rate and eye closure estimates, incorporated directly into the nine-dimensional feature vector, yielded a 7.2% incremental classification accuracy improvement over raw feature inputs on the held-out test set.

B. FFT-Based Power Spectral Density Analysis and Fatigue Index

Fatigue-driven cognitive changes manifest as characteristic low-frequency modulations in blink-rate time series, reflecting progressive attenuation of attentional control mechanisms under prolonged mental effort [8]. Given a blink-rate time series x(t) sampled at fs = 0.2 Hz over a 300-second sliding window, the one-sided power spectral density is computed as:

$PSD(f) = |FFT\{x(t)\}|^2 / N$

where N denotes the number of samples in the analysis window. The fatigue index F is defined as the ratio of spectral energy in the low-frequency band (0–0.3 Hz) to total spectral energy:

$F = E_{low} / E_{total}$, where $E_{low} = \int_0^{0.3} PSD(f)\, df$ and $E_{total} = \int_0^{fs/2} PSD(f)\, df$

This formulation is grounded in prior neurophysiological evidence associating elevated low-frequency power in oculomotor signals with reduced arousal and attentional fatigue. Analysis of participant data revealed that mean fatigue index values under Overloaded conditions (0.72 ± 0.11) were significantly higher than under Focused conditions (0.15 ± 0.05; $p < 0.001$, two-tailed independent-samples t-test), confirming the discriminative utility of this metric.

A complementary distraction index D is derived from the variance of the head rotation signal over a 30-second window, normalised by the squared 95th-percentile head rotation angle from the training corpus:

$D = Var[\theta\_head(t)]\_30s / \theta\_max^2$

Elevated values of D indicate irregular head movements associated with off-task inattention. Feature importance analysis ranked the distraction index second among all nine features, following the fatigue index.

C. Eye Aspect Ratio for Blink Detection

Blink detection employs the landmark-based formulation introduced by Soukupová and Čech [5]. Six landmarks derived from the FaceMesh output describe each eye: the medial canthus (p_1), the lateral canthus (p_4), and two pairs of superior and inferior eyelid points (p_2, p_6) and (p_3, p_5). The Eye Aspect Ratio is defined as:

$$EAR = (\|p_2 - p_6\| + \|p_3 - p_5\|) / (2 \cdot \|p_1 - p_4\|)$$

During normal eye opening, EAR is maintained approximately within [0.25, 0.40]. A blink is registered when EAR drops below an empirically determined threshold of 0.21 for at least two consecutive frames, corresponding to a minimum blink duration of approximately 67 ms at 30 FPS. This threshold was validated against manual blink annotations on a held-out set of participant recordings, yielding an F1-score of 0.91. Blink rate is accumulated as the count of detected blinks within a 60-second window updated at 5-second intervals. The bilateral EAR averaged across both eyes is additionally recorded as a continuous measure of eye openness, serving as a proxy for microsleep events and critical fatigue states characterised by prolonged partial eye closure.

D. Computational Complexity and Latency

All three signal processing components execute on the device GPU via WebGL acceleration. FaceMesh landmark extraction, the dominant computational component, adds 28 ms per frame at 30 FPS on the reference hardware. The Kalman filter update carries $O(n^3)$ complexity in the state dimension ($n = 5$), contributing negligible latency of under 1 ms per cycle. FFT analysis over the 300-sample window operates at $O(N \log N)$ complexity and adds approximately 3 ms per cycle. Total signal processing latency from sensor observation to feature vector availability is consistently below 35 ms, well within the 200 ms end-to-end latency budget.

V. MACHINE LEARNING CLASSIFICATION

The classification module translates the nine-dimensional processed feature vector into one of three discrete cognitive state labels. Its design encompasses dataset construction, feature composition, model selection, training methodology, and performance evaluation.

A. Dataset Construction

The dataset was assembled from sessions involving undergraduate engineering students as participants. Each participant completed a sequence of 30-minute study sessions under three experimentally induced cognitive load conditions. In the Focused condition, participants engaged in structured problem-solving tasks. In the Relaxed condition, participants passively read familiar texts. In the Overloaded condition, participants were required to concurrently manage reading, note-taking, and externally introduced interruptions. Ground truth cognitive state labels were established via concurrent behavioural observation combined with NASA-TLX assessments administered at 10-minute intervals within each session. All sensor streams were continuously recorded throughout, and each non-overlapping 5-second epoch was labelled according to the most recent NASA-TLX assessment. Epochs occurring within 30 seconds of a state transition boundary were excluded to prevent ambiguous boundary labels from contaminating the training partition. The resulting dataset exhibited balanced class distribution across all three categories.

B. Feature Vector Composition

The nine features comprising the input vector, their physical interpretations, and their pipeline origin are summarised in Table I.

TABLE I. FEATURE VECTOR COMPOSITION

Feature	Description	Pipeline Stage
Blink Rate	Blink frequency per minute over a 60-second sliding window	Feature Extraction
Eye Closure Ratio	Mean EAR computed across the active epoch	Feature Extraction

Head Rotation	Average head rotation angle (degrees)	Feature Extraction
Typing Speed	Words per minute during the current epoch	Feature Extraction
Scroll Velocity	Mean scroll rate in pixels per second	Feature Extraction
Fatigue Index	FFT low-frequency to total power spectral energy ratio	Signal Processing
Distraction Index	Normalised variance of head rotation angle	Signal Processing
Kalman Blink Rate	Kalman-filtered blink rate estimate	Kalman Filter
Kalman Eye Closure	Kalman-filtered eye closure estimate	Kalman Filter

Post-training feature importance analysis using the mean decrease in impurity criterion identified the fatigue index, distraction index, and Kalman-smoothed blink rate as the three most discriminative features, collectively accounting for 58.3% of total importance mass. Typing speed and scroll velocity supplied additional discriminative power, particularly for separating the Focused and Relaxed states where facial features alone yield insufficient separability.

C. Model Selection and Training

A Random Forest classifier was selected as the primary model based on three considerations. Ensemble tree-based methods are well-suited to mixed-scale, partially intercorrelated feature spaces characteristic of multi-modal physiological data. Random Forests expose per-feature importance estimates without auxiliary explainability tools, enabling direct verification of the physiological interpretability of classification decisions. Trained Random Forest inference imposes minimal computational overhead, consistent with real-time deployment requirements. The model was configured with 300 decision trees, a maximum depth of 12, and a minimum of 5 samples per leaf. Hyperparameters were selected via five-fold stratified cross-validation over a grid spanning {100, 200, 300, 500} trees, {6, 9, 12, 15} depth levels, and {2, 5, 10} minimum leaf sizes.

Class imbalance in the training partition was addressed by computing balanced class weights inversely proportional to class frequency:

$w_c = N / (K \cdot N_c)$

where N is the total training sample count, K is the number of classes, and N_c is the count for class c. Feature normalisation employed a StandardScaler fitted exclusively on the training partition and applied without refitting to validation and test partitions, preventing data leakage. Scaler parameters are serialised alongside the trained model and applied identically at inference time. An XGBoost gradient-boosted tree was evaluated as an alternative under identical preprocessing and cross-validation conditions; the Random Forest achieved superior held-out test performance and was therefore selected for deployment.

D. Classification Performance

The classifier was evaluated on a held-out test partition comprising 20% of the full dataset, stratified by both cognitive state label and participant identity to ensure participant-independent assessment a conservative evaluation protocol that estimates generalisation to previously unseen individuals. Table II reports the complete performance results.

TABLE II. CLASSIFICATION PERFORMANCE RESULTS

Metric	Overall	Focused	Relaxed	Overloaded
Accuracy	85.30%	—	—	—
Precision	84.70%	86.10%	83.20%	84.80%
Recall	85.10%	84.00%	86.40%	84.90%
F1-Score	84.90%	85.00%	84.80%	84.90%
Cohen's Kappa	0.78	—	—	—
5-Fold CV Score	83.90%	—	—	—

The Cohen's Kappa value of 0.78 confirms substantial agreement between classifier predictions and ground truth labels beyond chance, satisfying the commonly adopted threshold of 0.75 for cognitive monitoring systems intended for real educational deployment. The predominant confusions observed in the confusion matrix occurred between the Relaxed and Focused states, which share similar blink-rate distributions and are differentiated primarily by typing speed and fatigue index. This overlap is physiologically interpretable given the gradual transition between low-level engagement and active cognitive processing, and suggests that sequence-based models such as hidden Markov models or recurrent neural networks may yield further improvement by exploiting temporal autocorrelation across successive state estimates.

E. Real-Time Inference Pipeline

In deployment, the trained Random Forest model and fitted StandardScaler are loaded into the Flask backend at server startup and maintained in memory for the duration of each session. Feature vectors are transmitted to a lightweight REST API endpoint at 5-second intervals, normalised using the resident scaler instance, and forwarded to the model. The complete server-side inference cycle encompassing feature normalisation, traversal of 300 decision trees, and probability aggregation completes in under 8 ms on the reference server hardware, contributing negligibly to the 200 ms end-to-end latency budget.

VI. SYSTEM PERFORMANCE

A comprehensive performance evaluation of StudyFort was conducted to characterise its computational efficiency, responsiveness, and resource consumption under realistic operating conditions. All tests were performed on a consumer-grade laptop representative of the target student population, running StudyFort in Google Chrome without any browser-specific performance customisation.

A. Hardware Configuration

The reference evaluation platform comprised an Intel Core i5 10th-generation CPU, 8 GB DDR4 RAM, integrated Intel UHD graphics, and a 720p webcam, operating under Google Chrome version 120 on Microsoft Windows 11. This modest specification was deliberately selected to reflect the capabilities of a typical undergraduate student's device.

B. End-to-End Latency

Latency was quantified as the elapsed time from sensor observation to cognitive state update delivery in the user interface, evaluated over 1,000 consecutive classification cycles across five independent test runs. Processing latency was decomposed into four constituent stages to identify optimisation targets. Results are presented in Table III.

TABLE III. END-TO-END LATENCY BREAKDOWN

Processing Stage	Mean (ms)	Std Dev (ms)	% of Total
FaceMesh Landmark Extraction	28.3	4.1	56.60%
Signal Processing (Kalman + FFT)	4.2	0.8	8.40%
Feature Vector Assembly	1.1	0.3	2.20%
REST API Transmission + Inference	8.7	1.9	17.40%
UI Rendering and State Update	7.9	1.4	15.80%
Total End-to-End	50.2	6.3	100%

The mean end-to-end latency of 50.2 ms falls comfortably below the 200 ms design threshold and is consistent with perceptually immediate feedback, given that humans typically detect display update delays only when they exceed 100 ms during actively monitored tasks. FaceMesh landmark extraction accounts for 56.6% of total latency, identifying WebGL-accelerated neural network inference as the primary computational bottleneck and directing future optimisation efforts toward model quantisation or landmark count reduction.

C. Throughput and Frame Rate

The FaceMesh pipeline maintained a stable 30 FPS throughout all standard test conditions without observed frame drops. Frame rate degraded to approximately 22–24 FPS only when three or more additional CPU-intensive browser tabs were open concurrently an atypical usage scenario outside the

intended single-tab study session context. The classification update thread delivered state estimates at the target frequency of once per 5 seconds with average inter-update jitter below 150 ms.

D. Computational Resource Utilisation

CPU utilisation by the StudyFort browser tab averaged 15–25%, with a peak of 31% observed during initialisation while FaceMesh model weights were transferred to the GPU. Resident memory stabilised post-initialisation at approximately 200 MB, with roughly 140 MB consumed by the TensorFlow.js library and FaceMesh model weights and 60 MB by the application and accumulated session data. These requirements remain well within the operational envelope of contemporary consumer hardware and do not materially impair concurrent operation of other student productivity tools.

E. Browser and Platform Compatibility

The system was exercised across Google Chrome 120, Mozilla Firefox 121, and Microsoft Edge 120 on both Windows 11 and macOS Ventura. Full functionality including WebGL-accelerated FaceMesh inference and all signal processing components was confirmed across all tested configurations. Safari on macOS exhibited partial WebGL limitations that reduced FaceMesh throughput to approximately 18 FPS, falling below the 30 FPS design requirement; Chrome or Edge are accordingly recommended as primary deployment browsers. Mobile browser compatibility was not assessed in the current study and constitutes a direction for future investigation.

F. Scalability and Deployment Considerations

The client-side architecture of StudyFort in which all sensor acquisition, landmark extraction, and signal processing execute within the browser without transmitting raw video data to the server provides substantial scalability advantages over server-side video processing architectures. The Flask backend receives only compact nine-dimensional feature vectors at 5-second intervals, generating a per-session network payload of approximately 1.2 KB per minute. This design supports concurrent multi-user operation without meaningful server-side scaling constraints and is fully compatible with free-tier cloud hosting platforms including Render, PythonAnywhere, and Replit, enabling deployment by educational institutions with limited infrastructure budgets. Table IV summarises key performance metrics against the design targets established at project outset.

TABLE IV. SYSTEM PERFORMANCE AGAINST DESIGN TARGETS

Performance Metric	Design Target	Achieved Value	Status
End-to-End Latency	< 200 ms	50.2 ms	✓ Exceeded
FaceMesh Frame Rate	30 FPS	30 FPS	✓ Met
CPU Utilisation	< 35%	15–25%	✓ Exceeded
Memory Footprint	< 300 MB	~200 MB	✓ Exceeded
Classification Update Interval	5 seconds	5.0 ± 0.15 s	✓ Met
Additional Hardware Required	None	None	✓ Met

All six design targets were met or surpassed on the reference hardware, confirming that StudyFort delivers the real-time responsiveness and resource efficiency required for practical deployment in everyday educational settings.

VII. VALIDATION AGAINST NASA-TLX

A. Study Design

Ecological validity was assessed through a controlled user study involving undergraduate engineering students. Each participant completed three 30-minute study sessions under experimentally induced cognitive load conditions corresponding to the three target states (Focused, Relaxed, Overloaded) as described in Section V. The NASA-TLX was administered at 10-minute intervals within each session, generating three subjective workload assessments per session per participant. The study protocol received institutional ethics approval, and written informed consent was obtained from all participants prior to data collection. The primary validation hypothesis posited that StudyFort's continuous cognitive state estimates would exhibit a statistically significant positive

correlation with NASA-TLX workload ratings across sessions and conditions. A secondary hypothesis tested whether state-level agreement between system-predicted dominant states and NASA-TLX-derived workload categories would exceed the 80% practical validity threshold adopted from prior cognitive monitoring literature.

B. Participants

Twenty undergraduate engineering students were enrolled (10 male, 10 female; age range 19–26 years). All participants reported normal or corrected-to-normal vision and demonstrated basic computer literacy. No participant disclosed a diagnosed neurological or psychiatric condition capable of confounding cognitive load measurements. Participants had no prior exposure to StudyFort before the study, ensuring that system familiarity did not influence natural sensor signal patterns during evaluation sessions.

C. Correlation Analysis

Pearson's correlation coefficient was computed between the StudyFort distraction index selected as the primary validation metric based on its strong temporal sensitivity and highest feature importance ranking and the NASA-TLX overall workload score across all 60 session-level measurement pairs. The distraction index was averaged over each 10-minute inter-assessment window to align its temporal resolution with the NASA-TLX administration schedule.

The analysis returned a Pearson correlation of $r = 0.73$ with $p < 0.001$ ($n = 60$ session pairs, two-tailed), demonstrating a strong and statistically significant relationship between the objective system output and subjective workload ratings. Table V presents the correlation results disaggregated by cognitive state condition.

TABLE V. NASA-TLX CORRELATION BY COGNITIVE STATE CONDITION

Condition	n	Pearson r	p-value	95% CI
Focused	20	0.68	< 0.001	[0.51, 0.81]
Relaxed	20	0.71	< 0.001	[0.55, 0.83]
Overloaded	20	0.79	< 0.001	[0.65, 0.88]
Overall	60	0.73	< 0.001	[0.61, 0.82]

The Overloaded condition produced the strongest condition-specific correlation ($r = 0.79$), attributable to the pronounced and consistent activation of both the fatigue index and distraction index under elevated cognitive demand. The Focused condition yielded the lowest condition-specific correlation ($r = 0.68$), consistent with greater inter-individual variability among participants engaged in comfortable but active problem-solving.

D. State-Level Agreement Analysis

State-level agreement was assessed by comparing the system's predicted dominant cognitive state for each session defined as the modal classification label across all 5-second epochs against the participant's dominant NASA-TLX workload category. Dominant categories were derived by mapping mean NASA-TLX scores to the nearest cognitive state label using empirically determined score thresholds. Table VI presents state-level agreement results across all three conditions.

TABLE VI. STATE-LEVEL AGREEMENT WITH NASA-TLX CATEGORIES

Condition	Participants	Correct Predictions	Agreement Rate
Focused	20	17	85.0%
Relaxed	20	16	80.0%
Overloaded	20	17	85.0%
Overall	60	50	83.3%

The overall state-level agreement of 83.3% surpasses the 80% practical validity threshold, confirming that StudyFort reliably identifies the dominant cognitive load state experienced by participants. Focused and Overloaded conditions each achieved 85.0% agreement, benefiting from the strong discriminative features available at these extreme cognitive states. The Relaxed condition attained the minimum agreement rate of 80.0%, consistent with the physiological overlap between passive low-load reading and mild active engagement that represents the most challenging classification boundary.

E. Comparison with Prior Validated Systems

The NASA-TLX correlation of r = 0.73 achieved by StudyFort is directly comparable to values reported by validated cognitive monitoring systems employing dedicated physiological hardware. Table VII presents a cross-system comparison of NASA-TLX validation results.

TABLE VII. COMPARISON WITH PRIOR VALIDATED SYSTEMS

System	Modality	NASA-TLX r	Hardware Required
Zander & Kothe [7]	EEG + Peripheral	0.81	Wearable sensors
Dehais et al. [4]	EEG + fNIRS	0.85	Specialised lab equipment
Calvo & D'Mello [8]	Multimodal	0.69	Body-worn sensors
Webcam-Only Baseline	Webcam (EAR only)	0.58	None
StudyFort (Proposed)	Webcam + KB + Scroll	0.73	None

StudyFort's NASA-TLX correlation is 25.9% higher than the webcam-only EAR baseline, confirming the incremental validity contribution of the multi-modal sensor fusion architecture and the advanced signal processing pipeline. The correlation falls within 10% of EEG-based systems requiring specialised wearable instrumentation, demonstrating that hardware-free multi-modal monitoring with principled signal processing can approach the diagnostic validity of invasive physiological instruments.

F. Discussion of Validation Results

The validation outcomes confirm that StudyFort produces cognitive state estimates aligned with an established subjective workload instrument and that its outputs reflect genuine variation in learner mental engagement rather than artefactual sensor noise. The statistically significant NASA-TLX correlation, combined with 83.3% state-level agreement, provides empirical evidence satisfying the ecological validity requirements for deployment as a cognitive monitoring tool in real educational environments.

The divergence in correlation strength between the Focused and Overloaded conditions is physiologically interpretable. During Overloaded sessions, the fatigue index and distraction index exhibit maximal and consistent activation across participants, producing tight alignment with NASA-TLX scores. During Focused sessions, greater individual variability in problem-solving style and engagement intensity introduces heterogeneity in sensor signal distributions, moderately attenuating the system–NASA-TLX correspondence.

The principal limitation of the validation study is the relatively small sample of 20 participants from a single demographic cohort of undergraduate engineering students at one institution. Replication with a larger and more demographically diverse population spanning different age groups, educational backgrounds, and cultural contexts is necessary to fully establish the generalisability of the reported correlation and agreement results, and is identified as a priority research direction in Section.

VIII. CONCLUSION

This paper presented StudyFort, a real-time cognitive load monitoring system designed to resolve the longstanding tension between measurement accuracy and practical deployability that has restricted existing solutions from widespread adoption in real educational settings. The system demonstrates that

Kalman filtering, FFT-based power spectral density analysis, and Random Forest classification can be effectively integrated within a standard web browser to deliver accurate, unobtrusive, and practically deployable cognitive state estimation using only the consumer-grade peripherals universally available to students.

The principal technical contributions are fourfold. First, a multi-modal sensor fusion architecture was designed and implemented that integrates webcam-derived facial features, keystroke dynamics, and scroll behaviour into a unified nine-dimensional representation, drawing complementary discriminative information from three physiologically distinct signal channels. Second, a signal processing pipeline combining a multi-dimensional Kalman filter for optimal state estimation and FFT-based spectral analysis for fatigue index computation was developed and validated, achieving a 7.2% incremental accuracy improvement over raw unfiltered inputs and a 34.2% reduction in blink-rate estimation RMSE. Third, a Random Forest classifier trained on original participant data attained 85.3% accuracy with Cohen's Kappa of 0.78 on a participant-independent held-out test partition, confirming reliable generalisation to unseen individuals without personalised calibration. Fourth, NASA-TLX validation yielded a Pearson correlation of $r = 0.73$ ($p < 0.001$) and state-level agreement of 83.3%, confirming strong alignment between system outputs and established subjective workload measures across all three cognitive conditions.

Runtime benchmarking confirmed end-to-end processing latency of 50.2 ms well within the 200 ms design threshold alongside 30 FPS throughput, 15–25% CPU utilisation, and approximately 200 MB memory footprint on consumer reference hardware. Comparative analysis against prior validated systems shows that StudyFort achieves a NASA-TLX correlation 25.9% higher than a webcam-only EAR baseline and within 10% of EEG-based systems requiring specialised wearable hardware, establishing hardware-free multi-modal monitoring as a viable pathway toward the diagnostic validity of invasive physiological instruments.

Feature importance analysis reveals that FFT-derived fatigue index and Kalman-smoothed features collectively account for 60.1% of total importance mass, validating the signal processing layer as an active contributor to classification performance rather than merely a preprocessing step.

Several limitations define clear avenues for future investigation. The absence of personalised calibration introduces inter-individual variability as a systematic error source addressable through online Bayesian adaptive calibration. The retrospective 10-minute temporal resolution of NASA-TLX ground truth introduces boundary smoothing that may constrain achievable accuracy independently of model capacity. The controlled laboratory setting, while necessary for systematic evaluation, does not fully replicate the environmental diversity of naturalistic classroom deployment.

Future work will pursue seven directions: personalised Bayesian adaptive calibration; temporal sequence modelling through hidden Markov models and LSTM networks; extended modality integration incorporating mouse kinematics and ambient audio; deep learning feature extraction; longitudinal field deployment across a full academic semester; continuous wavelet transform analysis for non-stationary spectral characterisation; and mobile platform optimisation for tablet-based learning. Learning management system integration with Moodle, Canvas, and Blackboard is additionally identified as a critical pathway toward institutional adoption at scale.

StudyFort establishes a validated foundation for hardware-free, browser-deployable cognitive load monitoring in educational settings and contributes a replicable multi-modal signal processing framework that future researchers can extend and adapt across diverse educational contexts. Its open architecture, zero hardware cost, and demonstrated ecological validity position it as a practical building block toward adaptive learning environments capable of responding intelligently and continuously to the cognitive needs of individual learners in real time.

REFERENCES

[1] J. Sweller, "Cognitive load during problem solving: Effects on learning," Cognitive Science, vol. 12, no. 2, pp. 257–285, 1988.

[2] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," Advances in Psychology, vol. 52, pp. 139–183, 1988.

[3] T. Soukupová and J. Čech, "Real-time eye blink detection using facial landmarks," in Proc. 21st Computer Vision Winter Workshop (CVWW), Rimske Toplice, Slovenia, Feb. 2016, pp. 1–8.
[4] F. Dehais, M. Causse, F. Vachon, N. Régis, E. Menant and S. Tremblay, "Monitoring pilot's neurophysiological state in real flight conditions," IEEE Trans. Cogn. Develop. Syst., vol. 11, no. 1, pp. 8–19, Mar. 2019.
[5] T. O. Zander and C. Kothe, "Towards passive brain–computer interfaces: Applying brain–computer interface technology to human systems," J. Neural Eng., vol. 8, no. 2, p. 025005, Apr. 2011.
[6] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," IEEE Trans. Affect. Comput., vol. 1, no. 1, pp. 18–37, Jan.–Jun. 2010.
[7] R. E. Kalman, "A new approach to linear filtering and prediction problems," J. Basic Eng., vol. 82, no. 1, pp. 35–45, Mar. 1960.
[8] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," Math. Comput., vol. 19, no. 90, pp. 297–301, Apr. 1965.
[9] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.
[10] S. H. Fairclough, "Fundamentals of physiological computing," Interact. Comput., vol. 21, no. 1–2, pp. 133–145, Jan. 2009.
[11] F. Paas, J. E. Tuovinen, H. Tabbers and P. W. M. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," Educ. Psychol., vol. 38, no. 1, pp. 63–71, 2003.
[12] V. Vapnik, The Nature of Statistical Learning Theory. New York, NY, USA: Springer, 1995.
[13] G. Welch and G. Bishop, "An introduction to the Kalman filter," Univ. North Carolina at Chapel Hill, Chapel Hill, NC, USA, Tech. Rep. TR 95-041, 2006.
[14] A. Bulling, U. Blanke and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," ACM Comput. Surv., vol. 46, no. 3, pp. 1–33, Jan. 2014.
[15] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran and M. Grundmann, "BlazeFace: Sub-millisecond neural face detection on mobile hardware," in Proc. CVPR Workshop Comput. Vis. Augmented Virtual Reality, Long Beach, CA, USA, Jun. 2019.
[16] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, Oct. 2011.