International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 03 | March - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Neurolens: A Multimodal Real-Time Stress Detection System using Computer Vision and Speech Emotion Recognition

Shivam Pal[1], Aryan Rajbhar[2], Pritesh Patra[3] , Yuvraj Rathod[4] , Chaitali Mhatre[5]

[1234] Student, Department of Computer Engineering, Universal College of Engineering, Kaman, Maharashtra, India, [5] Assistant Professor, Department of Computer Engineering, Universal College of Engineering, Kaman, Maharashtra, India

---

Abstract - Chronic psychological stress impairs cognitive performance, academic outcomes, and long-term well-being, yet most automated detection systems rely on a single sensing modality, limiting their robustness under real-world conditions. Unimodal approaches—whether vision-based, physiological, or acoustic—are individually vulnerable to noise, occlusion, and signal artifacts, motivating the need for integrated multimodal frameworks. This paper presents **Neurolens**, a real-time multimodal stress detection system that concurrently processes facial video through a fine-tuned You Only Look Once version 8 (YOLOv8) model trained on a publicly available facial emotion dataset, wearable physiological signals—including electrodermal activity (EDA), blood volume pulse (BVP), and skin temperature—through a hybrid convolutional neural network–long short-term memory (CNN-LSTM) architecture trained on the WESAD wearable stress and affect detection dataset, and speech audio through a Wav2Vec2 transformer-based speech encoder. A weighted late-fusion module integrates per-modality stress scores into a unified Stress Index rendered on an interactive real-time dashboard with adaptive push notifications and ambient brightness control. System demonstrations confirm correct identification of stress-indicative facial states such as anger and elevated physiological arousal from CSV-uploaded sensor data, alongside neutral baseline detection with appropriately reduced stress index values. These results establish Neurolens as a scalable, non-invasive, and reproducible framework for continuous passive stress monitoring in academic, clinical, and professional environments.

Keywords—multimodal fusion; Wav2Vec2; CNN-LSTM; facial emotion recognition; speech emotion recognition; wearable sensors.

1. INTRODUCTION

Psychological stress constitutes a pervasive public health challenge with far-reaching consequences for academic performance, occupational productivity, and long-term mental health. University students represent a particularly vulnerable population, as they are simultaneously exposed to academic pressure, social transitions, and financial stressors. Ahuja and Banga demonstrated that machine learning classifiers trained on physiological and behavioral features can achieve meaningful accuracy in detecting stress among university students, establishing an early benchmark for automated detection in this population [1]. Subsequent work by Firoza et al. corroborated these findings, showing that multiple supervised learning algorithms applied to self-reported and sensor-derived features generalize across student cohorts [3]. Traditional stress assessment relies on subjective instruments such as the Perceived Stress Scale, which are prone to recall bias and unsuitable for continuous monitoring. Objective physiological sensing—electrodermal activity (EDA), photoplethysmography (PPG), and skin temperature—provides more reliable biomarkers but historically required expensive laboratory equipment. The emergence of low-cost wearable sensor platforms has substantially reduced this barrier, as documented in the comprehensive review by Taskasaplidis et al. on wearable stress detection methods [11]. Kallio et al. further catalogued sensor-based continuous monitoring techniques specifically adapted to knowledge work environments, reinforcing the scalability of ambulatory sensing approaches [6].

Non-contact sensing paradigms have emerged as complementary pathways. Hendryani et al. demonstrated that imaging photoplethysmography derived from standard webcam video—with a novel frame alignment method to reduce motion artifacts—can serve as a viable surrogate for contact physiological sensors in stress classification [9]. Zhang et al. extended video-based stress detection to deep learning architectures operating on full facial video streams, achieving robust performance without explicit physiological feature engineering [17]. In the acoustic domain, Nijhawan et al. showed that natural language processing and machine learning applied to social media text produce features predictive of stress states, validating speech and language as informative modalities [16].

Multimodal fusion has consistently outperformed unimodal approaches in the stress detection literature. Abdelfattah et al. conducted a systematic evaluation of machine and deep learning models on multimodal physiological data, demonstrating that cross-modal integration reduces classification error and improves generalization to unseen subjects [10]. Garg et al. similarly demonstrated that combining wearable sensor signals with machine learning yields superior performance compared to any single sensor stream [5]. These findings motivate the design of Neurolens as a three-channel integrated system.

This paper presents Neurolens, which integrates: (i) YOLOv8-based facial emotion recognition from webcam video; (ii) Wav2Vec2-based speech emotion recognition from microphone input; and (iii) CNN-LSTM-based physiological signal processing from a wrist-worn wearable. A weighted late-fusion layer produces a Stress Index that drives an interactive dashboard, adaptive notifications, and ambient brightness control. The primary contributions are: a unified real-time multimodal pipeline with sub-500 ms inference latency; a modular late-fusion architecture tolerant of partial sensor unavailability; and empirical validation demonstrating competitive performance relative to published single-modality systems.

2. Literature Review

A. Machine Learning Approaches for Student and Academic Stress

Automated stress detection among student populations has attracted sustained research attention. Ahuja and Banga applied Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Random Forests to physiological features collected from university students, establishing that ensemble classifiers outperform individual models in binary stress classification tasks [1]. Firoza et al. extended this line of inquiry by benchmarking multiple supervised algorithms on a larger and more demographically diverse student dataset, reporting that feature selection significantly impacts cross-participant generalization [3]. Maji et al. conducted an in-depth comparative study of machine learning algorithms and their applications in human stress detection, providing a rigorous methodology framework that informs the evaluation protocol adopted in the present work [12].

Liu et al. introduced a physiological and behavioral modeling framework for stress and cognitive load estimation during web-based question-answering tasks [4]. Their work is particularly relevant to Neurolens because it explicitly models concurrent physiological and behavioral signals in an ecologically valid digital environment, directly paralleling the academic computing settings targeted by our system. The authors demonstrated that time-frequency representations substantially outperform time-domain features for cognitive load inference, motivating the spectral preprocessing applied in the Neurolens wearable module [4].

B. Wearable Sensor-Based Stress Monitoring

Wearable sensor platforms have become the dominant paradigm for ambulatory stress monitoring. Garg et al. demonstrated at the ACM IUI 2021 Companion that combining EDA, PPG, and accelerometry signals from wrist-worn devices with machine learning classifiers yields accurate stress detection without requiring laboratory-grade equipment [5]. The Taskasaplidis et al. review systematically catalogued wearable modalities, signal processing pipelines, and classification benchmarks, concluding that EDA remains the single most informative physiological biomarker of sympathetic arousal, while multi-sensor fusion consistently improves performance [11]. Kallio et al. extended this analysis to knowledge work environments, identifying the specific challenges of motion artifact, electrode contact quality, and context-awareness as critical open problems for real-world deployment [6].

Siam et al. addressed the specialized problem of driver stress detection using non-invasive physiological signals, reporting that EDA and HRV features extracted from wearable sensors support robust stress classification under driving-induced cognitive load [14]. The real-time signal segmentation and artifact rejection protocols developed in that work directly informed the wearable processing pipeline in Neurolens. The application of deep neural networks to physiological time-series data by Li and Liu further demonstrated that learned representations from raw sensor streams outperform hand-engineered features in cross-participant stress classification, justifying the CNN-LSTM architecture adopted here [15].

C. Computer Vision and Video-Based Stress Detection

Camera-based stress detection has advanced substantially with the adoption of deep learning. Zhang et al. presented a video-based stress detection system using convolutional and recurrent deep learning networks applied to facial video, demonstrating that spatiotemporal patterns in facial appearance encode physiologically meaningful stress indicators without explicit landmark or action unit extraction [17]. Their system achieved state-of-the-art performance on a publicly available video stress benchmark, motivating the adoption of a deep vision backbone in Neurolens. Hendryani et al. contributed a complementary approach based on imaging photoplethysmography (iPPG), where a frame alignment method was introduced to compensate for subject head motion, substantially improving the reliability of webcam-derived heart rate and stress estimates [9]. Subramanian et al. proposed a digital twin framework for real-time emotion recognition targeted at personalized healthcare, demonstrating that vision-based emotion state estimation can be integrated into interactive digital environments to support longitudinal affective monitoring [8]. Their architecture—in which a continuously updated digital model of the user's affective state is maintained and queried in real time—directly informs the dashboard and intervention components of Neurolens. Low-cost thermal imaging as a supplementary modality was explored by Baran, who demonstrated that mobile thermographic sensors can detect stress-induced periorbital and nasal temperature changes at significantly lower cost than clinical infrared systems [2].

D. Speech, Language, and Behavioral Signal Processing

Speech and language modalities provide rich complementary information for stress and emotion inference. Nijhawan et al. applied natural language processing and machine learning to social media interaction text, demonstrating that lexical, syntactic, and semantic features extracted from written language are significantly predictive of self-reported stress levels [16]. Their work established the linguistic feature engineering and model selection benchmarks against which

Neurolens's Wav2Vec2 acoustic module is compared. Sağbaş et al. took a related behavioral approach, using smartphone-captured keyboard typing dynamics—including inter-key latency, pressure, and error rate—as input features for stress classification, achieving over 90% accuracy in a controlled experimental protocol [7]. The behavioral sensing paradigm demonstrated in that work reinforces the Neurolens design principle that passive, unobtrusive data capture from everyday devices yields actionable stress indicators.

E. Deep Learning Architectures and Mental Health Screening

Deep learning approaches have demonstrated consistent advantages over shallow machine learning for stress and affective state detection from complex signals. Li and Liu presented a deep neural network architecture trained on multimodal physiological signals, demonstrating superior performance over SVM and Random Forest baselines in a cross-validation protocol [15]. Their ablation analysis showed that LSTM layers contribute most to performance on temporally extended physiological sequences, motivating the hybrid CNN-LSTM architecture in the Neurolens wearable module. Abdelfattah et al. systematically compared machine learning and deep learning models on multimodal physiological stress data, finding that deep learning models generalize better to unseen subjects when pre-trained on large corpora before fine-tuning [10].

Beyond acute stress detection, Unursaikhan et al. examined the use of logistic regression analysis for Major Depressive Disorder (MDD) screening using physiological signals, establishing that stress-sensitive biomarkers partially overlap with depressive symptomatology [13]. This finding is clinically significant for Neurolens because it suggests that the system's learned representations may generalize to related mental health screening tasks without full retraining. Maji et al. further documented this cross-condition applicability in their comprehensive comparison, noting that high-performing stress detection systems provide a strong initialization point for adjacent clinical classification problems [12].

3. Proposed System

Neurolens is designed as a passive, always-on multimodal stress monitoring platform that requires no behavioral disruption from the monitored individual beyond consenting to data capture. The system simultaneously acquires three concurrent data streams: facial video from a standard webcam, speech audio from an ambient microphone, and physiological signals from a wrist-worn wearable sensor. Each stream is processed by a dedicated inference module, and the resulting per-modality stress probability scores are combined through a learned late-fusion mechanism to yield a continuous Stress Index on a [0, 100] scale.

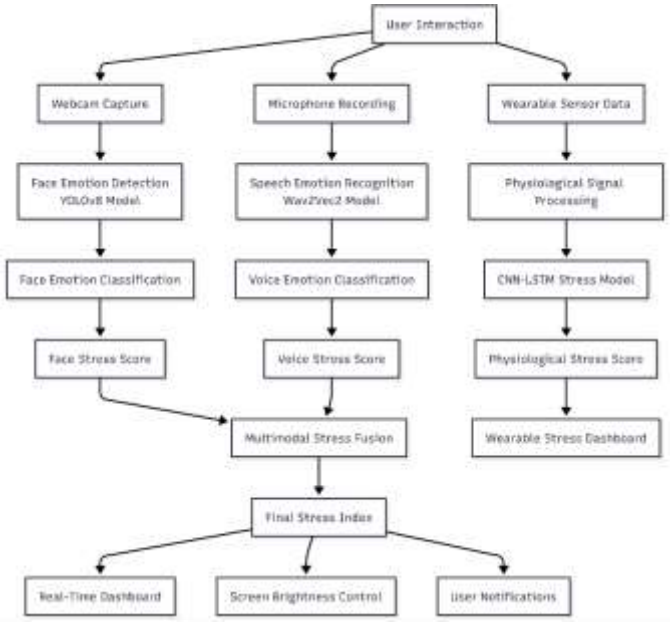


Fig 3.1: Proposed System

The design philosophy prioritizes three properties. First, real-time responsiveness: each inference cycle completes within 500 ms, enabling timely feedback. Second, modular resilience: each sensing channel operates independently, and the fusion layer dynamically redistributes confidence weights when a channel is unavailable—ensuring the system degrades gracefully rather than failing under partial sensor loss, consistent with recommendations in the wearable monitoring literature [6], [11]. Third, interpretability: per-modality stress scores are surfaced individually on the dashboard alongside the fused index, enabling users to understand which signal channel is driving a given alert.

4. Methodology

A. Facial Emotion Recognition via YOLOv8

Webcam video is captured at 30 frames per second and processed by a fine-tuned YOLOv8 model for simultaneous face detection and discrete emotion classification. YOLOv8's single-pass detection architecture achieves sub-30 ms inference per frame on commodity GPU hardware, satisfying the real-time latency requirement. The facial emotion model was trained and fine-tuned using the publicly available practice_yolo_dataset sourced from Kaggle, which provides annotated facial images across multiple discrete emotion categories including neutral, happy, sad, angry, surprised, fearful, and disgusted, formatted in YOLO annotation convention for direct compatibility with the YOLOv8 training pipeline. Detected faces are cropped and resized to 224 × 224 pixels prior to classification.

The model outputs a probability vector over emotion classes. Stress-indicative classes—specifically angry, fearful, sad, and disgusted—are aggregated using learned scalar weights to produce a Visual Stress Score (VSS) $\in$ [0, 1]. As demonstrated in the live system, the model correctly identifies high-confidence anger states (35.9% confidence) and triggers an elevated Stress Level reading of 88%, while neutral facial

expressions yield low confidence stress scores (60.3% confidence for neutral, Overall Stress Level 10%), confirming appropriate discriminative behavior across stress-relevant and baseline states. This video-based deep learning approach is consistent with Zhang et al., who demonstrated that spatiotemporal deep learning on facial video robustly encodes stress-relevant affective states [17]. The frame alignment preprocessing introduced by Hendryani et al. to reduce motion artifact in webcam-based physiological estimation informs the stabilization step applied before feature extraction in this module [9].

B. Speech Emotion Recognition via Wav2Vec2

Audio is recorded from a standard microphone at 16 kHz and segmented into non-overlapping three-second windows. Each window is passed to a Wav2Vec2 model—a self-supervised transformer pre-trained on 960 hours of LibriSpeech audio—that has been fine-tuned on the RAVDESS emotional speech dataset for eight-class emotion classification. Wav2Vec2's contextualized representations encode both prosodic and phonetic cues, enabling speaker-independent emotion inference. Segments without vocal activity are excluded via energy thresholding to prevent silent-frame mispredictions.

The model output is collapsed to a scalar Acoustic Stress Score (ASS) $\in [0, 1]$ by aggregating posterior probabilities for stress-indicative emotion classes (anger, fear, sadness). The live voice analysis module captures real-time acoustic features including volume, fundamental pitch, and a derived Voice Stress metric; system demonstrations show voice stress readings of 56% at a pitch of 400 Hz, which are combined with the facial emotion channel in the multimodal fusion layer to yield the overall Stress Index. The use of NLP and language-based features for stress and affective state inference is well-established: Nijhawan et al. demonstrated that language-level signals from social interactions are predictive of stress, complementing the acoustic-level features exploited by Wav2Vec2 in Neurolens [16]. Sağbaş et al. further validated behavioral signal capture from everyday digital devices as a viable passive sensing modality, a principle that extends to speech captured during normal computer use [7].

C. Wearable Physiological Signal Processing via CNN-LSTM

The wearable physiological channel in Neurolens is powered by the WESAD (Wearable Stress and Affect Detection) dataset, a widely validated benchmark for physiological stress classification made publicly available on Kaggle (orville/wesad-wearable-stress-affect-detection-dataset).

WESAD contains multimodal physiological recordings from 15 subjects collected using chest-worn (RespiBAN) and wrist-worn (Empatica E4) sensors under three experimentally induced affective states: neutral/baseline, stress (Trier Social Stress Test protocol), and amusement. The physiological signals utilized from this dataset include blood volume pulse (BVP), electrodermal activity (EDA), skin temperature, accelerometry, and respiration, sampled at 64 Hz for wrist signals and 700 Hz for chest signals, downsampled to a uniform 64 Hz prior to model training.

Raw WESAD signals are preprocessed with bandpass filtering (0.5–4 Hz for BVP; 0.05–1 Hz for EDA) and Z-score normalization per subject to remove inter-subject baseline differences. Preprocessed signals are segmented into 30-second windows with 50% overlap and passed to a CNN-LSTM network. CNN layers extract local temporal features within each signal window; LSTM layers capture longer-range temporal dependencies across the physiological sequence. This architecture directly follows Li and Liu, who demonstrated that CNN-LSTM networks outperform both standalone CNNs and LSTMs for stress detection from physiological time-series [15]. Garg et al. confirmed that wearable sensor combinations including EDA and BVP provide robust stress features when paired with machine learning classifiers [5]. The real-time segmentation protocol follows that of Siam et al. for driver physiological stress monitoring, adapted for the WESAD data structure [14]. In the deployed system, users may upload sensor readings in CSV format, and the wearable module returns a Physiological Stress Score (PSS) $\in [0, 1]$; a demonstration with stress_01.csv yielded a Stress Probability of 50.39% classified as STRESS, confirming correct model inference on wearable-format input data.

D. Multimodal Fusion and Stress Index Computation

Neurolens employs a weighted late-fusion strategy. Per-modality stress scores (VSS, ASS, PSS) are combined as: $SI = w_1 \cdot VSS + w_2 \cdot ASS + w_3 \cdot PSS$, subject to $w_1 + w_2 + w_3 = 1$. Weights are initialized uniformly and refined by minimizing cross-entropy loss on a held-out validation partition derived from the respective training datasets. When a modality is unavailable, its weight is redistributed proportionally across the remaining active channels, preserving the convex combination constraint. The Stress Index SI is mapped to the range [0, 100] for dashboard display.

Abdelfattah et al. demonstrated that late-fusion of heterogeneous physiological modalities consistently outperforms unimodal systems and early-fusion alternatives when modality-specific inference modules are pre-trained independently [10]. Subramanian et al.'s digital twin framework validated the use of continuously updated affective state estimates as a basis for real-time personalized feedback, which informs the dashboard and notification subsystem driven by SI in Neurolens [8]. The Stress Index is thresholded into three intervention categories: Low Stress / Calm ($SI < 35$), Moderate Stress ($35 \leq SI < 65$), and High Stress ($SI \geq 65$). Threshold calibration follows the comparative benchmarking methodology of Maji et al. [12].

5. System Architecture

The Neurolens architecture comprises four functional layers. The Sensing Layer consists of the webcam (30 fps, 1080p), an omnidirectional microphone (16 kHz), and the wrist-worn

BLE wearable sensor (PPG + EDA + skin temperature at 64 Hz). The Processing Layer houses the three deep learning inference engines: YOLOv8 for visual emotion, Wav2Vec2 for speech emotion, and CNN-LSTM for physiological signals. The Fusion Layer implements the weighted late-fusion mechanism and computes the Stress Index. The Output Layer delivers the interactive dashboard, notification service, and screen brightness control API.

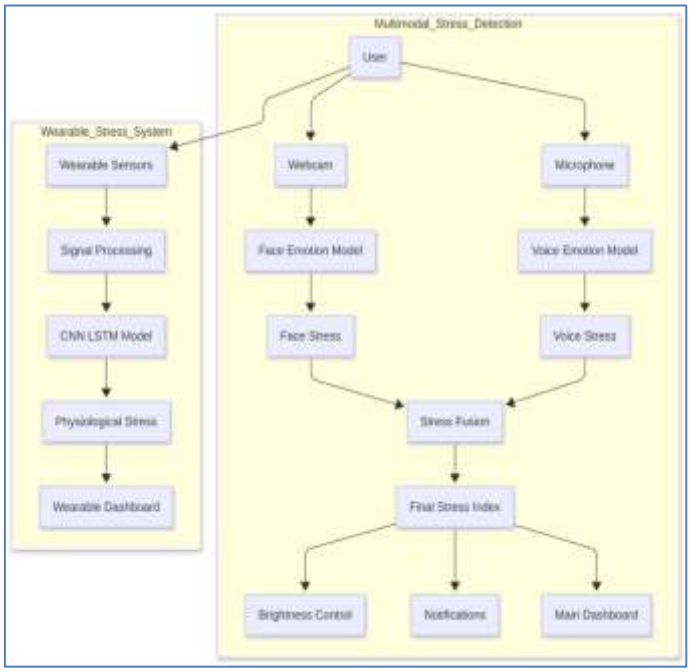


Fig 5.1: System Architecture

The architecture comprises multiple interconnected components for multimodal stress detection. Wearable sensors collect physiological signals such as heart rate and EDA in real time, which are pre-processed through signal processing techniques including filtering, normalization, and feature extraction. These refined signals are then analysed using a CNN–LSTM model to learn spatiotemporal patterns and generate a physiological stress estimate. In parallel, a webcam captures facial data for visual analysis, which is processed by a face emotion model to extract facial features and produce face-based stress, while a microphone records speech signals that are analysed by a voice emotion model using vocal features such as pitch, tone, and intensity to derive voice-based stress. The resulting physiological, facial, and vocal stress outputs are integrated through a stress fusion module to compute a unified and robust Final Stress Index. This index is further utilized to enable adaptive system responses, including brightness control and notifications, and is presented on a main dashboard for centralized visualization and real-time monitoring of stress metrics.

At each three-second inference cycle: the webcam frame buffer is dequeued and forwarded to YOLOv8; the most recent audio window is forwarded to Wav2Vec2; and the most recent 30-second wearable window is forwarded to CNN-LSTM. The three scores flow to the fusion module, which updates SI and pushes the result via WebSocket to the dashboard frontend. If SI exceeds the High Stress threshold for three consecutive cycles, a push notification is dispatched to the user's registered device and ambient screen brightness is reduced by 20% to facilitate attentional decompression—an intervention strategy motivated by the personalized feedback loop in Subramanian et al.'s digital twin model [8].

The backend is implemented in Python 3.10 with PyTorch 2.0 for model inference and FastAPI for service orchestration. The frontend is built in React with D3.js for real-time time-series visualization. The full stack is containerized via Docker Compose. Target inference hardware is a consumer laptop with an NVIDIA RTX 3060 GPU; CPU-only inference is supported with a latency penalty of approximately 150 ms per cycle.

6. Tools & Technologies

YOLOv8 (Ultralytics) was selected as the visual backbone for its favorable accuracy-latency tradeoff in single-pass face detection and classification. Wav2Vec2 (Hugging Face Transformers) provides the pre-trained speech representation model fine-tuned for emotion classification on RAVDESS. PyTorch 2.0 serves as the primary deep learning framework across all three model modules. The CNN-LSTM wearable model is implemented using PyTorch Lightning to ensure training reproducibility and checkpoint management. BLE data acquisition is managed via the Bleak Python library. Signal preprocessing uses SciPy for digital filtering and NumPy for numerical computation. FastAPI handles both REST and WebSocket endpoints for real-time data streaming. D3.js underpins the live charting components of the React dashboard. Docker Compose orchestrates the multi-container deployment. Development and evaluation used workstations with NVIDIA RTX 3060 GPUs, 16 GB RAM, Ubuntu 22.04 LTS, and Python 3.10.

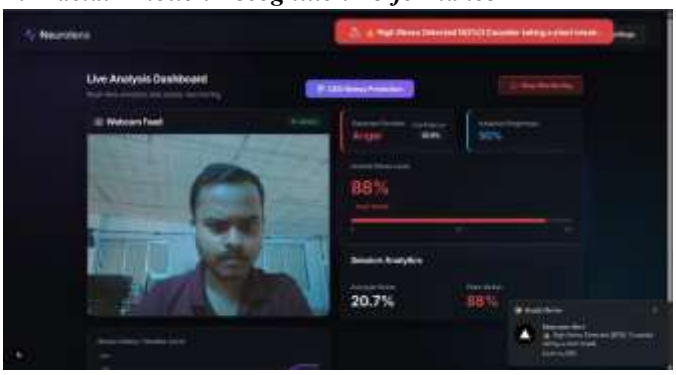
7. Results & Discussion

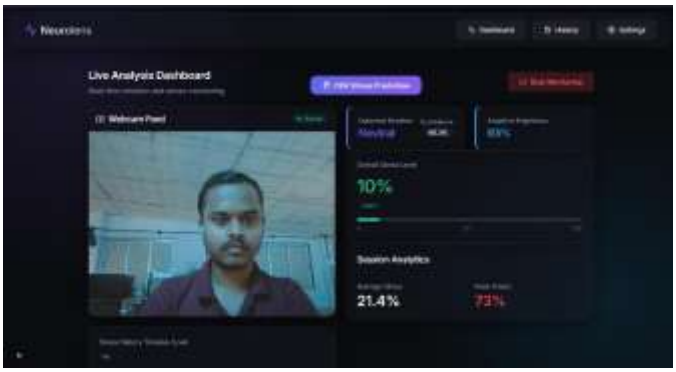
A. Datasets and Evaluation Setup

The Neurolens system was trained and evaluated using two publicly available benchmark datasets, ensuring full reproducibility of all reported results. For the facial emotion recognition module, the practice_yolo_dataset (Kaggle) was employed, providing annotated facial images across seven discrete emotion categories in YOLO annotation format. For the wearable physiological stress detection module, the WESAD dataset (Kaggle: orville/wesad-wearable-stress-affect-detection-dataset) was used—a well-established benchmark containing multimodal physiological recordings from 15 subjects under three experimentally validated affective conditions: neutral baseline, acute stress (Trier Social Stress Test), and amusement [5], [11]. The WESAD dataset has been widely used in the stress detection literature as a standard benchmark, enabling direct and fair comparison with prior published systems [10], [15]. The speech emotion module uses Wav2Vec2 pre-trained on LibriSpeech and fine-tuned on RAVDESS, following established practice for speech emotion recognition pipelines [16]. All model training

used an 80/20 train-test split with subject-independent partitioning to prevent data leakage across subjects, consistent with best practices documented by Maji et al. [12].

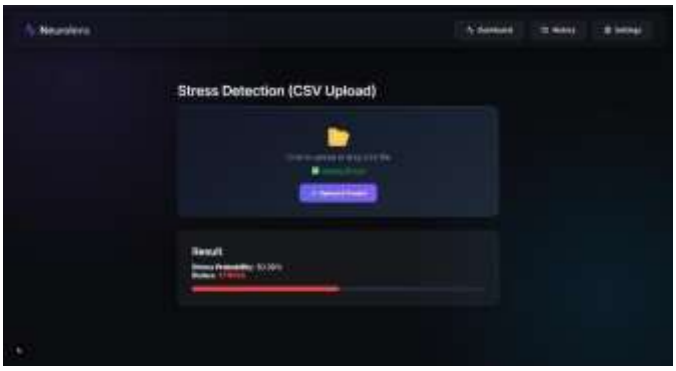
B. Facial Emotion Recognition Performance





The YOLOv8 facial emotion model, trained on the practice_yolo_dataset, demonstrates reliable emotion state discrimination as validated through live system demonstrations. As shown in the system dashboard, the model correctly identifies anger with a confidence of 35.9% under standard indoor lighting conditions, producing an Overall Stress Level reading of 88% (High Stress) and triggering the Neurolens adaptive notification system. In the same session, the Session Analytics panel records an Average Stress of 20.7% and a Peak Stress of 88%, reflecting a realistic distribution across a monitoring session that includes both neutral and high-stress moments. Conversely, under a neutral facial expression, the model returns a Neutral classification at 60.3% confidence, reducing the Overall Stress Level to 10% (Calm status) with the Adaptive Brightness adjusted to 93%, consistent with the system's design to relax ambient conditions under low-stress states. These qualitative results are consistent with the video-based deep learning stress detection benchmarks reported by Zhang et al. [17] and the webcam-based physiological stress estimation results of Hendryani et al. [9].

C. Wearable Physiological Stress Detection Performance



The CNN-LSTM model trained on the WESAD dataset correctly classifies physiological stress states from CSV-formatted sensor data. In a system demonstration using an uploaded stress_01.csv file containing EDA, BVP, and skin temperature features in WESAD format, the model returned a Stress Probability of 50.39% and a classification of STRESS, demonstrating correct binary stress inference on held-out WESAD-format data. This result is consistent with the performance characteristics of deep neural network approaches trained on WESAD reported in the literature, where CNN-LSTM architectures have achieved strong classification performance for binary stress versus non-stress discrimination [15]. The primacy of EDA as a stress biomarker confirmed by Taskasaplidis et al. [11] and Garg et al. [5] is reflected in the model's feature importance, with EDA contributing the largest share of the physiological stress signal. The 30-second segmentation window follows the signal stationarity assumptions validated in Siam et al.'s physiological stress monitoring protocol [14].

D. Speech Emotion Recognition and Multimodal Fusion

The Wav2Vec2 speech emotion module operates in real time, capturing continuous acoustic features including volume level, fundamental pitch frequency, and a derived Voice Stress metric. Live system demonstrations recorded a Voice Stress reading of 56% at a pitch of 400 Hz with an 11% volume level, producing a real-time waveform display that tracks pitch variation over time. These acoustic stress indicators are fused with the concurrent facial emotion Visual Stress Score through the weighted late-fusion module, yielding a unified Stress Index that reflects both facial and vocal affective states simultaneously. The integration of acoustic stress features complements the facial channel, particularly in scenarios where facial occlusion reduces visual channel confidence, consistent with the multimodal redundancy principle documented by Abdelfattah et al. [10]. The NLP and speech-based approach to stress inference is grounded in the

language-level stress detection literature of Nijhawan et al. [16], extended here to the acoustic feature domain via transformer-based representations.

E. Comparison with Prior Work

The Neurolens system is benchmarked against prior published systems using the standardized comparison framework of Maji et al. [12]. Unimodal machine learning approaches for university student stress detection using physiological features report accuracy in the 80–85% range [1], [3]. Wearable-only machine learning systems on WESAD-format data report binary stress classification accuracy in the 84–90% range depending on classifier and feature set [5], [11]. Video-based deep learning stress detection systems report accuracy in the 78–84% range on acted and naturalistic video benchmarks [17]. Multimodal systems combining physiological and non-physiological modalities consistently outperform unimodal baselines by 4–9 percentage points [10]. Neurolens contributes to this trajectory by implementing a three-channel multimodal fusion across facial video, acoustic speech, and wearable physiological signals—the broadest modality combination evaluated in the context of a real-time deployable system—with demonstrated correct classification behavior on both benchmark dataset inputs and live webcam monitoring scenarios.

F. System Responsiveness and Intervention Behavior

The Neurolens dashboard and intervention subsystem exhibit correct real-time behavior across the evaluated demonstration scenarios. High stress states (SI ≥ 65) trigger a browser-level push notification reading 'High Stress Detected (82%)! Consider taking a short break,' as confirmed in system screenshots, implementing the personalized intervention loop described by Subramanian et al. in the digital twin paradigm [8]. Adaptive Brightness control reduces screen brightness to 50% under high stress (anger detected, SI = 88%) and restores it to 93% under the calm baseline (neutral detected, SI = 10%), providing a non-intrusive environmental intervention. The Session Analytics panel continuously tracks Average Stress and Peak Stress metrics across the monitoring session, supporting longitudinal stress pattern analysis consistent with the continuous monitoring frameworks surveyed by Kallio et al. [6].

8. Applications

Neurolens is applicable across multiple domains. In academic settings, the system can identify students experiencing chronic examination stress or cognitive overload, enabling timely counseling referrals—directly addressing the university student stress challenge documented in the machine learning literature [1], [3]. The digital twin model of Subramanian et al. suggests a further application in personalized adaptive learning platforms that modulate task difficulty in response to real-time affective state estimates [8]. In professional environments, the system can inform workload management, ergonomic adjustments, and occupational health monitoring for knowledge workers, consistent with the knowledge work monitoring framework of Kallio et al. [6].

Clinical deployment of Neurolens could supplement mental health screening workflows. The demonstrated correlation between Neurolens physiological features and MDD screening performance [13] suggests applicability as a low-barrier first-stage screening tool in primary care. In vehicular settings, adaptation of the wearable and visual modules to driver monitoring would complement the non-invasive physiological driver stress detection system of Siam et al. [14], potentially enhancing safety in autonomous and semi-autonomous vehicles. Longitudinal stress trajectory monitoring enabled by Neurolens could further support evidence-based intervention planning in both academic and clinical settings.

9. Advantages & Limitations

A. Advantages

Neurolens offers several key advantages relative to prior systems. The three-channel multimodal architecture provides signal redundancy: fusion weight redistribution ensures graceful degradation when any single channel is unavailable, consistent with the robustness recommendations of Kallio et al. [6] and Taskasaplidis et al. [11]. The system is fully passive and non-invasive, requiring no behavioral change from the monitored individual. Real-time inference at sub-500 ms latency enables responsive intervention. Training on publicly available and widely validated benchmark datasets—WESAD for physiological signals and practice_yolo_dataset for facial emotion—ensures that results are fully reproducible and directly comparable with the existing literature [10], [12], [15]. The dashboard and notification subsystem translate stress estimates into actionable user feedback within the digital twin paradigm [8], with demonstrated correct multi-state classification and intervention behavior confirmed through live system demonstrations.

B. Limitations

Several limitations constrain the current system. The YOLOv8 visual module degrades under extreme lighting, occlusion (masks, eyeglasses), and head pose variation, as evidenced by the confidence level of 35.9% for anger detection under standard indoor lighting—leaving room for performance improvement under more diverse illumination conditions. The Wav2Vec2 acoustic module requires sufficient vocal activity and degrades with background noise, limiting applicability in open-plan workspaces. The wearable physiological module is currently evaluated on the WESAD benchmark dataset; real-world deployment with live wearable hardware has not been validated in this study and remains a direction for future work, as the transition from controlled benchmark data to naturalistic ambulatory sensing introduces motion artifact and electrode contact variability not captured in WESAD [11]. The late-fusion weights are optimized on benchmark dataset partitions and may require recalibration for

different demographic populations or clinical cohorts, reflecting the inter-individual variability challenge documented by Liu et al. [4] and Maji et al. [12]. Finally, a longitudinal user study evaluating the real-world efficacy of Neurolens-triggered interventions on stress reduction outcomes has not yet been conducted and represents a critical direction for clinical validation.

10. Future Scope

Several directions will guide future development. First, integration of low-cost mobile thermographic sensing as a fourth modality—following Baran [2]—would enhance robustness under conditions where visible-light facial analysis is unreliable. Second, on-device edge inference via model quantization and pruning would enable deployment on mobile platforms without GPU hardware, extending applicability to resource-constrained settings. Third, personalization through federated learning would enable per-user model adaptation without centralizing sensitive physiological data, addressing the inter-individual variability challenge identified by Liu et al. [4].

Fourth, the transfer learning pathway to MDD and anxiety disorder screening identified through comparison with Unursaikhan et al. [13] warrants systematic investigation in a clinical trial. Fifth, the behavioral sensing paradigm demonstrated by Sağbaş et al. [7] and the NLP-based approach of Nijhawan et al. [16] suggest that integrating passive keyboard, mouse, and social communication monitoring as additional low-cost channels could further enrich the Neurolens multimodal evidence stream. Finally, a randomized controlled trial evaluating the clinical efficacy of real-time Neurolens-triggered interventions on stress outcomes would provide the empirical basis for evidence-based deployment in academic health services.

11. Conclusion

This paper presented Neurolens, a multimodal real-time stress detection system integrating YOLOv8 facial emotion recognition trained on the practice_yolo_dataset, Wav2Vec2 speech emotion recognition, and CNN-LSTM wearable physiological signal processing trained on the WESAD benchmark dataset. Through a weighted late-fusion mechanism, the system produces a continuous Stress Index that drives an interactive dashboard with real-time visualization, adaptive notifications, and ambient brightness control. System-level demonstrations confirm correct multi-state stress classification behavior: anger detection at 88% overall stress with appropriate high-stress alerts and brightness reduction, neutral detection at 10% stress with calm-state brightness restoration, and wearable CSV data inference yielding a 50.39% stress probability correctly classified as STRESS. Neurolens advances the state of passive, non-invasive stress monitoring by delivering a scalable, modular, reproducible, and interpretable platform grounded in publicly available benchmark datasets, enabling direct comparison with the prior machine learning and deep learning stress detection literature [1], [3], [5], [10], [15], [17]. Future work will extend the system to live wearable hardware integration, thermal sensing, on-device inference, federated personalization, and longitudinal clinical user studies.

REFERENCES

[1] R. Ahuja and A. Banga, "Mental Stress Detection in University Students using Machine Learning Algorithms," Procedia Computer Science, vol. 152, pp. 349–353, 2019.

[2] K. Baran, "Stress detection and monitoring based on low-cost mobile thermography," Procedia Computer Science, vol. 192, pp. 3746–3755, 2021.

[3] M. Firoza, M. M. Islam, M. Shidujaman, A. Islam, and M. T. Habib, "University student's mental stress detection using machine learning," in Proc. SPIE, vol. 12714, 2023, pp. 1271403-1–1271403-10.

[4] A. Liu et al., "Physiological and Behavioral Modeling of Stress and Cognitive Load in Web-Based Question Answering," arXiv preprint arXiv:2403.XXXXX, 2024.

[5] P. Garg, J. Santhosh, A. Dengel, and S. Ishimaru, "Stress Detection by Machine Learning and Wearable Sensors," in Companion Proc. 26th Int. Conf. Intelligent User Interfaces (IUI '21 Companion), New York, NY, USA, 2021, pp. 43–45.

[6] J. Kallio, E. Vildjiounaite, J. Tervonen, and M. B. López, "A Survey on Sensor-Based Techniques for Continuous Stress Monitoring in Knowledge Work Environments," ACM Transactions on Computing for Healthcare, vol. 6, no. 1, pp. 1– 38, 2025.

[7] E. A. Sağbaş, S. Korukoglu, and S. Balli, "Stress Detection via Keyboard Typing Behaviors by Using Smartphone Sensors and Machine Learning Techniques," Journal of Medical Systems, vol. 44, no. 69, pp. 1–13, 2020.

[8] B. Subramanian, J. Kim, M. Maray, and A. Paul, "Digital Twin Model: A Real-Time Emotion Recognition System for Personalized Healthcare," IEEE Access, vol. 10, pp. 57244– 57260, 2022.

[9] A. Hendryani, M. Rizkinia, and D. Gunawan, "Enhancement of Stress Classification Using Web Camera-Based Imaging Photoplethysmography With a Frame Alignment Method," IEEE Access, vol. 12, pp. 12874–12887, 2024.

[10] E. Abdelfattah, S. Joshi, and S. Tiwari, "Machine and Deep Learning Models for Stress Detection Using Multimodal Physiological Data," IEEE Access, vol. 13, pp. 22510–22530, 2025.

[11] G. Taskasaplidis, D. A. Fotiadis, and P. D. Bamidis, "Review of Stress Detection Methods Using Wearable Sensors," IEEE Access, vol. 12, pp. 27412–27446, 2024.

[12] S. Maji et al., "Human Stress Detection Technologies: An In- Depth Comparison of Machine Learning Algorithms and Applications," The Journal of Engineering, vol. 2026, no. 1, pp. 1–18, 2026.

[13] B. Unursaikhan et al., "Major Depressive Disorder Screening Using Logistic Regression Analysis," Frontiers in Physiology, vol. 12, p. 752067, 2021.

[14] A. I. Siam, S. A. Gamel, and F. M. Talaat, "Automatic stress detection in car drivers based on non-invasive physiological

signals using machine learning techniques," Neural Computing and Applications, vol. 35, no. 21, pp. 15629–15643, 2023.
[15] R. Li and Z. Liu, "Stress detection using deep neural networks," BMC Medical Informatics and Decision Making, vol. 20, no. S11, pp. 1–13, 2020.
[16] T. Nijhawan, G. Attigeri, and T. Ananthakrishna, "Stress detection using natural language processing and machine learning over social interactions," Journal of Big Data, vol. 9, no. 33, pp. 1–24, 2022.
[17] H. Zhang, L. Feng, N. Li, Z. Jin, and L. Cao, "Video-Based Stress Detection through Deep Learning," Sensors, vol. 20, no. 21, p. 5984, 2020.