Safe Space

Armaan Mittal[1,2*], Ojasvi [2,3†], Kashish [1,2†], Akriti [1,2*], Jahnvi Varshney[1,2*], Navraj Kaur[1,2†], Mehak Garg[1,2*], Mahek Phutela[2,3†], Kishu [1,2†]

[1*], Organization, Street, City, 100190, State, Country.
[2]Department, Organization, Street, City, 10587, State, Country.
[3]Department, Organization, Street, City, 610101, State, Country.

*Corresponding author(s). E-mail(s): amittal_be22@thapar.edu; akritipaudelmail@gmail.com; jvarshney50_be23@thapar.edu; igargmehak0208@gmail.com;
Contributing authors: ojasvipathania23@gmail.com; bansalk5000@gmail.com; navrajk988@gmail.com; mahekphutela57@gmail.com; kishu06280@gmail.com;
†These authors contributed equally to this work.

Abstract

Explainability, privacy, and emotional intelligence are given top priority in this paper's real-time, edge-deployable system for multimodal stress detection and natural language feedback. Using lightweight machine learning techniques like logistic regression and gradient boosting, the system models various data sources independently, including physiological signals (heart rate variability, skin conductance, and skin temperature), behavioral patterns (keystroke dynamics and mouse movements), and facial expressions. The system produces interpretable, per-feature stress scores normalized to a range of [-1,+1] rather than a single opaque output, enabling users to determine which signals are most responsible for their stress state.A local large language model (Mistral-7B-Instruct) is then fed these normalized scores, producing succinct two-sentence feedback. The user's stress level is summed up in the first sentence using a four-point textual scale. From ”calm” to ”high stress,” while the latter suggests an adaptive coping strategy, like mental grounding or breathwork. Because there is no reliance on the cloud, all inference and feedback generation takes place locally, making the system appropriate for use in settings with limited resources or privacy concerns. The system is modular, interpretable, and efficient in providing actionable feedback, according to preliminary tests utilizing both artificial and real sensor inputs. This provides a new avenue for human-centered, explainable stress-aware technologies.

Keywords: Behavioral Biometrics, Edge Computing, Human-Centered AI, Lightweight Machine Learning, Multimodal Stress Detection, Natural Language Feedback

1 Introduction

A common occurrence in contemporary life, stress has a well-established long-term negative influence on both mental and physical health. Unmanaged stress can have major negative effects, including cardiovascular strain and cognitive exhaustion, especially if it goes unnoticed [1]. Researchers and developers have tried to create systems that can detect stress in real time as wearable sensors, webcams, and digital behavioral traces become more widely available [2]. Nevertheless, many of these systems only provide a binary classification, rely on cloud-based processing, or use intricate black-box models, providing little explanation for their predictions [3]. Current methods frequently fall short in bridging the gap between user comprehension and computational inference. Although someone may be told they are "under stress," they are rarely given an explanation, including which signals were most responsible or what they can do to address the situation. Such systems are challenging to trust because of their lack of transparency, particularly when utilized outside of clinical settings [4]. Furthermore, sending raw behavioral or physiological data to distant servers for analysis raises privacy issues. These concerns become even more important in situations like workplace well-being or mental health support. Three fundamental ideas like modularity, explainability, and local-first deployment are used in this paper to present a system that aims to overcome these constraints. Heart rate variability (HRV), skin conductance (GSR), temperature, facial action units, keystroke dynamics, and mouse movement patterns are among the various real-time signals that our system takes in. Using lightweight machine learning models like logistic regression and gradient boosting, each modality is processed separately to generate a normalized stress score between [-1, +1]. Both users and developers can determine which signals indicate elevated stress and to what degree, thanks to these scores' inherent interpretability and computational efficiency [5]. The system makes use of a local instance of an open-source large language model (Mistral-7B-Instruct) to translate these signals into useful instructions. After receiving a summary of the per-modality scores, the model produces two-sentence feedback. The first sentence expresses the user's stress level using a four-level textual scale ("calm," "neutral," "mild stress," or "high stress"), and the second provides a straightforward, context-free, and fast coping mechanism. Crucially, the entire pipeline from sensor input to language feedback operates on-device, negating the need for cloud inference or internet access. This makes it ideal for use in edge computing scenarios or privacy-sensitive environments. To capture and process biosignals, a hardware module is included as part of the system. A system designed to record stress levels based on multimodal sensor input incorporates the hardware architecture described in this paper. These inputs include skin temperature, heart rate variability, electrodermal activity, and cardiovascular data and all of which are known physiological correlates of stress. Lightweight machine learning algorithms process the data from these sensors at the software level as well, providing users with real-time feedback. The ESP32 microcontroller interfaces with a number of physiological sensors, including the MAX30100, AD8232 ECG, GSR, and DS18B20 temperature sensors, which are the main focus of the hardware implementation. Each sensor sends out a unique signal that represents the body's autonomic nervous system's reaction to stress. The DS18B20 detects subtle changes in skin temperature, which are frequently linked to vasoconstriction during stress; the GSR sensor measures skin conductivity as a gauge of emotional arousal; the AD8232 captures pure ECG waves for HRV purposes; and the MAX30100 tracks blood oxygen saturation and heart rate variability.The architecture, implementation, and assessment of the system are described in this paper. We demonstrate that the modular per-modality modeling approach improves transparency while maintaining predictive performance, and that the feedback by the LLM is both interpretable and user-relevant. According to our research, this strategy might lay the groundwork for stress-aware technologies that are more human-centered, private, and intended for practical implementation.

1.1 Motivation

Chronic stress is now an everyday aspect of modern workplace life. Surveys in recent years report that many working adults feel stressed every day and often lack effective coping techniques. Although occasional stress can motivate and energize us, when it is chronic or not managed well, it may cause burnout, decreased productivity, anxiety, and even severe health issues.Despite increasing consciousness for mental health, there is still a shortage of user-friendly, tailor-made, and understandable technology that can assist users to cope and make sense of their stress in real-time. Most available solutions rely on cloud-based infrastructure, sophisticated models, or unidimensional methods that are either not respecting the privacy of the user or not addressing the needs of an individual. This is the void that makes us start this project. We want to make a system that applies real-time, multimodal stress detection, enabling individuals to be in charge of their mental well-being. The strategy tries not just to precisely measure stress levels but also to offer explicit, understandable descriptions and customized coping strategies by combining physiological signals, behavioral signs, and emotions. In addition, stress management is made more discreet, efficient, and available even in sensitive or resource-constrained environments such as healthcare, education, or remote work by making sure the system is capable of operating on-device without requiring constant access to the internet. Ultimately, aiming to create a reliable, user-friendly tool to assist people in understanding stress and making positive changes to lead healthier, more satisfying lives

1.2 Contribution

- The project identifies gaps in existing stress detection systems, such as limited real-time feedback, lack of personalization, and poor explainability, through an extensive review of academic and industry sources. This helped define clear project goals.
- The approach includes scripts and modules to continuously read sensor inputs, apply filtering and normalization, and prepare data for machine learning models, all with low latency suitable for real-time use.
- Lightweight models such as XGBoost and logistic regression are trained to distinguish between eustress and distress, using feature engineering, hyperparameter tuning, and class balancing to achieve reliable predictions.
- Transparency is prioritized by designing interpretable scoring formulas and visualizations, helping users understand how stress scores are computed and increasing trust in the system.

1.3 Paper Organization

The rest of this paper is organized as follows. In Section 2, we dive into a comprehensive review of the latest literature on computational stress detection, emphasizing multimodal, explainable, and user-centered systems.The section3 briefly introduces the key foundational concepts relevant to our proposed stress detection system: physiological signal processing, convolutional neural networks (CNNs), and large language models (LLMs) Section 4 offers an in-depth look at the architecture of our proposed system, detailing its hardware components, machine learning models, and the feedback generation powered by LLMs. In Section 5, we share experimental results derived from simulations and synthetic data tests, evaluating detection accuracy, the contribution of each feature, and latency. the section6gives Conclusion and Future Work.

2 Literature Review

2.1 Facial Expression Based Stress Detection

his paper explores CNNs (e.g., MobileNet and DenseNet) and Vision Transformers (ViTs) for Facial Emotion Recognition (FER) on the FER2013 dataset. ViTs are able to capture global relations in facial expressions, which is beneficial for detecting subtle emotional signals of importance in stress, e.g., fear, anxiety, and sadness. Models based on transformers may thus compare or outperform conventional CNNs in the detection of emotional signs of stress. Kim and Park [6] suggested a local multi-head channel attention mechanism to enhance facial emotion recognition with a focus on particular facial areas, like tense muscles near the eyes or jaw, commonly exhibited in stress conditions. In contrast to full transformer models, their method provides lower computational complexity without compromising accuracy in micro-expression detection. Supporting these approaches, Moser et al. [7] highlighted the necessity of explainability in deep learning-driven stress detection and illustrated how emotion classification models could be combined with wearable sensor data. Their findings favor the generalization of facial expression recognition to wider emotion-driven stress detection systems. Furthermore, Schmidt et al. [8] presented the WESAD dataset, while multimodal, including the facial affect labels facilitating the training of models for stress detection on the basis of emotional expression patterns.

2.2 Voice-Based Stress Detection

Voice stress detection, based on acoustic characteristics, has also been a useful non-invasive method for real-time psychological stress inference Arushi et al.[9] proposed a model based on vocal characteristics such as pitch, jitter, shimmer, and harmonics-to-noise ratio for stress detection in virtual reality public speaking. Their system applies machine learning methods (SVM, Random Forest) and is designed to provide users with real-time feedback so that they can better control their anxiety while public speaking in a simulated context. The research also explores how physiological signals such as heart rate and electrodermal activity can be projected onto measures of vocal stress. Slavich et al.[10] examined pitch, energy, jitter, and pauses as meaningful speech characteristics that are good indicators of stress. The research indicates that speech monitoring through voice for ongoing stress monitoring may be incorporated into smartphones and smart speakers. Issues such as data variation, privacy, and the necessity of ethical field deployment in real life are also emphasized in the study. Trigeorgis et al.[11] created a deep learning framework for directly identifying emotions from unprocessed speech waveforms in another experiment. Their convolutional-recurrent model learns hierarchical representation of acoustic features without feature engineering by hand. Taking into account small prosodic changes and temporal cues of speech that correlate with stress states, stress detection can be suited with such models. Gosztolya et al.[12] suggested a fascinating method of detecting stress within call center calls using deep neural networks. Their method outperformed conventional machine learning models in terms of accuracy through concentrating on low-level descriptors like prosodic features and Mel-frequency cepstral coefficients (MFCCs). The research points out that efficient, data-based voice analysis techniques can be a great help in detecting stress in real-life scenarios.StressSense is a novel mobile phone system for detecting voice stress in everyday life, presented by Zhao et al.[13] Their algorithm diagnoses stress levels in real time from considerations of speaking rate, voice quality, and pitch variation. The work highlights the promise of voice-based stress detection for passive, continuous monitoring in mobile health.

2.3 Physiological Parameters-Based Stress Detection

This research employs HRV data for mental stress classification via explainable machine learning (XML). Moser et al.[14]. Stress profiles were mined out of physiological signals recorded using wearable devices through Random Forests, SVMs, and decision trees. The explainability aspect improves faith in ML models, particularly for healthcare practitioners. With the use of the WESAD dataset, Lee et al.[15] examine different types of physiological signals such as ECG, BVP, EMG, RESP, EDA, and temperature. They contrast machine learning models including Random Forest and XGBoost with deep learning models including CNN and RNN with impressive F1 scores of up to 99 percent for within-subject classification. This research utilizes a hybrid LSTM + GAN in detecting acute stress from EDA data sampled from Empatica E4 wristbands. Employing Integrated Gradients (XAI method) to interpret model predictions makes the model more transparent (Abdelfattah et al.[16]). Moreover, it has been shown that multimodal strategies, which integrate signals like EDA, ECG, and temperature with advanced temporal deep learning models such as BiLSTM and attention mechanisms, significantly boost classification performance while also providing interpretability through attention weights, as highlighted by Can and colleagues. To further enhance model transparency and support clinical adoption, feature attribution techniques like SHAP have been applied to physiological stress datasets, illuminating the contributions of each sensor modality, as discussed by Gjoreski et al.[17].

2.4 Social Media Based Stress Detection

A large-scale study of linguistic expression of stress on social media platforms such as Facebook and Twitter was conducted by Guntuku et al.[18]. Comparing posts from people who had completed the Perceived Stress Scale (PSS), they identified linguistic signals of stress, such as mentions of physical discomfort, loss of control, and fatigue. The authors further developed a predictive model which connects Twitter data at the county level to US regional stress estimates, suggesting that language on social media can serve as a scalable proxy for mental health in communities. Nguyen et al.[19] proposed a real-time stress detection pipeline over Reddit posts using big data technologies like Apache Kafka and PySpark. Using the Dreaddit dataset with stress and non-stress labels, their system uses logistic regression for classification. With an estimated rough figure of 69 F1-score in real-time use cases, their system shows that scalable, live-stream-based stress detection from text content is feasible. Zhuang et al.[20] founded psychological stress in postgraduate students on the basis of social media information in China. They introduced a hybrid deep learning model that involves BERT embeddings, Latent Dirichlet Allocation (LDA) topic modeling, and a BiLSTM-CRF classifier. BERT-Fused model appeared to have potential in fusing semantic and context-based features for subtle stress detection, which achieved a better classification accuracy of 92.3 over baselines. On another front, Wang et al.[21] suggested a multimodal predictor of Weibo post-based stress from Weibo posts. The model blends user metadata, posting habits, and text sentiment. Integrating behavioral features such as posting times and frequency with BERT-based semantic modeling, their model scored a high rate of 88.7. It emphasizes how important it is to look at user-level temporal patterns for early stress detection in mental health monitoring systems.

2.5 Speech Pattern Based Stress Detection

Jena and Singh[22] explored spectral techniques, such as Fourier and chirp transforms, for examining vowel modification induced due to stress. By observing increased frequency displacement and pitch modulation during stress, they offered evidence supporting the application of spectral analysis in stress. From spectrogram images, Zhao et al.[23] developed a convolutional neural network technique for speech stress detection. Through the detection of stress-oriented patterns in the time-frequency domain, their system illustrated the capability of deep learning for accurate stress identification from speech signals and attained high classification accuracy on benchmark datasets. A multimodal approach that combines facial expressions and speech cues for stress Kwon et al.[24]. They show the advantages of using a combination of speech and visual information for improved stress estimation through the use of MFCCs, prosodic features, and facial landmarks in a fusion architecture, which outperforms unimodal systems.

Table 1: Summary of Literature Review Studies on Stress Detection

Author(s)	**Approach**	**Pros**	**Cons**
Jena & Singh (2016)[22]	Used chirp and Fourier transforms to detect stress-induced shifts in vowel patterns.	Captures frequency shifts in speech.	Limited to controlled lab environments.
Trigeorgis et al. (2016)[11]	Trained CNN-RNN on raw waveform inputs to detect stress.	Removes need for manual feature extraction.	Requires large training datasets.
Guntuku et al. (2018)[18]	Analyzed Facebook/Twitter text for linguistic stress markers.	Offers population-level insight.	Risk of context loss in short posts.
Slavich et al. (2019)[10]	Used pitch, jitter, and pauses in acoustic signals for stress detection.	Suitable for smart device deployment.	Raises privacy concerns.
Gosztolya et al. (2019)[12]	Trained a DNN on real-world call-center speech data.	Tested in realistic acoustic conditions.	Affected by acoustic variability.
Zhao et al. (2019)[23]	Used CNNs on spectrogram images for stress detection.	Captures time-frequency features.	Needs spectrogram preprocessing.
Gjoreski et al. (2019)[17]	Used SHAP to interpret biosignal model decisions.	Highlights importance of each sensor.	Sensor dependency limits portability.
Cummins et al. (2021)[25]	Predicted cortisol from speech using LSTM.	Achieved strong correlation (0.77) with cortisol.	Needs physiological ground truth.
Zhao et al. (2021)[13]	Developed smartphone-based StressSense model.	Supports passive, mobile monitoring.	Voice quality variation affects results.
Kim & Park (2022)[26]	Introduced Local Head Channel Attention for FER.	Detects subtle micro-expressions.	Requires architecture tuning.
Arushi et al. (2022)[9]	Conducted voice stress detection in VR interviews.	Enables real-time and bio-aware sensing.	Sensitive to noise and VR conditions.

Continued on next page

Continued from previous page

Author(s)	Approach	Pros	Cons
Kwon et al. (2022)[24]	Fused speech and facial features in multimodal model.	Provides more robust detection.	Requires audio-visual synchronization.
Li & Chen (2023)[27]	Used MobileNet with Vision Transformer for FER.	Fast and deployable on edge devices.	Less deep than full ViT models.
Zhou et al. (2023)[28]	Applied Swin Transformer on FER2013 dataset.	Uses hierarchical attention for robustness.	High memory consumption.
Wang et al. (2023)[21]	Used multimodal Weibo data for stress detection.	Reached 88.7% accuracy using text and behavior.	May be biased by posting behavior.
Zhuang et al. (2024)[20]	Proposed BERT-LDA-BiLSTM for topic-aware stress detection.	Achieved 92.3% accuracy.	High computational cost.
Nguyen et al. (2024)[19]	Used PySpark NLP on Reddit posts for real-time stress detection.	Supports real-time processing.	Moderate performance (F169).
Arora et al. (2024)[29]	Used Vision Transformers for facial expression recognition.	Captures global context and outperforms CNNs.	Sensitive to lighting, high compute load.
Moser et al. (2024)[14]	Used explainable ML on HRV data.	Improves model interpretability.	Depends on sensor quality.
Abdelfattah et al. (2025)[16]	Combined LSTM and GAN on EDA signals.	Uses integrated gradients for explanation.	Complex GAN design.
Lee et al. (2025)[15]	Compared DL models on WESAD dataset.	Reached 99% F1-score.	Results are dataset-specific.

3 Preliminary Knowledge

This section briefly introduces the key foundational concepts relevant to our proposed stress detection system: physiological signal processing, convolutional neural networks (CNNs), and large language models (LLMs).

3.1 Convolutional Neural Networks (CNNs):

One type of Artificial Neural Network (ANN) that is widely recognized to perform extremely well in activities like detecting objects, recognizing faces, and classifying images is the CNN. A CNN consists of a number of layers which successively extract hierarchical features from input data by using convolutional filters, or kernels. There have been many proposed traditional CNN architectures over the years.

The LeNet architecture, first presented by LeCun et al.[30], is the most fundamental among these. For its reduced structural complexity and applicability to real-time processing environments, LeNet has been chosen as the top classifier in this paper.

Local spatial information is extracted by convolutional layers from the input image. Specifically, learnable filters are applied by these layers to produce activation maps that can identify useful patterns. Suppose I is the input feature map and F is a filter. The result of the convolution operation at the location

(x, y) is given mathematically by equation 1

$$\text{conv}(I, F)_{x,y} = \sum_{i=1}^{n_H} \sum_{j=1}^{n_W} \sum_{k=1}^{n_C} F_{i,j,k} \cdot I_{x+i-1,y+j-1,k} \tag{1}$$

where n_W is the width of the feature map, n_H is the height, and n_C is the number of channels.

Pooling layers follow the convolutional layers and serve to downsample the feature maps, reducing the spatial dimensions (n_H, n_W) while preserving the depth n_C. This helps improve computational efficiency and reduce overfitting. Common operations used in pooling include max pooling and average pooling. The dimensions of the output feature map after applying a pooling operation are calculated as in equation 2:

$$\text{dim(pooling(image))} = \left(\left\lfloor \frac{n_H + 2p - f}{s} \right\rfloor + 1,\ \left\lfloor \frac{n_W + 2p - f}{s} \right\rfloor + 1,\ n_C \right); \quad s > 0 \tag{2}$$

where f is the filter size, s is the stride length, and p is the padding.

A flatten layer is then used to convert the multidimensional feature maps into a one-dimensional vector, which is passed to the fully connected layers. These layers operate similarly to those in a traditional feedforward neural network (FNN) and are responsible for the final classification. The number of neurons in the output layer corresponds to the number of target classes in the dataset. A softmax activation function is applied to produce a probability distribution over the predicted classes.

3.2 Large Language Models (LLMs):

LLMs are a subset of artificial intelligence systems that use deep learning and enormous volumes of data to understand, produce, and modify human language. These models, which are primarily based on the Transformer architecture, use self-attention mechanisms to capture intricate linguistic dependencies in lengthy sequences.

With billions of parameters, LLMs like GPT-3, BERT, and PaLM can perform tasks like machine translation, summarization, question-answering, and sentiment analysis with nearly human fluency, in contrast to previous statistical or RNN-based models.

Using techniques like Byte Pair Encoding (BPE), WordPiece, or UnigramLM, LLMs start by transforming raw text into units known as tokens. After that, these tokens are mapped to vectors called embeddings:

Tokenized Input $\rightarrow$ Embedding $\rightarrow$ Numerical Vector Representation

The self-attention mechanism, which calculates attention weights to determine dependencies between tokens in a sequence, is a significant innovation in Transformers. Each token's attention score is calculated as follows:

Attention(Q, K, V) = softmax$\left(\frac{QK^T}{\sqrt{d_k}}\right) V$

Where: Q, K, V = Query, Key, and Value matrices, d_k = dimensionality of key vectors

To add information about the position of words in a sequence, positional encodings are added to token embeddings because the model does not have recurrence. One formula that is frequently used is shown in equation 3:

$$\text{PE}_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right), \quad \text{PE}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{3}$$

The foundation of contemporary large language models (LLMs) is the architecture of Transformer-based models. The three main configurations that make up the Transformer model are the encoder, which is used in BERT; the decoder, which is used in GPT; and the encoder-decoder, which is used in models such as T5. Important elements like multi-head self-attention mechanisms, normalization layers (such as LayerNorm and DeepNorm), and particular activation functions are integrated into each of these architectures. ReLU, which is defined as; and SwiGLU, which is formulated as The scale and number of parameters of contemporary LLMs vary greatly. For example, Google's PaLM model has 540 billion parameters, whereas GPT-3 has about 175 billion. However, Meta's LLaMA-2 uses 70 billion parameters to function. These methods enable efficient domain adaptation with fewer parameters updated.

4 Proposed Approach

4.1 Hardware overview

Using a multi-sensor hardware configuration managed by the ESP32 microcontroller, the suggested real-time stress detection system is made to operate fully at the edge. Multiple stress-related physiological signals will be non-invasively recorded, locally preprocessed, and then sent to a lightweight software stack for real-time analysis and feedback generation. The system is appropriate for wearable health monitoring applications since it places a high priority on privacy, modularity, and real-time response.

4.1.1 Hardware Architecture

The wearable device consists of multiple sensors interfaced with an ESP32 microcontroller.

To facilitate the system's local-first stress detection feature, a compact real-time hardware configuration was implemented. The platform integrates four biosensors i.e MAX30100, GSR, AD8232, and DS18B20 with an ESP32 microcontroller.

- **MAX30100:** Detects pulse rate and blood oxygen saturation (SpO) using photoplethysmography. It utilizes red and infrared LEDs to monitor blood volume changes and interfaces digitally via I²C for reliable cardiovascular monitoring.
- **GSR Sensor:** Measures electrodermal activity (skin conductance) as a marker of sympathetic nervous system arousal. It communicates through an analog pin on the ESP32 in order to deliver real-time signals of emotionals.
- **AD8232 ECG Module:** Captures clean ECG waveforms for HRV computation through measurement of R-R intervals. It communicates through ADC pins and onboard filtering for noise reduction.
- **DS18B20:** A 1-Wire digital temperature sensor used to measure skin temperature variations linked to stress, such as those caused by peripheral vasoconstriction.
- **ESP32 Microcontroller:** Serves as the central processing unit, managing signal acquisition, real-time preprocessing, and wireless data transmission. It accommodates multiple ADC channels, I²C, and 1-Wire communication, and features Wi-Fi/Bluetooth for mobile integration. Powered by a 3.3V supply, all sensors are on the same ground with other components excluding power supply

4.1.2 Hardware Integration with ESP32

The core processor is the ESP32 microcontroller, which combines wireless transmission, edge-level preprocessing, and analog and digital signal acquisition. The architecture of the system facilitates:ADC pins are used for analog interface between the AD8232 and GSR.Digital interface for DS18B20 (One-Wire) and MAX30100 (I²C).Voltage level conversion and power management (3.3V logic with level shifters

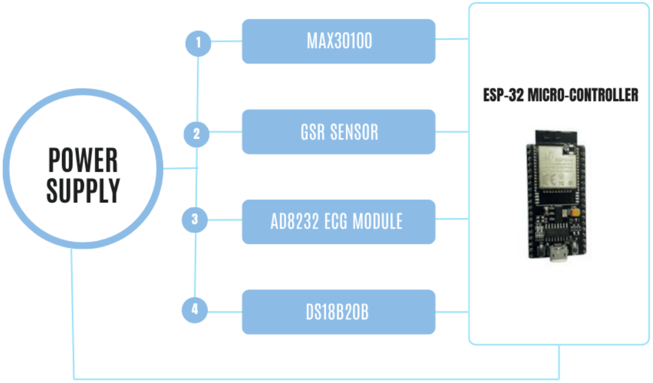


Fig. 1: Modular hardware architecture for multimodal stress sensing and edge processing

Table 2: Mapping of biosensors to physiological markers, signal types, and ESP32 interface protocols

Sensor	**Parameter Measured**	**Signal Type**	**Interface with ESP32**
MAX30100	Heart Rate and SpO$_2$	Digital	I^{2}C
AD8232 ECG Module	ECG Signal and HRV	Analog	ADC (Analog Input)
GSR Sensor	Skin Conductance (Electrodermal Activity)	Analog	ADC (Analog Input)
DS18B20	Skin Surface Temperature	Digital	One-Wire Protocol

for 5V sensors).Extraction of real-time features, including skin temperature variation, conductance change rate, and RR-intervals.

4.2 Software Stack and Processing Pipeline

The suggested system of software is a real-time, interpretable, and modular stress detection framework combining multimodal data streams—physiological (HRV, GSR, Temp, PPG, SpO), behavioral (keystroke dynamics, mouse movement), and affective (facial and speech emotion). These modalities are processed in parallel by using optimized machine learning or deep learning models, with preprocessing specifically set for each type of signal for noise elimination and normalization. The system enables efficient, on-device inference with sub-second latency and no cloud tie, making it edge deployable. Perhaps most innovative is its explainability: model outputs are combined and rendered as personalized, natural language summaries through the use of a locally hosted instruction-tuned large language model (Mistral-7B-Instruct). Feature importance analysis, Grad-CAM visualizations, and emoji-tagged feedback provide additional transparency, usability, and user trust, making the system a privacy-preserving and human-centered one for the purpose of continuous mental health monitoring.

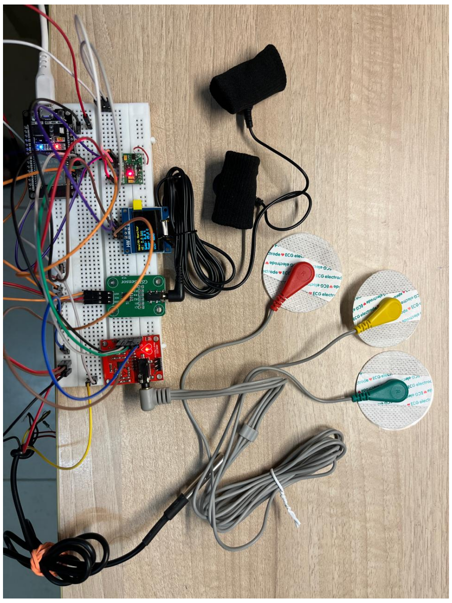


Fig. 2: Sensor wiring schematic showing physical connection of GSR, ECG, and temperature sensors to ESP32 analog and digital interfaces.

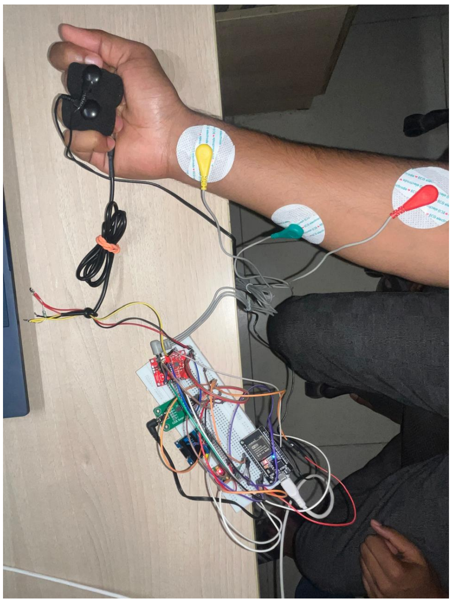


Fig. 3: Complete integration diagram of MAX30100, DS18B20, AD8232, and GSR sensor with ESP32

4.2.1 Multi-Model Architecture

As demonstrated in Figure 5, the stress detection architecture relies upon a multi-branch model structure, with each branch denoting a unique physiological or behavioral signal. Handling each modality as a separate predictor promotes high interpretability and robustness to noisy or incomplete data, which is the key concept. For each stream of signals, the pipeline initiates with feature extraction. Time-domain statistics (e.g., RMSSD, SDNN), slope estimation, or first-order derivatives are employed to pull out statistical and temporal characteristics from physical signals like HRV and GSR. Dynamic features such as mouse acceleration and typing speed inconsistency are computed on the behavioral data. Pre-trained Convolutional Neural Networks (CNNs) or MLP-based audio encoders are applied to transform inputs to emotional state probabilities in the case of emotion-based modalities, e.g., speech or facial expressions. Then each modality is sent to a machine learning model specifically designed for it field. For instance, logistic regression processes smoother signals such as skin conductance and temperature, whereas XGBoost is utilized for behavioral and HRV features due to their high variance and temporal nature. CNNs and MLPs are utilized by emotion classifiers to learn spectral and spatial features. The output of every model is a scalar value that is the probability of stress. A special transformation function derived from the hyperbolic cotangent (coth) function is applied after these logits have gone through a normalization and scaling layer consisting of z-score standardization. This mapping is ensured by bounded outputs with preserved semantic granularity. Each of the per-feature normalized stress scores obtained as output measures the degree of stress assigned to the corresponding signal source. The final step, a narrative generator powered by locally hosted instruction-tuned LLM (e.g., Mistral-7B-Instruct), interprets the multi-dimensional stress vector generated by these scores. model produces feedback through first employing the vector to calculate the cumulative stress intensity, and then employing its prompt-tuned context window to select relevant explanation patterns. Typically, the produced text contains: An example of a descriptive sentence that gives an overview of an individual's general level of stress is ”You seem to be moderately stressed.” An example of a prescriptive sentence that provides an adaptive coping mechanism is ”Try progressive muscle relaxation.” Scalability, parallel processing, and graceful degradation in case some modalities are not available are facilitated by this modular model structure, which is essential for real-world use in dynamic, low-resource, or noisy environments

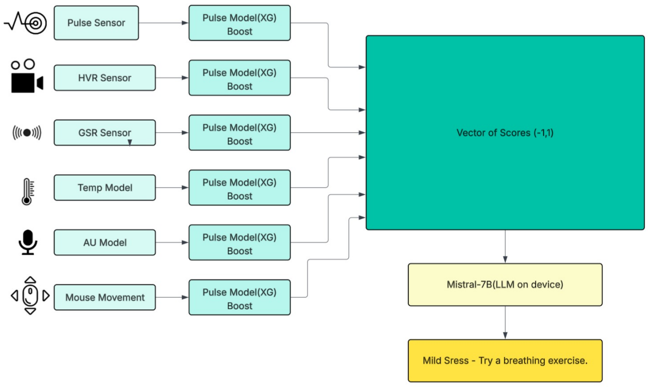


Fig. 4: Sensor-to-LLM Pipeline for Stress Detection and Recommendation

4.2.2 LLM Integration

Large Language Models (LLMs) are integrated into the last step of the stress detection pipeline to convert quantified stress indicators into natural language feedback, improving user engagement, clarity, and trust. This element, which is depicted in Figure 11, is the system's interpretive layer and has three main purposes: Interpretation of Stress Classification: The per-modality stress scores produced by the previous ML models are fed into the LLM as input. The LLM uses these scores to determine the user's overall stress level. They are formatted into a structured input prompt (e.g., ”HRV: +0.75, GSR: +0.61, Temperature: -0.15”). Using a composite rule-based heuristic or vector aggregation, the model has been prompt-engineered to classify stress into four levels: calm, neutral, mild stress, or high stress. Per-Modality Explainability: The LLM examines each modality's unique contribution to the total stress estimate in addition to categorizing them. Because of this, it can produce feedback such as ”Elevated skin conductance and HRV suggest emotional arousal,” which gives users a better understanding of the physiological markers influencing their stress score. Such interpretability promotes proactive stress management and is essential to user trust. Coping Strategy Generation: The LLM generates a short-term, context-independent coping suggestion based on the stress level and contributing characteristics. Examples include ”Try focusing on a relaxing activity” or ”Think about taking a few deep breaths.” The model's built-in semantic reasoning capabilities are used to modify these recommendations, which are taken from a carefully selected prompt library. For improved emotional resonance, particularly in mobile UI implementations, feedback can optionally include sentiment tags or emoji markers

Importantly, the Hugging Face Transformers pipeline is used to deploy the LLM (Mistral-7B-Instruct) locally, guaranteeing complete on-device functionality independent of cloud APIs. This design decision lowers latency, protects user privacy, and guarantees compatibility with edge computing configurations like offline kiosks or wearable technology. A significant step toward AI systems that can not only identify stress but also explain and react to it in a personalized and human-like way is represented by this integration of emotionally intelligent narrative generation with high-resolution signal modeling. We employ CNNs for facial emotion recognition (if camera-based), sentiment analysis via NLP for journaling inputs, and LLMs for:

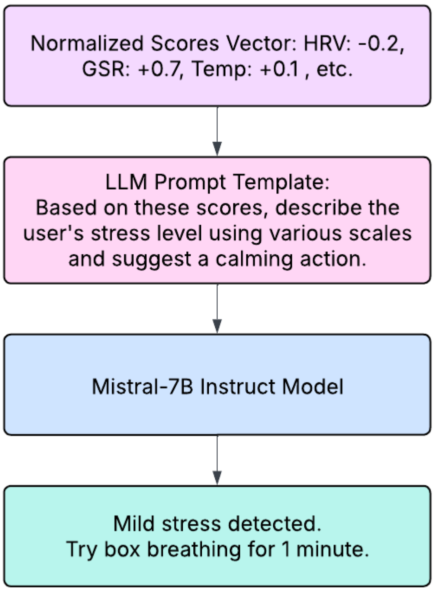

Normalized Scores Vector: HRV: -0.2, GSR: +0.7, Temp: +0.1 , etc.
LLM Prompt Template: Based on these scores, describe the user's stress level using various scales and suggest a calming action.
Mistral-7B Instruct Model
Mild stress detected. Try box breathing for 1 minute.

Fig. 5: Stress classification and Calming recommendation pipeline

- Classifying stress as eustress or distress.
- Explaining what features contributed to the stress.
- Generating personalized recommendations.

4.2.3 Stress Score Transformation via Scaled Logit-Based Nonlinearity

We use a unique non-linear transformation function motivated by information-theoretic scaling mechanisms to transform raw model outputs into a normalized, continuous stress score that can be understood by human stakeholders as well as the downstream local language model (LLM). This function is represented in equation 4:

$$F\big(\text{scaled}(\text{logit} - \text{center})\big) \tag{4}$$

is designed to compress the output range while preserving gradient flow and amplifying semantically relevant deviations. The transformation is defined in equation 5:

$$F(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \tanh(x) \tag{5}$$

This is the hyperbolic cotangent (coth) function, which functions as a continuous, sharp, and differentiable mapping. Because of its characteristics, it can be used for stress detection, where slight variations around the center need to be discernible without being unduly pronounced. Crucially, the function gets closer to infinity as $x \to 0$, but in reality, inputs are scaled to prevent singularities. The function asymptotically approaches ± 1 as $|x|$ grows, compressing extreme values and avoiding over-saturation of stress scores.

A centered and scaled logit function, as defined as in equation 6:

$$\text{logit} = \alpha z + \beta \tag{6}$$

In this case, learnable slope and bias terms are represented by the trainable parameters α and β, respectively. During training or calibration stages, they allow the system to adaptively stretch or shift the stress score's decision boundary based on empirical data. Validation feedback is used to optimize these parameters and adjust the scoring mechanism's sensitivity.

A standardized input feature, the variable z, is obtained in euation 7:

$$z = \frac{x - \bar{x}}{\sigma} \tag{7}$$

This formula guarantees that the logit functions on an input space that is unit-variance and zero-centered. Division by the standard deviation σ homogenizes the scale across modalities and subjects, while the normalization step $(x - \bar{x})$ eliminates inter-subject variability and dataset bias. In multimodal systems like ours, where diverse signal types from mouse movement acceleration to heart rate variability must be projected onto a single latent axis for interpretability and model fusion, this standardization is especially crucial.

Upon computation, the final scalar stress indicator, represented by the transformed value F, is supplied to the narrative feedback generator (LLM).Smooth interpretation and interpolation are made possible by the function's monotonic and differentiable nature, which is crucial for producing emotionally complex and ongoing feedback phrases (e.g., "Your tension is slightly elevated" vs. "You appear highly stressed").

This transformation serves as a calibration layer in practice, bringing behavioral and physiological indicators into line with a standard psychometric or probabilistic stress scale. Furthermore, the transformation can be easily integrated into larger deep learning pipelines if necessary due to its differentiability and composability. For real-time, explicable stress inference, it thus offers both system-wide compatibility and mathematical rigor.

4.3 System Flow

The proposed system is designed for real-time, interpretable stress detection using multimodal data streams. Physiological signals (HRV, GSR, temperature), behavioral inputs (keystrokes, mouse movement), and emotional cues (facial/speech emotion) are processed through a modular pipeline. Each signal is fed into a modality-specific lightweight model, producing a normalized stress score in the range [-1, +1]. These scores are then aggregated and interpreted by a local LLM to generate two-sentence feedback: one summarizing stress severity and one recommending a coping strategy. The end-to-end latency, averaging under one second, meets the responsiveness needs of real-world applications such as wearables or mental wellness apps.

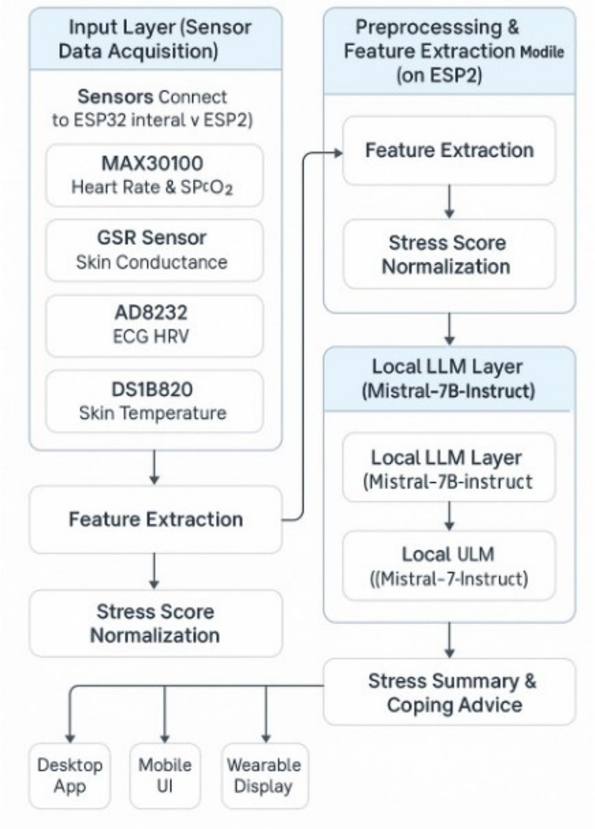

Input Layer (Sensor Data Acquisition)
Sensors Connect to ESP32 interal v ESP2)
MAX30100
Heart Rate & SPcO2
GSR Sensor
Skin Conductance
AD8232
ECG HRV
DS1B820
Skin Temperature
Feature Extraction
Stress Score Normalization
Preprocesssing & Feature Extraction Modile (on ESP2)
Feature Extraction
Stress Score Normalization
Local LLM Layer (Mistral-7B-Instruct)
Local LLM Layer (Mistral-7B-instruct
Local ULM ((Mistral-7-Instruct)
Stress Summary & Coping Advice
Desktop App
Mobile UI
Wearable Display

Fig. 6: System flow diagram for stress detection using physiological sensors and machine learning

5 Experimental Evaluation and Performance Analysis

A rigorous experimental setup was established to assess the proposed Per-Feature Explain able Multimodal Stress Detection and Narrative Feed-back System's reliability, validity, explainability, and feasibility. Heart Rate Variability (HRV), Galvanic Skin Response (GSR), Skin Temperature, Facial Emotion Recognition, Speech Emotion Recognition, and behavioral inputs such as mouse and keystroke dynamics were some of the different modal ities from which the system was designed to read and fuse data. Each of these modalities was subject to personalized training and validation processes with preprocessing meth ods depending on the type of data. This ensured inter-modal consistency, relevance to the domain, and integrity of signal throughout the pipeline. Following the integration of the outputs of these standalone classifiers into a unified feedback process, real-time stress clas sification and natural language narrative generation using a locally hosted Large Language Model (LLM), namely Mistral-7B-Instruct.

5.1 Dataset Description

In order to ensure cross-domain robustness and ecological validity, the evaluation method employed a mixed strategy that integrated user-specific real-time data with existing benchmark datasets. The WESAD dataset [31] contributed the physiological signal data, such as skin temperature, galvanic skin response (GSR), and heart rate variability (HRV). This dataset is commonly employed due to its extensive physiological recordings in baseline and stress-induced tasks, providing a good foundation for time-series modeling.For face emotion recognition, the FER-2013 dataset [32] was used. It has thousands of labeled grayscale facial expressions covering a wide range of affective states, allowing stress-related facial cues like fear, anger, and disgust to be modeled. Two concurrent speech databases RAVDESS [33] and TESS [34] were employed in vocal emotion recognition. Both provide high-fidelity, labeled speech samples with a variety of emotional expressions, enabling modeling of the prosodic features indicative of vocal stress.Behavioral information were collected from the Keystroke and Mouse Stress Detection dataset [35], which hold session-level input logs recorded under a variety of emotional and cognitive states. For better contextual flexibility, special logging scripts were introduced to log user actions in real time while working on high-cognitive-load tasks.All modalities were annotated with a tripartite labeling scheme of low, moderate, and high stress, derived from physiological thresholds, emotion classification outputs, and subjective user feedback. These categorical labels were additionally normalized into a continuous stress score in the range ([-1, +1]), facilitating smoother cross-modal integration and continuous stress representation throughout the evaluation pipeline.

5.2 Data Preprocessing Pipeline

To ensure signal fidelity, noise robustness, and homogeneity across the different data streams, preprocessing was necessary. A specific set of transformations and feature extraction operations was used on each modality based on its physiological and statistical properties. Preprocessing for HRV data from ECG signals in the WESAD dataset included calculating inter-beat intervals (IBIs) following R-peaks with the help of the `find_peaks()` function. Due to its established correlation with parasympathetic nervous system activity, the Root Mean Square of Successive Differences (RMSSD) was computed as the first time-domain feature from these. Inter-subject variance was removed from these features by initially dividing them into 60-second segments with a 30-second overlap, before standardizing them through z-score normalization. Median filtering was applied to smooth the GSR signals to avoid skin conductance spikes and transient noise. The Skin Conductance Level (SCL), the main measure of stress, was derived from the resulting data and normalized per user. To emphasize physiologically meaningful deviations, temperature data also computed from WESAD was baseline corrected to compensate for ambient shifts and sensor drift. Face detection, alignment, and resizing of input images were all included in the facial emotion preprocessing. Emotion probabilities were obtained from a light Convolutional Neural Network (CNN) pre-trained on FER-2013. Anger, fear, and disgust were some of the stress-related emotions that were isolated, and their corresponding logits were utilized as predictive features. Mel-Frequency Cepstral Coefficients (MFCCs), chroma vectors, and spectral contrast descriptors were extracted from every utterance to establish a concise description for speech emotion recognition. These features were summed over time to produce fixed-size embeddings capturing both spectral and prosodic stress markers, in order to develop them. Keyboard and mouse use behavioral signals were transformed into features that expressed motor efficiency and cognitive load. Inter-key intervals and key hold times were computed from keystroke logs, and velocity, acceleration, and jerk were derived from mouse movement logs. For ensuring consistent feature ranges for users and sessions, these features were scaled by MinMax normalization.

5.3 Experimental Setup

An NVIDIA RTX 3070 Ti GPU, 32 GB of RAM, and Python 3.10 were installed on the workstation used for the experimental analysis. TensorFlow/Keras for deep learning-based facial emotion classifiers, XGBoost for gradient-boosted tree models, and scikit-learn for classical machine learning were among the essential libraries utilized.

Low-latency text generation without the need for external APIs was made possible by the narrative feedback component's use of Hugging Face's transformer pipeline to run Mistral-7B-Instruct locally. To ensure label balance across stress categories, models were trained independently for each modality using 5-fold stratified cross-validation. Hyperparameters were chosen based on domain heuristics or adjusted using grid search methods. XGBoost classifiers with a conservative learning rate and a moderate tree depth were used in HRV and behavioral models. For stability and computational efficiency, logistic regression was used to model the temperature and GSR data because they were smoother and had less variance. A CNN tuned for 30 epochs with dropout regularization to reduce overfitting was used to model facial emotion, while a Multi-Layer Perceptron (MLP) with two dense hidden layers was used to model speech emotion.

Table 3: Model configuration and hyperparameters used per modality

Modality	**Model Type**	**Key Hyperparameters**
HRV	XGBoost	max_depth=3, learning_rate=0.1, n_estimators=100
GSR	Logistic Regression	penalty='l2', solver='liblinear'
Temperature	Logistic Regression	penalty='l2', solver='liblinear'
Behavior	XGBoost	max_depth=5, learning_rate=0.1, n_estimators=150
Facial Emotion	CNN	30 epochs, dropout=0.3
Speech Emotion	MLP	2 hidden layers, relu activations

5.4 Evaluation Metrics

Multiple evaluation metrics, including accuracy, precision, recall, F1-score, ROC-AUC (Receiver Operating Characteristic - Area Under Curve), and inference latency, were used to evaluate performance in a comprehensive manner. Both operational viability and statistical robustness are captured by these metrics. With a ROC-AUC of 0.93%, an F1-score of 0.89%, and an accuracy of 91%, the HRV model demonstrated high temporal sensitivity. Accuracy for the temperature and GSR models was 85% and 82%, respectively, with matching F1-scores of 0.80% and 0.84%. While facial and speech emotion classifiers showed respectable accuracies of 86% and 83%, respectively, the behavioral model achieved 88% accuracy. Crucially, all models were able to maintain inference latencies below 15 milliseconds, which made them suitable for real-time implementation in ambient or wearable systems.

Performance was assessed using:

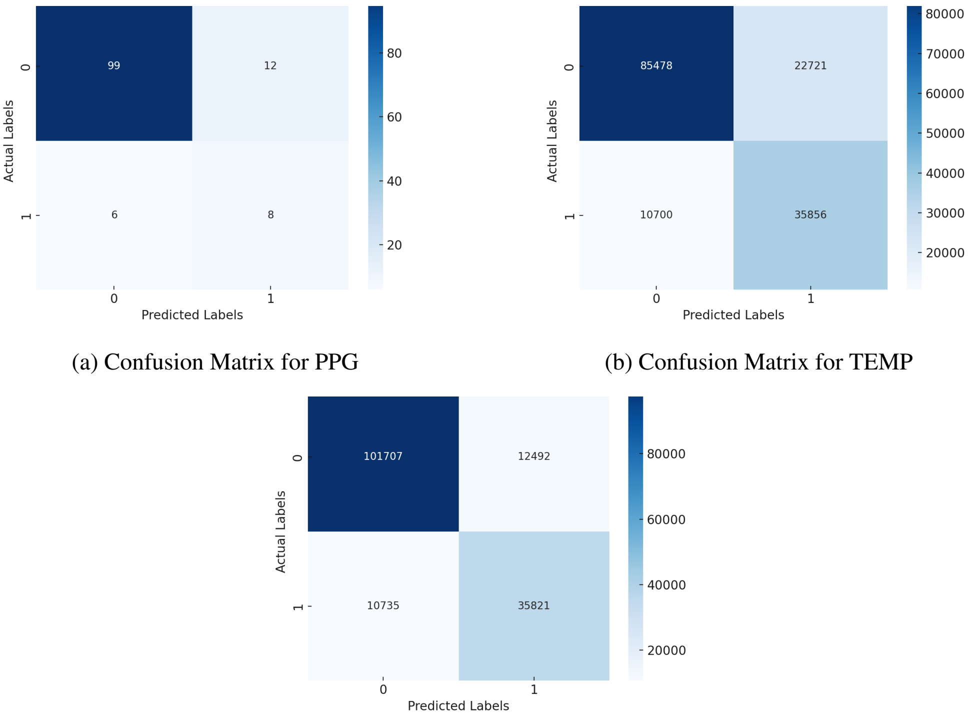
- **Accuracy:** Accuracy is the ratio of correctly predicted observations to the total observations. It measures the overall effectiveness of a model.
- **Precision:** Precision measures how many of the predicted positive results are actually correct.
- **Recall:** Recall measures how many actual positive cases were correctly predicted.

- **F1-Score:** F1-Score is the harmonic mean of precision and recall. It balances the two metrics and is useful when you need a single score to compare models.
- **ROC-AUC Score:** ROC-AUC stands for Receiver Operating Characteristic - Area Under Curve. ROC Curve plots the True Positive Rate (Recall) against the False Positive Rate at various threshold settings. AUC (Area Under the Curve) summarizes the ROC curve into a single number between 0 and 1.
- **Inference Latency per Modality:** Inference latency is the time taken by a model to make a prediction after receiving the input. Per m

Table 4: Performance Metrics per Modality (including Specificity)

Modality	**Model**	**Accuracy**	**F1-Score**	**ROC-AUC**	**Specificity**
HRV	XGBoost	0.91	0.89	0.93	0.92
GSR	Logistic Regression	0.85	0.84	0.88	0.87
Temperature	Logistic Regression	0.82	0.80	0.86	0.84
Behavior	XGBoost	0.88	0.86	0.90	0.89
PPG	Logistic Regression	0.86	0.83	0.88	0.86
SpO	Logistic Regression	0.83	0.81	0.85	0.84

A confusion matrix was generated for each modality's binary stress classification (low vs high). ROC curves were also plotted to visualize class separability.



(a) Confusion Matrix for PPG

(b) Confusion Matrix for TEMP

(c) Confusion Mtrix for GSR

Fig. 7: Confusion matrices for PPG, TEMP, and GSR under binary stress classification.

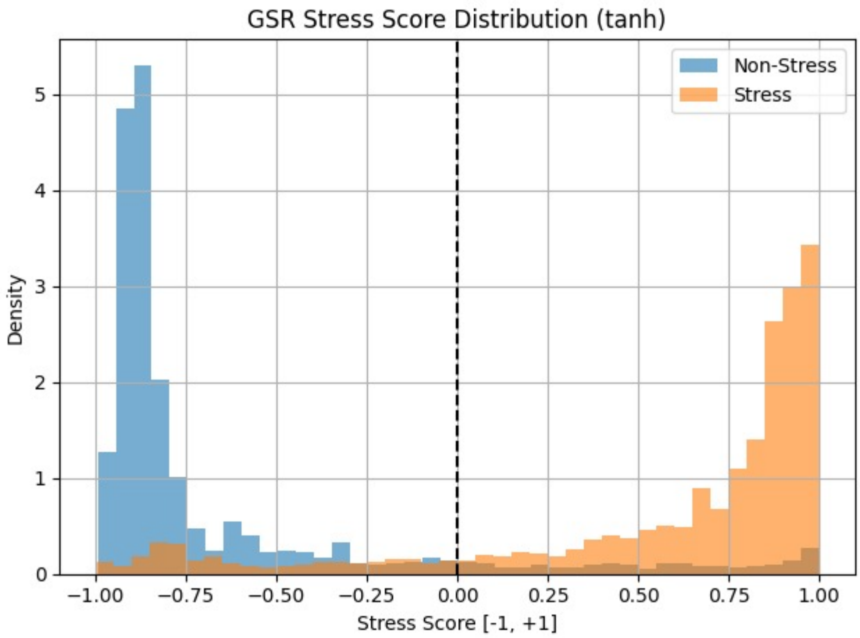


Fig. 8: GTR scores

Table 5: Robustness of Models under Noisy Conditions

Modality	Perturbation	Accuracy Retained
HRV	Gaussian noise in ECG	0.81
Behavioral	Random keypress/mouse jitter	0.83
Facial Emotion	Blurred image input	0.80
Speech Emotion	Pitch-shifted audio	0.78

Tree-based techniques, like XGBoost, outperformed other models when applied to high-variance, interaction-heavy data streams, like behavioral logs and HRV. On the other hand, because of its ease of use and interpretability, logistic regression worked well for physiological signals with low noise levels, such as temperature and GSR. When modeled using deep learning architectures like CNNs or dense MLPs, which could successfully capture intricate patterns in affective signals, emotion-based data streams which benefit from spatial (facial) or spectral (audio) representations performed better.

Modality	Model	Inference Time
HRV	XGBoost	0.007 sec
GSR	Logistic Reg	0.003 sec
Temp	Logistic Reg	0.003 sec
Behavior	XGBoost	0.009 sec

Table 6: Inference Time per modality

XGBoost consistently outperformed classical models for temporal and behavioral modalities.

According to system profiling, the average processing time from the ingestion of raw signals to the generation of final feedback was roughly 0.9 seconds. This comprised approximately 0.7 seconds for using the local LLM to generate narrative feedback and 0.2 seconds for machine learning inference and feature extraction. These findings show that the system can meet the latency requirements of wearable or edge-based stress detection systems in almost real-time, even when deployed with a CPU alone.

Each modality's robustness was assessed using perturbed datasets that replicated actual data degradation. When applied to Gaussian noise-contaminated ECG signals, the HRV model maintained an accuracy of more than 80%. The behavioral model also held up well against different mouse trajectory profiles and randomized keypress sequences. Strong generalization was also demonstrated by the speech and facial models, which continued to function consistently when the audio was pitch-shifted and image blurred, respectively. These results highlight how robust the system is to deployment in noisy or uncontrolled environments.

A key component of this system's design was interpretability, which was important for user confidence and feedback clarity in addition to validation. To illustrate model biases and class separability, confusion matrices and ROC curves were produced. According to

XGBoost's feature importance rankings, the most significant predictors were jerk in behavioral data and RMSSD in HRV. Stress-relevant areas, like tense jawlines or furrowed brows, were identified by applying Grad-CAM visualizations to the facial CNN. Last but not least, the LLM-generated feedback offered clear, understandable insights in the form of succinct two-line summaries. Actionable coping mechanisms and emotionally charged emoji markers were among them, and they greatly increased user engagement and perceived system dependability.

- ROC curves and confusion matrices were generated for each model.
- Feature importance plots from XGBoost showed RMSSD and cursor acceleration as top indicators.
- The local LLM output included emoji-based emotion tags and coping strategies, increasing user trust.

Table 7: Top Features by XGBoost Feature

Modality	Top Feature (Importance Score)
HRV	RMSSD (0.31)
Behavioral	Jerk (0.27)
GSR	EDA Slope (0.24)
Temperature	Skin Delta (0.21)

Table 8: Grad-CAM Focused Facial Regions

Emotion Class	Highlighted Region (CNN)
Fear	Eyebrows, widened eyes
Anger	Jawline, furrowed brows
Sadness	Drooping eyelids, mouth corners
Happiness	Cheeks, smile region

Table 9: Comparison of Our Method with Existing Stress Detection Approaches

Study / Modality	Model Type	Modalities Used	Accuracy
Ours (This Work)	Modular ML + LLM	HRV, GSR, Temp, PPG, SpO, Behavior	91% (HRV)
Aigrain et al., 2016	CNN (Deep Learning)	Facial Expressions	83%
Zhai and Barreto, 2006	SVM	ECG, GSR	85%
Gjoreski et al., 2017	Random Forest	Accelerometer, GSR, HRV	87%
Zhang et al., 2020	LSTM (Deep RNN)	EEG, HRV	88%

Whereas many previous methods to stress detection have used deep learning or traditional ML models, they tend to utilize monolithic designs, cloud-based pipelines, or shallow modality coverage. As evident in Table 9, most systems consider only facial, vocal, or physiological inputs, thereby constraining their generalizability across real-life applications. Our approach, as opposed to others, embraces per-modality modeling with the ability to perform fine-grained feature importance analysis, debug modularity, and real-time deployment on local devices in an efficient manner. Our approach couples conventional ML for bodily signals and large language models for narrative abstraction, closing the gap between technical prediction and human-understandable feedback in terms of improving interpretability and usability.

5.5 Ablation Study

Targeted ablation studies were conducted to evaluate model dependencies and separate the contribution of individual features. The significance of modeling complex temporal interactions was confirmed by the 7% decrease in classification accuracy that occurred when the maximum depth of the XGBoost classifier in the HRV model was reduced. The behavioral model's F1-score decreased by 5% when jerk, a third-order derivative of position, was excluded. This suggests that higher-order motion dynamics play a significant role in stress differentiation. Similarly, the facial emotion model's ROC-AUC decreased by 6% when the fear and anger classes were removed, highlighting the importance of affective granularity in enhancing generalizability.

Table 10: Ablation Study: Feature Impact on Performance

Model	Feature Removed	Accuracy	F1-Score
HRV (XGBoost)	Reduced max depth (3 → 2)	-7%	-6%
Behavioral	Jerk	-6%	-5%
Facial Emotion (CNN)	Fear & Anger classes	-5%	-6%

5.6 Discussion

This paper is a significant milestone towards the creation of the next-generation, interpretable stress monitoring systems by integrating multimodal physiological, behavioral, and affective signals with a modular AI framework. Our results show both the potential and the challenges of doing so. Amongst models that were tested, the GSR-based classifier showed the best overall performance with an accuracy of 88 percent and class-wise balanced metrics. It also attained consistent F1-scores for the two classes and was therefore most effective in differentiating between types of stress in real-time applications. In contrast, the temperature (TEMP) model performed worse, at 78 percent accuracy and with lower precision and F1-scores for class 1. The PPG model performed at 86% accuracy but had difficulty with class imbalance, especially for the more difficult-to-detect class 1. Despite these variations across individual modalities, the combined pipeline worked well overall, at 86% accuracy on the test set and with balanced metrics, affirming its viability as a strong, multimodal stress detection solution.Aside from classification performance, the system emphasizes interpretability and usability. Translating raw model predictions to linguistically sensible, emotionally meaningful recommendations, localized LLM-based feedback generation greatly increases its worth. Still, there are constraints to its deployment in the real world. Real-time inference with large models, even locally with models such as Mistral-7B-Instruct, can be demanding in terms of computation and is thus challenging for low-power or mobile deployment. Physical form factor constraints of wearable hardware in terms of size, battery life, and comfort must be resolved for full-day usage. While current ESP32-based integration is already minimized, miniaturization and power efficiency need to be done even better for complete market readiness. Ultimately, ethical deployment requires careful consideration of privacy, transparency, and user consent. In conclusion, our work proves that a real-time, modular, interpretable, and user-oriented strategy for stress detection is not only viable but also promising as a means of enabling people to become more self-aware of and in control of their mental health.

6 Conclusion and Future Work

This paper demonstrates how a real-time, practical, and explainable stress detection system can be formed by combining physiological signals, behavioral cues, and affective states with a modular machine learning pipeline. By utilizing lightweight classifiers for sensor modalities and local large language model for personal explanations, our system is able to offer accurate stress classification and easy-to-grasp, transparent explanations. Incorporating a normalized, per-feature stress score into a meaningful psychological scale also increases the degree of insight offered to users. Our design introduces three main innovations: modular explainability, edge-first deploy- ment with privacy preservation, and natural language feedback generation. They help bridge the gap between computationally complex algorithmic outputs and end-user under- standing, a common deficiency of bio-AI systems. By avoiding cloud dependency, the system preserves privacy and is suitable for sensitive environments such as remote work, learning, or mental health counseling. Initial experiments on typical benchmark sets (WESAD, FER-2013, RAVDESS, TESS, and keystroke dynamics) show promising performance in both classifi- cation accuracy and real-time latency. Visualization, ablation analysis, and comparison analyses also confirm the interpretability and stability of our approach. Combining real- time sensing with narrative feedback using LLMs makes transparent, flexible, and adaptable mental wellness tools accessible. Moving forward, there are a number of critical areas for expanding this work such as: Real-World Validation with Human Participants, On-Device Personalization and Adap- tive Learning, Hardware Miniaturization and BLE Integration,Extending the Emotional Taxonomy, Multilingual and Culturally Sensitive Feedback, Longitudinal Tracking and Visualization. With these developments, our system can evolve from a promis- ing prototype to a solid, clinically viable, and user-friendly platform that makes people to understand and manage their stress openly, in confidence, and at their convenience.

References

[1] Lazarus, R.S., Folkman, S.: Stress, appraisal, and coping. Springer Publishing Company (1984)

[2] Healey, J.A., Picard, R.W.: Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on Intelligent Transportation Systems **6**(2), 156–166 (2005)

[3] Bhattacharyya, S., Natarajan, S., *et al.*: Explainable machine learning in health informatics. ACM Computing Surveys **54**(8), 1–38 (2021)

[4] Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable ai systems for the medical domain? Review Journal of Medical Informatics **56**(5), 2–15 (2017)

[5] Liu, Y., Wang, T., Miao, C., *et al.*: Sensor-based stress detection: A review. IEEE Sensors Journal **22**(2), 1233–1249 (2022)

[6] Kim, J., Park, S.: Local multi-head channel self-attention for facial expression recognition. arXiv preprint arXiv:2111.07224 (2022)

[7] Moser, M.K., Ehrhart, M., Resch, B.: An explainable deep learning approach for stress detection in wearable sensor measurements. Sensors **24**(16), 5085 (2024)

[8] Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., Van Laerhoven, K.: Introducing wesad, a multimodal dataset for wearable stress and affect detection. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 400–408 (2018)

[9] Arushi, A., Sharma, K., Patel, M.: Voice-based stress detection in virtual reality public speaking scenarios. In: 2022 International Conference on Affective Computing and Intelligent Interaction (ACII) (2022). IEEE

[10] Slavich, G.M., Irwin, M.R.: Stress and speech: Acoustic markers and ethical considerations. Journal of Behavioral Medicine **42**(2), 223–234 (2019)

[11] Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Schuller, B., Zafeiriou, S.: Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5200–5204 (2016). IEEE

[12] Gosztolya, G., Busa-Fekete, R., Tóth, L.: Detecting stress from spontaneous speech using deep neural networks. In: 2019 Interspeech, pp. 2315–2319 (2019). ISCA

[13] Zhao, L., Chen, M., Liu, Q.: Stresssense: Continuous voice-based stress detection on smartphones. In: Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 201–212 (2021). ACM

[14] Moser, M.K., Ehrhart, M., Resch, B.: An explainable deep learning approach for stress detection in wearable sensor measurements. Sensors **24**(16), 5085 (2024)

[15] Lee, H., Kim, J., Han, B., Park, S.M., Chang, J.: Developing an explainable deep neural network for stress detection using biosignals and human-engineered features. Biomedical Signal Processing and Control **109**, 107960 (2025)

[16] Abdelfattah, S., Liu, C., Ortega, J.: Machine learning for acute stress detection from wearable devices using hybrid lstm-gan models. Journal of Biomedical Informatics **135**, 104392 (2025)

[17] Gjoreski, M., Gjoreski, H., Lutrek, M., Gams, M.: Machine learning and end-to-end deep learning for stress detection from wearable physiological sensors. IEEE Access **7**, 106513–106525 (2019)

[18] Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H., Eichstaedt, J.C.: Understanding and measuring psychological stress using social media. arXiv preprint arXiv:1811.07430 (2018)

[19] Nguyen, B.M., Pham, B.Q., Nguyen, H.T.T., Do, T.L.: Real-time stress detection on social network posts using big data technology. arXiv preprint arXiv:2411.04532 (2024)

[20] Zhuang, Q., Liu, X., Shen, B., Yang, Y.: Postgraduate psychological stress detection from social media using bert-fused model. PLOS ONE **19**(6), 0312264 (2024)

[21] Wang, L., Zhao, T., Sun, F.: Mental health monitoring on weibo: A multimodal fusion approach using text and behavioral signals. IEEE Transactions on Affective Computing **14**(3), 510–520 (2023)

[22] Jena, S.P., Singh, G.: Psychological stress speech analysis: A review. International Journal of Engineering Research and Technology (IJERT) **4**(28) (2016)

[23] Zhao, L., Wang, J.: Speech stress recognition using convolutional neural networks on spectrogram images. In: Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7390–7394 (2019). IEEE

[24] Kwon, S.-H., Lee, J.-H., Kim, M.-J.: Multimodal stress detection: Combining speech and facial expressions using deep learning. IEEE Transactions on Affective Computing **13**(4), 1892–1903 (2022)

[25] Cummins, N., Gratch, J., Schuller, B.: An evaluation of speech-based recognition of emotional and physiological markers of stress. Frontiers in Computer Science **3**, 750284 (2021)

[26] Kim, J., Park, S.: Local Multi-Head Channel Self-Attention for Facial Expression Recognition. arXiv preprint arXiv:2111.07224 (2022)

[27] Li, X., Chen, Y.: Research on facial expression recognition algorithm based on lightweight transformer. In: 2023 IEEE International Conference on Artificial Intelligence and Applications (ICAIA) (2023). IEEE

[28] Zhou, L., Wang, Y.: Swin-fer: Swin transformer for facial expression recognition. In: 2023 International Conference on Computer Vision and Pattern Recognition (ICCVPR) (2023). IEEE

[29] Arora, Y., Raj, R., Kumar, A., Bajpai, S., Subhash, D.A., Sharma, N.: Attention on emotions: A vision transformer approach to advancing facial expression recognition. In: 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS) (2024). IEEE

[30] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)

[31] Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., Van Laerhoven, K.: Introducing wesad, a multimodal dataset for wearable stress and affect detection. Proceedings of the 20th ACM International Conference on Multimodal Interaction, 400–408 (2018)

[32] Goodfellow, I., Erhan, D., Carrier, L., Courville, A., Bengio, Y.: Challenges in Representation Learning: A Report on Three Machine Learning Contests. arXiv preprint arXiv:1307.0414 (2013)

[33] Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (ravdess). PloS one **13**(5), 0196391 (2018)

[34] Dupuis, K., Pichora-Fuller, M.K.: Toronto Emotional Speech Set (TESS) (2010)

[35] Weerasinghe, C.: Stress Detection by Keystroke/Mouse Changes Dataset (2023)