Intelligent Human Systems Integration (IHSI), Vol. 200, 2026, 617–627
https://doi.org/10.54941/ahfe1007120

Integrating Uncertainty-Aware Stress Detection With Spoken Dialogue-Based Interaction for Human-Centered Stress Management

Prachi Sheth, Jordan Schneider, and Teena Hassan

Hochschule Bonn-Rhein-Sieg University of Applied Sciences, Sankt Augustin, Germany

ABSTRACT

Stress significantly affects mental and physical health, contributing to anxiety, depression, and cardiovascular issues. Existing stress management tools, such as meditation apps or therapy, often rely on self-reports or rigid routines, making them less effective in real-life, dynamic situations. Physiological signals, such as heart rate, respiration rate, electrodermal activity, etc. provide objective markers for stress detection; however, inter-individual variability and signal noise introduce critical uncertainty into automated predictions. For trustworthy and reliable stress management, this work proposes a human-centered approach where the system's uncertainty is explicitly quantified and used to proactively solicit assistance from the user. We integrate physiological computing and machine learning with a dialogue system. Machine learning models, including random forest and convolutional neural networks trained on the WESAD dataset, classify stress versus non-stress states using features extracted from physiological signals. Experiments demonstrated that a 60-second window provided the best trade-off between temporal resolution and classification accuracy, with the random forest achieving 76% accuracy and the convolutional neural network achieving 75%. To account for uncertainty, entropy-based and Monte Carlo dropout methods estimate the confidence of each prediction. These confidence levels guide the dialogue manager, which tailors responses to users; high-confidence predictions trigger immediate, context-appropriate recovery interventions. While, low-confidence cases prompt a clarifying dialogue, actively enabling the user to confirm or correct the system's state prediction. Experiments validated the classification accuracy and, more importantly, demonstrated that uncertainty quantification successfully identified instances requiring this active human intervention. This system establishes a robust, human-in-the-loop cycle, offering a necessary setting for adaptive, personalized, and trustworthy stress management in critical healthcare and high-pressure settings.

Keywords: Physiological computing, Human-computer interaction, Stress management, Dialogue systems, Uncertainty quantification

INTRODUCTION

Stress represents a pervasive challenge to human well-being, severely impacting mental and physical health, productivity, and overall quality of life. Prolonged exposure is a known contributor to chronic conditions, including

Received October 10, 2025; Revised November 10, 2025; Accepted November 25, 2025; Available online February 1, 2026

anxiety disorders, depression, and cardiovascular issues. Consequently, developing effective, non-intrusive management strategies is crucial for both individual health and societal performance. Traditional methods for stress management, such as mindfulness applications, guided meditation, or scheduled therapy, are often beneficial but inherently limited (Huberty et al., 2019). They typically rely on retrospective self-reports or rigid routines, which fail to provide timely support during moments of acute, real-life stress. Advances in wearable physiological sensing offer a solution by providing objective, real-time data on the body's stress response via signals like Electrodermal Activity (EDA), Electrocardiogram (ECG), and Respiration Rate (RESP). However, this approach faces significant challenges: the considerable variance in physiological responses between individuals and the potential for unreliable predictions when signals are ambiguous or overlap with non-stress conditions (e.g., elevated heart rate from exercise or EDA changes from heat). Therefore, a multimodal approach that combines multiple signals is essential for capturing robust stress patterns and ensuring reliable support for the user (Bobade, 2020).

To address this need for reliable, real-time support, a dual-model strategy was employed for stress classification. A Random Forest (RF) classifier, using hand-crafted features, offers fast, lightweight, and interpretable inference suitable for real-time systems. Simultaneously, a 1D Convolutional Neural Network (CNN) is used to learn complex temporal patterns directly from raw signal windows, providing a potentially higher performance representation. The research utilized the publicly available WESAD (Wearable Stress and Affect Detection) dataset (see Table 1), chosen for its synchronized and labeled multimodal physiological data, enabling a comparative analysis of both approaches (Schmidt et al., 2018). The Leave-One-Subject-Out (LOSO) cross-validation method was applied to ensure the robustness and generalization of the models across subject variability (Bobade & Vani, 2020).

Table 1: WESAD dataset summary.

Features	Details
Dataset	WESAD
Label	Baseline, Stress, Amusement
No. of Subjects	Total - 17 (2 discarded)
Modalities	ECG, EMG, ACC, BVP, GSR, RESP
Study Environment	Laboratory

Classification uncertainty is a key challenge; physiological markers overlap with non-stress states, causing mislabels and unreliable feedback; without explicit handling, generic systems fail to adapt and engagement declines. To address this, we implement Uncertainty Quantification (UQ), the RF uses the entropy of predicted class probabilities, and the 1D CNN uses Monte Carlo Dropout to estimate a confidence distribution. Predicted

labels and confidence scores feed an integrated dialogue manager that adapts interactions, while the dialogue system provides personalized, human-like support that can enhance comfort and companionship beyond digital-only interfaces. By combining wearable sensing, uncertainty-aware machine learning, and dialogue, the system delivers real-time, personalized, and engaging stress support as an empathetic companion.

RELATED WORK

The integration of physiological signals, such as heart rate variability (HRV), EDA, and respiration rate with computational systems has been widely studied for stress monitoring and management. Prior work has highlighted the potential of using physiological data to design adaptive interventions, where stress detection enables timely support (Klęczek et al., 2024). Several studies emphasize the role of wearable sensors in capturing these signals to improve the accuracy of real-time stress detection (Abd Al-Alim et al., 2024) and (Verma & Tiwary, 2014). In particular, researchers have shown that combining multiple modal inputs provides better discrimination of stress states compared to unimodal inputs for real-time stress detection (Zhai et al., 2005), (Schmidt et al., 2018) and (Ghaderi et al., 2015).

Machine learning methods such as Support Vector Machine (SVM), k-Nearest Neighbours (kNN), Artificial Neural Network (ANN), reported high accuracy in around 99% in controlled and 97% in real-world environment (Ashwin et al., 2022). Deep learning methods, such as CNN and Long Short-Term Memory (LSTM) further advance performance by learning features directly from raw signals (Mane & Shinde, 2023). However, they are often computationally expensive, making them less practical for wearable or embedded systems. In contrast, traditional machine learning models like RF and SVMs provide faster, more interpretable results, making them suitable for real-time applications. Hybrid and multimodal approaches have also been explored, reaching accuracies as high as 94%, though at the cost of increased complexity (Elzeiny & Qaraqe, 2021).

One of the main challenges in stress classification is managing uncertainty in predictions. Bayesian methods have been proposed to explicitly capture uncertainty and improve the reliability of model outputs (de Berker et al., 2016). Beyond detection, studies on mental health chatbots, such as Tess highlight the potential of conversational systems to improve engagement and deliver psychological support at scale (Dosovitsky & Bunge, 2021).

METHODOLOGY

This study followed a structured methodology combining physiological signal processing, machine learning, uncertainty quantification, and conversational interaction design. The workflow began with data acquisition, continued with preprocessing and feature extraction, and concluded with stress classification, uncertainty handling, and dialogue integration.

Physiological Signals

Two physiological signals were used: EDA and Inter-Beat Interval (IBI). EDA reflects changes in skin conductance caused by sympathetic nervous system activity. It consists of a slow tonic component (skin conductance level) and faster phasic responses (skin conductance responses). Under stress, tonic levels and SCR frequency typically increase, reflecting heightened arousal (Hosseini et al., 2022) and (Fernandes et al., 2014). IBI represents the time interval between consecutive heartbeats, derived from the Blood Volume Pulse (BVP) signal (Elzeiny & Qaraqe, 2021). During stress, increased sympathetic activation shortens IBIs (Milstein & Gordon, 2020) and (Kyriakou et al., 2019), while reduced heart rate variability (HRV) indicates lower adaptability of the autonomic nervous system (Aygun et al., 2020). Combining EDA and IBI could potentially provide a more reliable stress estimate because they capture complementary physiological responses. EDA reflects sympathetic arousal, while IBI captures cardiac balance between sympathetic and parasympathetic activity. Together, they help distinguish stress from confounding factors like physical exertion.

Figure 1 and Figure 2 illustrate the raw and pre-processed physiological signals used in this study, including EDA and IBI derived from BVP. These visualizations highlight clear variations between baseline and stress conditions, showing increased stress periods. Such trends confirm that physiological responses captured by wearable sensors provide reliable indicators of stress, forming the basis for feature extraction and subsequent classification.

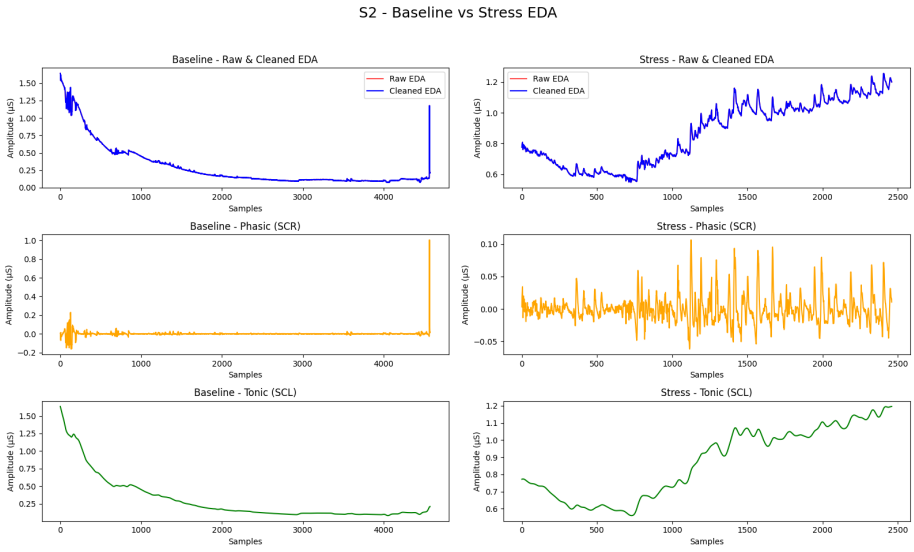

Figure 1: EDA plot of Subject 2; comparing baseline and stress conditions.

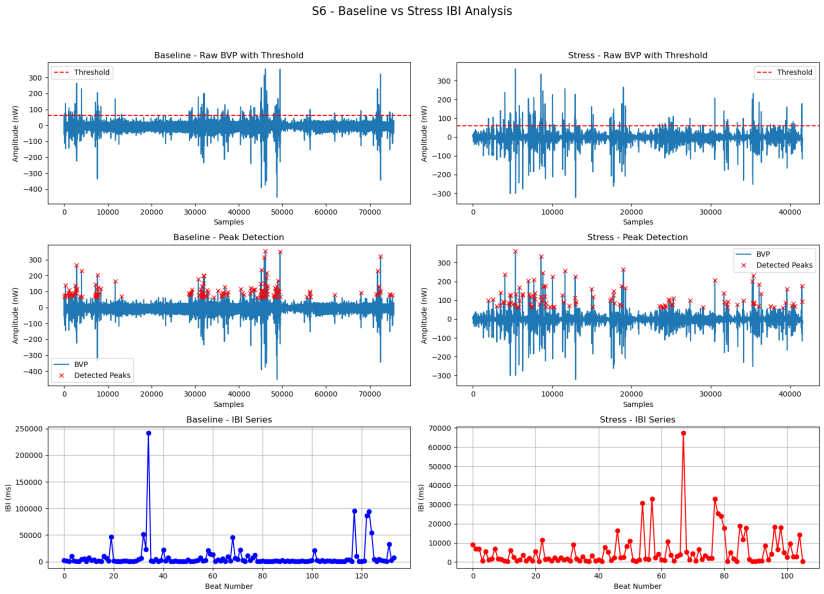

Figure 2: IBI plot of Subject 6; comparing baseline and stress conditions.

Dataset and Preprocessing

The WESAD dataset was used, which includes labeled segments for baseline, stress, amusement, and meditation (Schmidt et al., 2018). Only baseline and stress conditions were used, forming a binary classification problem. Physiological signals were collected with the Empatica E4 wristband (EDA - 4 Hz; BVP - 64 Hz). To align data, labels were downsampled to match EDA's sampling rate and interpolated for BVP. Signals were kept at their original frequencies to preserve data integrity. To capture both short- and long-term patterns, features were extracted using window sizes of 30–120 seconds.

Feature Extraction

EDA signals were pre-processed using NeuroKit2 (Makowski et al., 2021) for noise removal and decomposed into tonic and phasic components.

Table 2: Extracted EDA features.

EDA Features	Description
Mean, Std, Min, Max, Range	Basic signal statistics
Skewness, Kurtosis	Shape of the signal distribution
Number of SCR Peaks	Count of phasic responses
Mean SCR Amplitude	Average peak height
Tonic SD	Variation in the slow EDA component

For IBI, peaks were detected from BVP using an adaptive thresholding approach that accounts for individual variability (Jia et al., 2022). A total of 16 features extracted from EDA and IBI signals were used for evaluation as shown in Tables 2 and 3.

Table 3: Extracted IBI features.

IBI Features	Description
Mean, Std, Min, Max, Range, Median	Statistical measures of heartbeat intervals

Machine Learning Models

Stress classification was performed using an RF with mutual-information-based feature selection (top K = 5, 10, 15, total_features) (Deshpande & Ragha, 2023), where K represents no. of features. Random Forest was chosen for its robustness and interpretability (Campanella et al., 2023) and (El Haouij et al., 2018). To capture temporal dependencies, an Enhanced 1D CNN model was implemented using 16 combined EDA and IBI features as it can extract local patterns and combine them to complex interactions (Benita et al., 2024) and (Sánchez-Reolid et al., 2022). The architecture included convolutional layers, batch normalization, pooling, dropout, and fully connected layers. To address the significant class imbalance across the combined feature set (2983 baseline vs. 1832 stress instances), a dynamic class-weighting scheme was implemented within the LOSO cross-validation. Early stopping was used to avoid overfitting.

Uncertainty Quantification and Dialogue System

UQ enables confidence-aware, adaptive interaction via a Dialogue Manager. In Random Forests, uncertainty is the entropy of class probabilities aggregated across trees (Shaker & Hüllermeier, 2020); in the 1D CNN, it is estimated with Monte Carlo Dropout by running multiple stochastic forward passes to capture prediction variance (Gal & Ghahramani, 2015; Abdar et al., 2021). The Dialogue Manager uses these scores to decide engagement: high confidence (low uncertainty) prompts commitment to the predicted stress state, while low confidence triggers a request for user input or a deferred response. The system regulates its interactions by tracking the last response sent to a user, maintaining a counter of consecutive uncertain predictions to avoid query repetition, and monitoring a flag that indicates whether it is currently awaiting user feedback. Dialogue scenarios in Table 4 and Table 5 show that high confidence leads to proactive but non-intrusive support, whereas high uncertainty elicits a cautious, cooperative strategy; either continued silent observation or direct feedback requests, demonstrating how UQ enables adaptive, trustworthy human-centered interaction.

Table 4: High confidence (low uncertainty).

User State	Agent Utterance	User Response – Agent Action
Stressed	"You seem stressed. Would you like a short breathing exercise?"	Yes – starts breathing exercise. No – acknowledges and resumes monitoring.
Calm	"You seem calm - nice. Keep doing what you're doing!"	No response required – continues monitoring.

Table 5: Low confidence (high uncertainty).

User State	Agent Utterance	User Response – Agent Action
Uncertain	"I'm getting mixed signals… Let me observe a bit more." (1st–2nd uncertain reading)	No response required – continues monitoring.
	"I'm uncertain about your stress level. Can you please tell me how you feel?" (after 3+ uncertain readings)	Stressed – initiates breathing exercise. Calm – acknowledges and resumes monitoring.

EVALUATION

The evaluation compared two models: an RF with feature selection, and an Enhanced 1D CNN; integrated with uncertainty quantification. Models were tested using window sizes of 30s, 45s, 60s, 75s, 90s, and 120s. Performance metrics included Accuracy, Stress F1, Baseline F1, and predictive uncertainty.

While using RF, feature selection using mutual information was applied to improve computational efficiency and performance. Interestingly, feature selection did not always improve performance; in some cases, models with fewer selected features performed worse than those using more (see Table 6). While accuracy was the primary metric considered in this work, in practical deployment, considerations such as prediction time or computational efficiency may make feature selection more valuable, even when accuracy is only slightly affected.

Table 6: Random forest performance for different window sizes.

Window Size	K Features	Accuracy	Stress F1	Baseline F1
30	15	0.7385	0.7254	0.7505
45	15	0.7402	0.6954	0.7735
60	16	0.7650	0.7248	0.7950
75	15	0.7241	0.6429	0.7752
90	15	0.7142	0.6336	0.7658
120	10	0.7333	0.6486	0.7851

The 1D CNN captured temporal dependencies in the features and used the full feature set without selection.

Table 7: 1D CNN performance for different window sizes.

Window Size (s)	Accuracy	Stress F1	Baseline F1
30	0.7086	0.6958	0.7203
45	0.7525	0.7307	0.7710
60	0.7536	0.7095	0.7861
75	0.7034	0.6387	0.7485
90	0.6988	0.6321	0.7451
120	0.6103	0.5250	0.6696

Across models, as shown in Tables 6 and 7 combining EDA and IBI features with a non-overlapping 60s window gave the best performance, so 60s was adopted for UQ. UQ used predictive entropy for both RF and 1D CNN, with a 0.45 threshold to balance accurate stress/baseline classification and graceful handling of uncertainty. Tables 8 and 9 show RF made more confident correct predictions (124 correct & certain) than CNN (35), but also some confident errors (13 incorrect & certain), whereas CNN was more conservative, assigning many samples to uncertain (226 correct & uncertain). Thus, RF favors decisiveness with moderate risk, while 1D CNN favors caution, leading to more frequent but safer deferrals in an uncertainty-aware system.

Table 8: Uncertainty quantification results for the RF model.

Category	Count
Total	349
Correct & Certain	124
Correct & Uncertain	139
Incorrect & Certain	13
Incorrect & Uncertain	73

Table 9: Uncertainty quantification results for the 1D CNN model.

Category	Count
Total	349
Correct & Certain	35
Correct & Uncertain	226
Incorrect & Certain	1
Incorrect & Uncertain	87

CONCLUSION

This study demonstrates that physiological signals from wearable sensors can effectively distinguish between baseline and stress conditions, even with a relatively small dataset. Incorporating uncertainty quantification could be helpful for deciding when to engage users through the dialogue system, reducing the risk of unreliable feedback and improving trust. The findings highlight that the choice of window size and features strongly influences model performance, offering guidance for designing efficient stress classification frameworks. Overall, the work establishes a foundation for real-time, and uncertainty-aware stress monitoring systems. The main contribution lies in integrating uncertainty quantification into both classical and deep learning models, enabling confidence-aware predictions that guide adaptive interaction through the dialogue system. This uncertainty-driven approach connects machine learning outcomes with user engagement, supporting a more reliable and human-centered stress management process.

Future research should focus on improving adaptability of stress management through active learning with context awareness, allowing the model to query users when uncertainty is high and to use this feedback to improve future performance. Combining physiological data with behavioral or contextual information could enrich stress representation and improve accuracy of detecting stressful periods. Extending the framework for real-time use with wearable sensors in natural settings will help in evaluating the effectiveness and usability of the system in real world scenarios and applications.

ACKNOWLEDGMENT

This work has been funded by the German Aerospace Center as part of a cooperation between their Institute for AI Safety and Security and the Hochschule Bonn-Rhein-Sieg University of Applied Sciences. (Reference: Einzelprojektvereinbarung 67350636; December 2024).

REFERENCES

Abd Al-Alim, M., Mubarak, R., M. Salem, N., & Sadek, I. (2024). A machine-learning approach for stress detection using wearable sensors in free-living environments. Computers in Biology and Medicine, 179, 108918. https://doi.org/10.1016/j.compbiomed.2024.108918

Ashwin, V. H., Jegan, R., & Rajalakshmy, P. (2022). Stress Detection using Wearable Physiological Sensors and Machine Learning Algorithm. 2022 6th International Conference on Electronics, Communication and Aerospace Technology, 972–977. https://doi.org/10.1109/iceca55336.2022.10009326

Aygun, A., Ghasemzadeh, H., & Jafari, R. (2020). Robust Interbeat Interval and Heart Rate Variability Estimation Method From Various Morphological Features Using Wearable Sensors. IEEE Journal of Biomedical and Health Informatics, 24(8), 2238–2250. https://doi.org/10.1109/jbhi.2019.2962627

Benita, D. S., Ebenezer, A. S., Susmitha, L., Subathra, M. S. P., & Priya, S. J. (2024). Stress Detection Using CNN on the WESAD Dataset. 2024 International Conference on Emerging Systems and Intelligent Computing (ESIC), 308–313. https://doi.org/10.1109/ESIC60604.2024.10481604

Bobade, P., & Vani, M. (2020). Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data. 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 51–57. https://doi.org/10.1109/ICIRCA48905.2020.9183244

Campanella, S., Altaleb, A., Belli, A., Pierleoni, P., & Palma, L. (2023). A Method for Stress Detection Using Empatica E4 Bracelet and Machine-Learning Techniques. Sensors, 23(7), 3565. https://doi.org/10.3390/s23073565

de Berker, A. O., Rutledge, R. B., Mathys, C., Marshall, L., Cross, G. F., Dolan, R. J., & Bestmann, S. (2016). Computations of uncertainty mediate acute stress responses in humans. Nature Communications, 7(1). https://doi.org/10.1038/ncomms10996

Deshpande, H. S., & Ragha, L. (2023). A hybrid random forest-based feature selection model using mutual information and F-score for preterm birth classification. International Journal of Medical Engineering and Informatics, 15(1), 84. https://doi.org/10.1504/ijmei.2023.127257

Dosovitsky, G., & Bunge, E. L. (2021). Bonding With Bot: User Feedback on a Chatbot for Social Isolation. Frontiers in Digital Health, 3. https://doi.org/10.3389/fdgth.2021.735053

El Haouij, N., Poggi, J.-M., Ghozi, R., Sevestre-Ghalila, S., & Jaïdane, M. (2018). Random forest-based approach for physiological functional variable selection for driver's stress level classification. Statistical Methods & Applications, 28(1), 157–185. https://doi.org/10.1007/s10260-018-0423-5

Elzeiny, S., & Qaraqe, M. (2021). Automatic and Intelligent Stressor Identification Based on Photoplethysmography Analysis. IEEE Access, 9, 68498–68510. https://doi.org/10.1109/ACCESS.2021.3077358

Fernandes, A., Helawar, R., Lokesh, R., Tari, T., & Shahapurkar, A. V. (2014). Determination of stress using Blood Pressure and Galvanic Skin Response. 2014 International Conference on Communication and Network Technologies, 165–168. https://doi.org/10.1109/CNT.2014.7062747

Ghaderi, A., Frounchi, J., & Farnam, A. (2015). Machine learning-based signal processing using physiological signals for stress detection. 2015 22nd Iranian Conference on Biomedical Engineering (ICBME), 93–98. https://doi.org/10.1109/icbme.2015.7404123

Hosseini, E., Fang, R., Zhang, R., Parenteau, A., Hang, S., Rafatirad, S., Hostinar, C., Orooji, M., & Homayoun, H. (2022). A Low Cost EDA-based Stress Detection Using Machine Learning. 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2619–2623. https://doi.org/10.1109/BIBM55620.2022.9995093

Huberty, J., Green, J., Glissmann, C., Larkey, L., Puzia, M., & Lee, C. (2019). Efficacy of the Mindfulness Meditation Mobile App "Calm" to Reduce Stress Among College Students: Randomized Controlled Trial. JMIR mHealth and uHealth, 7(6), e14273. https://doi.org/10.2196/14273

Jia, D., Zhang, X., Zhou, J. T., Lai, P., & Wei, Y. (2022). Dynamic thresholding for video anomaly detection. IET Image Processing, 16(11), 2973–2982. https://doi.org/10.1049/ipr2.12532

Klęczek, K., Rice, A., & Alimardani, M. (2024). Robots as Mental Health Coaches: A Study of Emotional Responses to Technology-Assisted Stress Management Tasks Using Physiological Signals. Sensors, 24(13), 4032. https://doi.org/10.3390/s24134032

Kyriakou, K., Resch, B., Sagl, G., Petutschnig, A., Werner, C., Niederseer, D., Liedlgruber, M., Wilhelm, F., Osborne, T., & Pykett, J. (2019). Detecting Moments of Stress from Measurements of Wearable Physiological Sensors. Sensors, 19(17), 3805. https://doi.org/10.3390/s19173805

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. H. A. (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. Behavior Research Methods, 53(4), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y

Mane, S. A. M., & Shinde, A. (2023). StressNet: Hybrid model of LSTM and CNN for stress detection from electroencephalogram signal (EEG). Results in Control and Optimization, 11, 100231. https://doi.org/10.1016/j.rico.2023.100231

Milstein, N., & Gordon, I. (2020). Validating Measures of Electrodermal Activity and Heart Rate Variability Derived From the Empatica E4 Utilized in Research Settings That Involve Interactive Dyadic States. Frontiers in Behavioral Neuroscience, 14. https://doi.org/10.3389/fnbeh.2020.00148

Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., & Van Laerhoven, K. (2018). Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. Proceedings of the 20th ACM International Conference on Multimodal Interaction, 400–408. https://doi.org/10.1145/3242969.3242985

Sánchez-Reolid, R., López de la Rosa, F., López, M. T., & Fernández-Caballero, A. (2022). One-dimensional convolutional neural networks for low/high arousal classification from electrodermal activity. Biomedical Signal Processing and Control, 71, 103203. https://doi.org/10.1016/j.bspc.2021.103203

Verma, G. K., & Tiwary, U. S. (2014). Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. NeuroImage, 102, 162–172. https://doi.org/10.1016/j.neuroimage.2013.11.007

Zhai, J., Barreto, A. B., Chin, C., & Li, C. (2005). Realization of stress detection using psychophysiological signals for improvement of human-computer interactions. Proceedings. IEEE SoutheastCon, 2005., 415–420. https://doi.org/10.1109/SECON.2005.1423280