Towards Personalisable Stress Management using Physiological Computing and Conversational Robots

Prachi Sheth

***Abstract*—Stress is a major challenge for both mental and physical health, and managing it effectively requires personalized support. Physiological signals offer a non-invasive way to monitor stress and provide timely feedback, but variability in these signals and uncertainty in model predictions make reliable stress detection difficult. This study presents a multimodal, uncertainty-aware stress management system that combines physiological computing with a conversational agent. The system uses two signals, Electrodermal Activity and Inter-Beat Intervals; to classify into stress and non-stress states. Two models were tested, i.e., a Random Forest classifier and a one-dimensional Convolutional Neural Network. To handle uncertainty, the models incorporate Uncertainty Quantification methods, which flag predictions made with low confidence. Experiments across different time windows showed that both models performed best at a 60-second window, where the Random Forest reached 76% accuracy and the Convolutional Neural Network reached 75% accuracy. The system also includes a dialogue manager that uses model predictions to provide users with adaptive stress management strategies, such as breathing exercises. These results show the potential of combining physiological sensing, machine learning with uncertainty handling, and conversational support to create personalized stress management solutions.**

***Index Terms*—Physiological Computing, Uncertainty Quantification, Conversational Agents.**

I. Introduction

Stress is one of the most common challenges affecting people's mental and physical health, with significant impacts on overall well-being and daily productivity. Prolonged stress can lead to anxiety, depression, cardiovascular problems, and reduced quality of life [1], [2]. Therefore, managing stress effectively is not only important for individuals but also for society as a whole.

Traditional methods for stress management, such as mindfulness apps, guided meditation, or therapy, can be helpful, but they often rely on self-reporting or scheduled activities [3], [4]. This limits their ability to provide real-time support when stress actually occurs. Advances in wearable sensors and machine learning have enabled the detection of stress through physiological signals like Electrodermal Activity (EDA), Electrocardiogram (ECG), Heart Rate (HR), etc [5]. These signals provide objective data that can be used to detect stress in real time. However, there are challenges, such as variance of physiological signals between individuals, and machine learning models may produce uncertain or unreliable predictions if these variations are not handled carefully. A multimodal approach, which combines multiple signals, is therefore more effective in capturing stress patterns and reducing uncertainty in model predictions [6], [7].

To investigate this, a dual-method approach was employed; a deep learning model for high-performance stress prediction using raw signals, and a machine learning model for real-time, lightweight inference using extracted features. While deep learning models, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) provide high accuracy and automatic feature learning, they require significant computational resources, making them less practical for embedded or wearable systems. In contrast, machine learning models like Support Vector Machine (SVM) and Random Forest (RF) offer faster training and inference, are interpretable, and are more suitable for real-time applications [8].

For evaluation, publicly available datasets were taken into consideration, the WESAD (Wearable Stress and Affect Detection) dataset was chosen for its synchronized and labeled multimodal physiological data, making it well-suited for a comparative analysis of both approaches [6]. The Leave-One-Subject-Out (LOSO) cross-validation approach was applied to account for subject variability and ensure robustness [7]. Model performance was measured using accuracy and F1-scores, with experiments conducted across different window sizes to determine which configuration works best when EDA and inter-beat intervals (IBI) signals are used together.

TABLE I: WESAD Dataset Summary

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

At the same time, conversational agents and robots offer new opportunities for personalized stress support. Unlike mobile apps or chatbots alone, robots can interact with people in shared social spaces, mimicking human gestures and behaviors, and creating a stronger sense of comfort and companionship [9], [10]. Research has shown that people often respond more positively to robots than to digital interfaces such as tablets, especially in healthcare contexts [11], [12]. This makes robots promising companions for delivering stress management interventions in a more engaging and human-like way. Building on this, the system integrates uncertainty-aware machine learning models with a dialogue manager, allowing not only stress detection but also confidence estimation in

*Submitted to the Department of Computer Science at Hochschule Bonn-Rhein-Sieg in partial fulfilment of the requirements for the degree of Master of Science in Autonomous Systems

†Supervised by Prof. Dr. Teena Hassan (HBRS) and Jordan Schneider (HBRS)

‡Submitted in September 2025

predictions. When stress is detected, the system can provide adaptive interventions such as guided breathing exercises through either a chatbot or a socially interactive robot.

By combining wearable sensing, machine learning with uncertainty quantification, and conversational agents, this research aims to explore how technology can support people in managing stress in real-time. The ultimate goal is to make stress management more accessible, personalized, and effective, and robots serve as supportive, empathetic companions.

A. Motivation

Physiological signals are a powerful resource for detecting stress because they capture the body's natural responses to challenging situations in a non-invasive way. Changes in signals, such as heart rate, skin conductance, or respiration provide objective markers that are often more reliable than self-reported data, which can be biased or delayed. However, stress detection based on these signals is not always straightforward. For example, an elevated heart rate alone may not necessarily indicate stress; it could simply reflect physical activity, such as running up a lot of stairs. This highlights the importance of a multimodal approach that integrates multiple signals for robust stress classification. Physiological responses vary not only between individuals but also within the same person across different contexts. Without accounting for this variability, models risk making unreliable predictions. Incorporating uncertainty quantification helps identify when the model lacks confidence in its output. Effective stress management also requires meaningful interaction. Here, a dialogue manager plays a crucial role by engaging with users in natural conversation and offering adaptive support strategies, such as guided breathing exercises.

B. Problem Statement

Most existing systems rely heavily on self-reports or scheduled activities, which fail to capture stress in real time and miss opportunities for timely intervention. Furthermore, these tools often adopt a generic, one-size-fits-all approach, offering little adaptation to individual needs and resulting in reduced user engagement over time. A critical challenge in stress classification lies in the uncertainty of physiological signals. Stress responses vary greatly across individuals, and physiological markers often overlap with non-stress conditions. For example, changes in EDA, which measures skin conductance, may increase not only during psychological stress but also due to physical factors, such as heat or movement. Similarly, sometimes stress, is difficult to distinguish from non-stress. Without mechanisms to handle uncertainty, classification models risk mislabeling stress and delivering unreliable feedback to users.

C. Proposed Approach

The approach includes below modules:

- Stress Classification: A dual-model strategy was employed. A RF classifier is used for its robustness with hand-crafted features and innate ability to provide probability estimates. A 1D CNN is used to learn temporal patterns directly from preprocessed signal windows, offering a different and potentially more powerful representation.
- Uncertainty Quantification (UQ): Model-specific UQ techniques were implemented. For the RF, confidence is derived from the entropy of the predicted class probabilities. For the 1D CNN, Monte Carlo Dropout was used, performing multiple stochastic forward passes during inference to obtain a confidence distribution.
- The predicted stress label and its confidence score are directed to a dialogue manager:
  - High Confidence: The system proceeds with a context-appropriate intervention (e.g., "You seem stressed. Would you like a short breathing exercise?").
  - Low Confidence: The system defaults to a safe, clarifying dialogue to disambiguate the user's state (e.g., "I am uncertain about your stress level. Can you please tell me how you feel?"). This ensures the system remains helpful and never imposes an incorrect intervention.

II. 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 [9]. Several studies emphasize the role of wearable sensors in capturing these signals to improve the accuracy of real-time stress detection [13], [14]. 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 [15].

Machine learning approaches have become central to stress classification tasks. Studies demonstrate that algorithms such as, SVM, k-Nearest Neighbors (kNN), and Artificial Neural Networks (ANN) can achieve high prediction accuracy up to 99% in controlled environments and around 97% in real-world settings [16]. Wearable sensors capturing electroencephalogram (EEG), ECG, and EDA have been used effectively with these models, reinforcing the role of multimodal physiological data in stress research [17]. Deep learning methods, such as CNN and LSTM further advance performance by learning features directly from raw signals [18]. 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 [8].

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 [19]. Other approaches employ multimodal data fusion, leveraging complementary signals to reduce ambiguity in classification tasks. These techniques help avoid misclassifications when stress is difficult to differentiate [6].

Beyond detection, conversational agents have been applied to support stress management, particularly among populations vulnerable to social isolation [20]. For instance, conversational AI systems have been shown to provide companionship and emotional support to older adults, reducing feelings of loneliness. Similarly, studies on mental health chatbots, such as Tess highlight the potential of conversational systems to improve engagement and deliver psychological support at scale [21]. More recently, researchers have begun exploring the role of social robots as embodied companions for stress management. Studies suggest that robots can provide engaging, empathetic, and interactive support, which may be more effective than traditional interfaces, such as tablets or mobile apps. For example, interventions using robots with university students have demonstrated positive physiological responses, indicating the promise of robotic embodiment in stress management interventions [9].

III. Background

Stress is a multidimensional experience involving both psychological and physiological components. Physiological sensors are devices that measure biological signals generated by the body, providing objective insights into a person's physical and psychological state [22]. Unlike self-reports, which are subjective and prone to bias, physiological signals offer real-time and continuous data that can be leveraged to detect stress. They are also non-invasive, as most measurements can be done with wearable or contactless sensors [23]. Commonly used physiological signals include EDA, heart rate and HRV derived from ECG or IBI, respiration rate, and EEG. These signals change in characteristic ways under stress. Wearable devices, such as the Empatica E4 wristband or chest-mounted sensors allow these signals to be collected in daily life. Most importantly, physiological signals have the benefit of being involuntarily triggered, making them difficult for a person to consciously control or manipulate [24]. By analyzing individual patterns and responses, systems can tailor stress management techniques to the user, enabling timely interventions that go beyond retrospective questionnaires or fixed check-ins. Despite their usefulness, physiological signals are inherently noisy and context-dependent. For example, elevated EDA levels may indicate stress, but they could also be caused by environmental temperature or physical activity. This introduces uncertainty in stress classification, where the system may be unsure whether a given pattern of signals reflects stress or another state. UQ refers to methods that capture and express this uncertainty in predictions rather than providing overly confident classifications [26]. Two main types of uncertainty are generally considered:

- Aleatoric Uncertainty: Arises from variability in the data itself, such as sensor noise or individual differences [27].

Fig. 1: Common Physiologicals sensors and wearable devices [25]

- Epistemic Uncertainty: Caused from limited knowledge of the model, such as insufficient training data or unobserved conditions. [27].

In this study, epistemic uncertainty is particularly important. Stress manifests differently across individuals and contexts, and no dataset can capture all possible variations. By modeling epistemic uncertainty, the system can identify cases where it is less confident, avoid misclassifications, and adaptively seek more data or clarification.

IV. Methodology

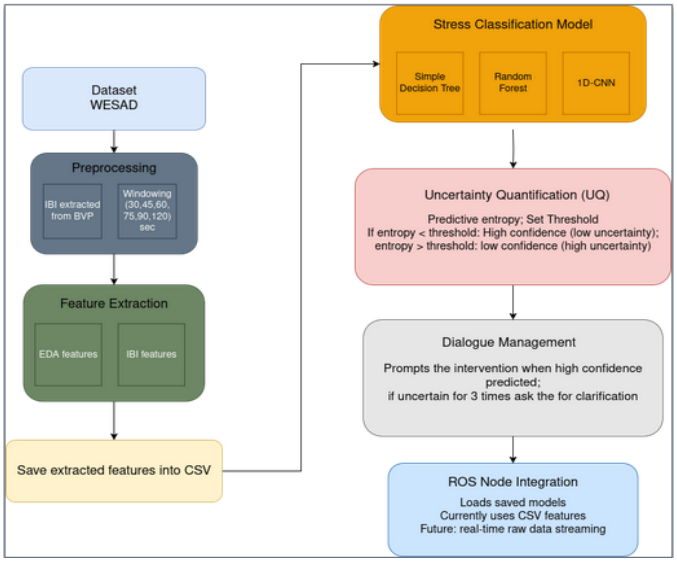

Fig. 2: Overview of the proposed stress detection and intervention pipeline

This study followed a structured methodology that combined physiological signal processing, machine learning, and conversational interaction design as shown in Fig. (2). The overall workflow began with signal acquisition, continued with dataset selection and preprocessing, and concluded with stress classification, uncertainty handling, and dialogue integration.

For stress detection, two physiological signals were chosen: EDA and IBI. EDA reflects changes in skin conductance driven by sweat gland activity, which is tightly regulated by the sympathetic branch of the autonomic nervous system. It

is measured in microsiemens (μS) and is widely recognized as a reliable indicator of psychological arousal, including stress [28], [5]. The EDA signal consists of two main components: a slowly varying tonic level, also called the skin conductance level (SCL), and faster phasic responses known as skin conductance responses (SCRs). Under stress, the tonic level typically rises, reflecting heightened sympathetic activation, while the frequency, amplitude, etc. of SCRs increases, capturing moment-to-moment fluctuations in arousal. These combined properties make EDA highly sensitive to emotional and cognitive stressors [29].

IBI, on the other hand, represents the time interval between consecutive heartbeats and which is here derived from the Blood Volume Pulse (BVP) signal using peak detection [30], [8]. As stress activates the sympathetic nervous system and suppresses parasympathetic activity, heart rate generally increases, leading to shorter IBIs [31], [32], [33]. Beyond mean heart rate, variability in IBIs commonly known as HRV is also informative. Stress is often associated with reduced HRV [32], reflecting a diminished ability of the autonomic nervous system to adapt to changing demands. Features such as the standard deviation of IBIs, median IBI, etc. help quantify these stress-related changes robustly, reducing sensitivity to outliers.

The combination of EDA and IBI enhances the robustness of stress detection because the two signals capture complementary aspects of physiological stress responses. While EDA is a direct marker of sympathetic arousal and emotional reactivity, IBI reflects the balance between sympathetic and parasympathetic regulation of cardiac activity. For example, an elevated heart rate alone could also indicate physical exertion, but if this is not accompanied by elevated EDA responses, it is less likely to be stress-induced. By integrating these signals, the system can more accurately differentiate stress from confounding factors like physical activity or transient arousal, thereby improving classification performance and mitigating uncertainty.

The dataset selected for this study was WESAD, which is a well-established multimodal dataset for stress research. The dataset provides labeled segments corresponding to different states: baseline (neutral), stress (induced using the Trier Social Stress Test), amusement (watching funny videos), and meditation (guided relaxation). For this project, the focus was restricted to baseline and stress conditions, which were mapped into a binary classification task: non-stress and stress. To ensure subject-independent evaluation, the LOSO cross-validation method was used, where the model was trained on data from all but one subject and then tested on the excluded subject. This process was repeated for each participant to validate the model's generalizability. This multimodal physiological signals were collected using the Empatica E4 wristband [34]. This device records EDA at a sampling rate of 4 Hz and BVP at 64 Hz, enabling the derivation of stress-related physiological measures. Since the dataset provides ground truth in the form of discrete labels marking baseline, stress, and other conditions, it was necessary to align these labels with the continuous sensor data. For the EDA pipeline, the label stream was downsampled to match the length of the EDA signal, ensuring proper alignment of conditions with the 4 Hz sampling rate. For the BVP pipeline, the labels were interpolated (upsampled) to match the higher 64 Hz sampling rate of the BVP signal. Importantly, the signals themselves were not resampled and were kept at their native recording frequencies to preserve data integrity.

To capture both short-term and long-term changes in physiology, feature extraction was conducted on multiple window sizes (30, 45, 60, 75, 90, and 120 seconds), with non-overlapping windows for consistency. For EDA, the raw signals were first segmented into baseline and stress segments and then preprocessed using the NeuroKit2 to remove noise and artifacts. From each window, features were extracted as provided below:

TABLE II: Extracted EDA Features

Feature Name	Description	Unit
eda_mean	Average skin conductance in the window	μS
eda_std	Standard deviation of EDA	μS
eda_min	Minimum EDA value in the window	μS
eda_max	Maximum EDA value in the window	μS
eda_range	Range of EDA values (max - min)	μS
eda_skewness	Skewness of EDA distribution	no unit
eda_kurtosis	Kurtosis of EDA distribution	no unit
SCR_Peaks_N	Number of skin conductance responses (phasic peaks)	count
SCR_Peaks_Amplitude_Mean	Mean amplitude of SCR peaks	μS
EDA_Tonic_SD	Standard deviation of the tonic EDA component	μS

Visualization of EDA data was performed by comparing raw and cleaned signals, as well as tonic and phasic components across baseline and stress conditions as shown in Fig. (3), allowing a clear inspection of stress-induced changes in skin conductance. For IBI, the features were extracted indirectly

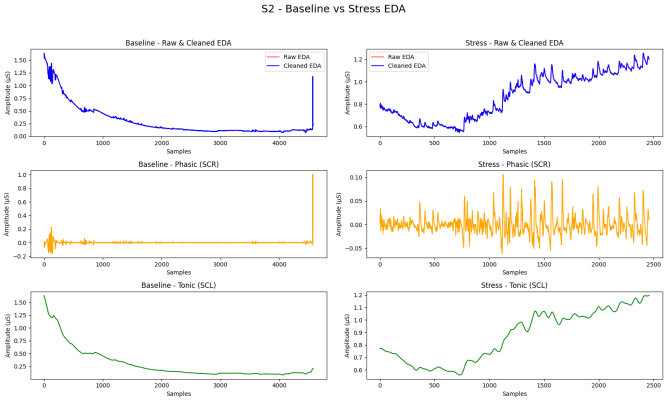

Fig. 3: EDA Plot of Subject 2; comparing baseline and stress conditions

from the BVP signal, a dynamic thresholding approach was

employed for peak detection rather than using a fixed threshold across all subjects [35]. Since BVP waveforms vary significantly between individuals due to differences in physiology and signal quality, a personalized threshold was necessary to ensure robust detection of heartbeats. The threshold was computed as:

$$T = \text{Median}(BVP) + k \times (Q_3 - \text{Median}(BVP)) \quad (1)$$

where Q_3 represents the 75th percentile of the BVP distribution and k is a scaling factor. The use of the median as the baseline reference makes the threshold more robust to outliers and transient spikes compared to the mean, which could be heavily influenced by noise. Incorporating the interquartile range (Q_3 - Median) allows the threshold to adapt to the variability of each subject's BVP signal. As a result, peaks corresponding to actual heartbeats could be detected more reliably, avoiding false positives caused by noise or motion artifacts while still being sensitive to genuine physiological variations. From the detected peaks, inter-beat intervals were computed, and statistical features were calculated for each window:

TABLE III: Extracted IBI Features

Feature Name	Description	Unit
ibi_mean	Average inter-beat interval in the window	seconds (s)
ibi_std	Standard deviation of IBI values	seconds (s)
ibi_min	Minimum IBI observed in the window	seconds (s)
ibi_max	Maximum IBI observed in the window	seconds (s)
ibi_range	Range of IBI values (max - min)	seconds (s)
ibi_median	Median IBI in the window	seconds (s)

Visualization included plots of raw BVP with threshold lines, peak detection, and IBI series representations for baseline and stress states as shown in Fig. (4), which illustrated the differences in beat-to-beat variability between conditions.

For the stress classification task, multiple machine learning models were implemented to analyze the extracted physiological features and evaluate their ability to discriminate between baseline and stress conditions. The workflow progressed from a simple baseline model to more sophisticated models incorporating feature selection and deep learning, allowing both interpretability and the modeling of complex temporal patterns in the data. Initially, a simple Decision Tree classifier was implemented as a baseline to evaluate the feasibility of distinguishing baseline and stress conditions using the combined EDA and IBI features. Decision Trees were chosen because they are computationally efficient, easy to interpret, and provide a quick assessment of whether the extracted features contain meaningful information for classification [6]. They also serve as a reference point to compare the performance gains achieved by more advanced models. The dataset was first loaded and merged across EDA and IBI features for each window size (30, 45, 60, 75, 90, 120 seconds). Labels were encoded numerically with baseline = 0 and stress = 1. To ensure a subject-independent evaluation, a LOSO cross-validation strategy was applied. This approach prevents information leakage between training and test sets and provides realistic estimates of how the model would perform on unseen subjects. The Decision Tree model was configured with a maximum depth of 5 and a minimum leaf sample of 5 to prevent overfitting while maintaining interpretability. This baseline analysis provided insights into feature relevance and the discriminative power of the extracted features across different temporal window sizes, establishing an initial benchmark for subsequent models.

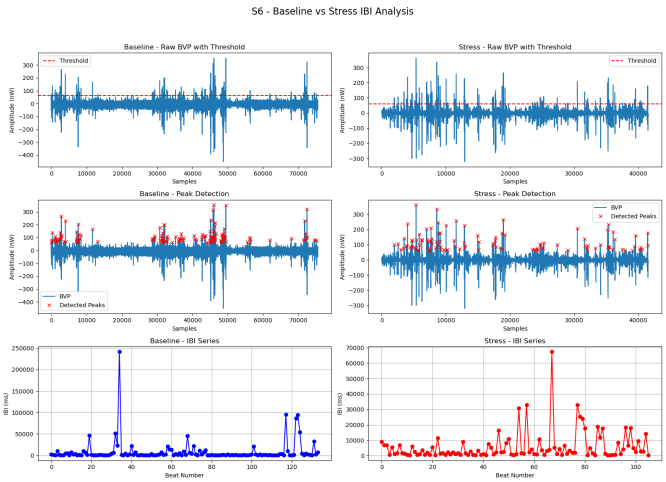

Fig. 4: IBI plot of Subject 6; comparing baseline and stress conditions

To improve classification performance and incorporate feature selection, a Random Forest classifier was employed in conjunction with global top-K feature selection using mutual information. Random Forest was selected due to its robustness, ability to handle non-linear interactions between features, and provision of interpretable feature importance measures [36], [6], [37]. The feature selection step involved standardizing the feature matrix and retaining only the most informative features for training, with K-values of 5, 10, 15, and the total number of combined features [38]. By systematically varying both the window size and the number of selected features, this model enabled a detailed analysis of how classification performance is influenced by temporal granularity and feature dimensionality. The resulting accuracy versus number of features plot allowed visualization of the impact of feature selection on model performance as showed in Fig. (6) Furthermore, the Random Forest model provided insights into the physiological markers most predictive of stress, highlighting key EDA and IBI features and their relative contributions to classification. This interpretability is particularly valuable for understanding the physiological underpinnings of stress.

For a more advanced approach capable of capturing subtle temporal patterns, 1D CNN, referred to as the Enhanced 1D CNN, was designed [39]. This model takes as input a 1D vector of 16 features, representing the combined set of extracted EDA and IBI features per sample. CNNs are well-

suited for sequential data because their convolutional layers can extract local patterns and hierarchically combine them to model complex interactions [40], [41]. The architecture consisted of two convolutional layers with batch normalization to stabilize training, followed by an adaptive average pooling layer to reduce the spatial dimension of the feature maps. A dropout layer was included to mitigate overfitting, and two fully connected layers mapped the extracted features to a binary output representing baseline or stress [42]. By treating the feature vectors as one-dimensional sequences, the CNN could learn temporal dependencies and subtle patterns within the physiological signals that might not be captured by classical machine learning models. Training was performed using early stopping on a validation split of the LOSO training set to prevent overfitting. The network was then retrained for a number of epochs equal to the best early-stopped epoch using the full train+validation set to leverage all available data. Class weights were applied in the cross-entropy loss function to account for potential imbalance between baseline and stress samples, ensuring that the network did not favor the majority class. A schematic diagram of the CNN architecture illustrates the convolutional layers, pooling, dropout, and fully connected layers, emphasizing the flow of feature extraction and classification.

In addition to standard classification, UQ was incorporated into both the RF and Enhanced 1D CNN models to assess the confidence of predictions and guide adaptive dialogue interventions. UQ provides critical information on when the model is unsure about its predictions, enabling the system to interact intelligently with users rather than issuing potentially misleading guidance. Different models require different UQ methods due to their inherent characteristics: for RF, which is an ensemble of decision trees, predictive probabilities naturally provide an estimate of uncertainty, making entropy over class probabilities an effective measure [43], [44]. In contrast, the CNN is a deep neural network where standard softmax outputs tend to be overconfident. Therefore, Monte Carlo (MC) Dropout is used during inference to approximate Bayesian uncertainty, allowing the network to capture epistemic uncertainty arising from limited data and model capacity [45], [46]. By tailoring UQ methods to each model type, the system obtains reliable confidence estimates suitable for triggering adaptive and trust-aware dialogues.

In a Random Forest classifier, each tree in the ensemble casts a vote for a class label, and the proportion of trees voting for each class forms the predicted probability vector for a given sample. For instance, if 70% of the trees vote for stress and 30% vote for baseline, the probability distribution is [0.3,0.7]. The uncertainty of the prediction was measured using Shannon entropy:

$$H(\mathbf{p}_i) = -\sum_c p_{i,c} \log p_{i,c} \tag{2}$$

where p_i is the predicted probability, c indexes the output classes; baseline and stress. The logarithmic term is crucial because it penalizes low probabilities and emphasizes the contribution of intermediate probabilities; distributions where one class dominates result in low entropy meaning high confidence, whereas nearly uniform distributions across

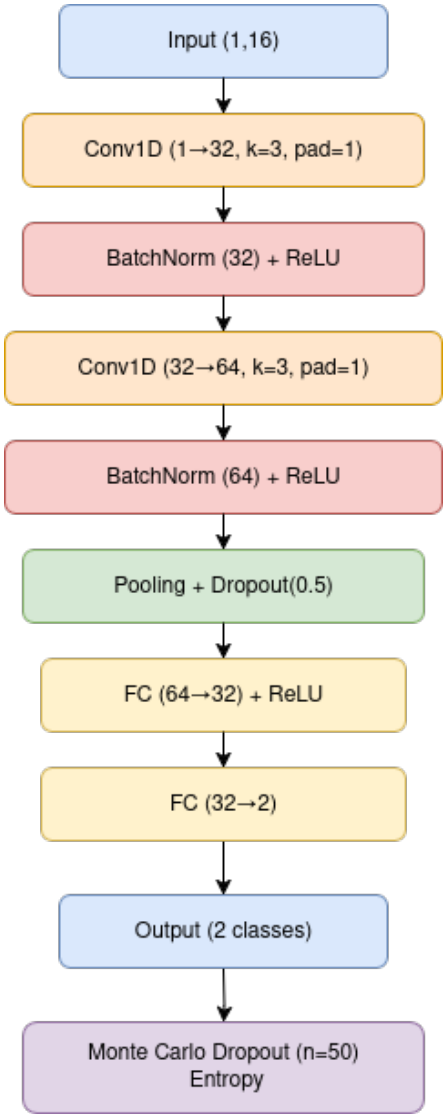

Fig. 5: Block architecture of the proposed 1D Convolutional Neural Network (1D CNN) for binary classification with UQ.

classes yield high entropy, i.e., low confidence. This allows the system to identify predictions that the Random Forest is uncertain about and flag them for adaptive intervention. The entropy values were plotted against the prediction accuracy for different thresholds to determine, which value is suitable for further analysis.

For the Enhanced 1D CNN, UQ is achieved using MC Dropout, which allows the network to estimate predictive uncertainty by performing multiple stochastic forward passes at inference time. During training, standard cross-entropy loss is applied with class weights to address potential imbalance:

$$L_{CE} = \frac{1}{N} \sum_{i=1}^{N} w_{y_i} \log p_{i,y_i} \tag{3}$$

where N is the number of samples, y_i is the true label, p_{i,y_i} is the predicted probability for the correct class, and w_{y_i} is the class weight for label y_i. Dropout is applied during training and also at inference, simulating sampling from an approxi-

mate posterior over the network's weights. Multiple stochastic forward passes (T samples) were performed for input, and the predictive probabilities were averaged to estimate the expected probability:

$$\mathbf{p}_i = \frac{1}{T} \sum_{t=1}^{T} \mathbf{y}_i^{(t)} \tag{4}$$

The predictive entropy is then computed as a measure of uncertainty; in the same way as Eq. (2). In both cases, entropy serves as a natural measure of uncertainty, but MC Dropout is particularly suited for deep learning models, capturing uncertainty arising from learned weights.

The Dialogue Manager is a core component designed to facilitate context-aware, adaptive interactions with the user based on both the predicted stress state and the model's uncertainty in its predictions. Its primary purpose is to decide when and how the system should communicate with the user, ensuring that the agent provides helpful guidance without issuing misleading or intrusive prompts. The Dialogue Manager uses the entropy values coming from the UQ to determine the confidence level of a prediction:

- High confidence (low entropy): The model is confident about the user's state.
- Low confidence (high entropy): The model is uncertain and requires adaptive dialogue interventions.

This allows the agent to modulate its responses based on the model's certainty, improving trust and user engagement. For each user, the Dialogue Manager maintains a state that includes the last response sent, the number of consecutive uncertain predictions; which is necessary so that the agent does not frequently ask the user for the uncertain state, a flag indicating whether it is awaiting user feedback, and a history of feedback received. Based on these states and the current prediction, the agent follows predefined interaction policies.

This Dialogue Manager is also capable of handling explicit user feedback. If the user confirms they are stressed, the agent offers targeted interventions, such as a breathing exercise. If the user indicates they are calm, the agent acknowledges this and continues monitoring without intervention. Feedback is logged to update the user state and guide future interactions, allowing the system to adapt dynamically to each individual's stress pattern. Formally, the response function of the Dialogue Manager can be represented as a mapping from entropy, predicted label, and user state to an agent response. By combining model predictions, entropy-based uncertainty, and user feedback, the Dialogue Manager creates a human-in-the-loop system that delivers trust-aware, personalized interventions while minimizing unnecessary or incorrect prompts. Table IV and Table V shows the flow of the dialogue from agent and user end.

TABLE IV: Dialogue Manager Response Logic

Condition	Agent's Response & Action
High Confidence + Calm (Low entropy, label = 0)	Response: "You seem calm - nice. Keep doing what you're doing!" Action: Do not wait. Continues monitoring.
High Confidence + Stressed (Low entropy, label = 1)	Response: "You seem stressed. Would you like a short breathing exercise?" Action: Waits for user feedback (yes/no).
Low Confidence (Uncertain, 1st or 2nd time)	Response: "I'm getting mixed signals... Let me observe a bit more." Action: Do not wait. Continues monitoring; counts uncertainty.
Low Confidence (Uncertain, 3rd+ time in a row)	Response: "I am uncertain about your stress level. Can you please tell me how you feel?" Action: Waits for user feedback (stress/calm).

TABLE V: User Feedback Handling

User Response	Agent Response
yes/no	yes: "Great! Let's begin with a breathing exercise..." no: "Okay, no problem..."
stressed/calm	stressed: Offers breathing exercise. calm: "Nice! Keep up the good work."

V. Evaluation

The evaluation was conducted across three models: a simple Decision Tree baseline, a Random Forest classifier with feature selection, and an Enhanced 1D CNN with MC Dropout for uncertainty quantification. Each model was tested using different window sizes (30s, 45s, 60s, 75s, 90s, and 120s) to analyze how temporal resolution influenced classification performance. Accuracy, F1-scores for both stress and baseline classes, and the reliability of predictions under uncertainty were used as the main evaluation metrics.

The Simple Decision Tree as a baseline provided a first indication of the separability between stress and baseline states. While its performance was modest, it served as an interpretable benchmark, showing stable but limited accuracy across windows.

TABLE VI: Simple Decision Tree Performance for Different Window Sizes

Window Size (s)	Accuracy	Stress F1	Baseline F1
30	0.6788	0.6682	0.6447
45	0.6810	0.5966	0.7013
60	0.7007	0.6474	0.7037
75	0.6790	0.5930	0.7025
90	0.6259	0.5490	0.6338
120	0.6962	0.6140	0.6895

The Random Forest achieved better classification outcomes than the baseline model, feature selection (top K features) using mutual information was applied. Feature v/s accuracy trends for RF across different feature set sizes can be found in Fig. (6) Interestingly, feature selection did not always improve performance; in some cases, models with fewer selected features performed worse than those using more. The performance gain over the simple Decision Tree baseline is clear; for the optimal 60-second window, Random Forest's accuracy (0.765) represents an improvement over the Decision Tree's accuracy (0.7007) for the same window. 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 VII: Random Forest Performance for Different Window Sizes

Window Size (s)	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

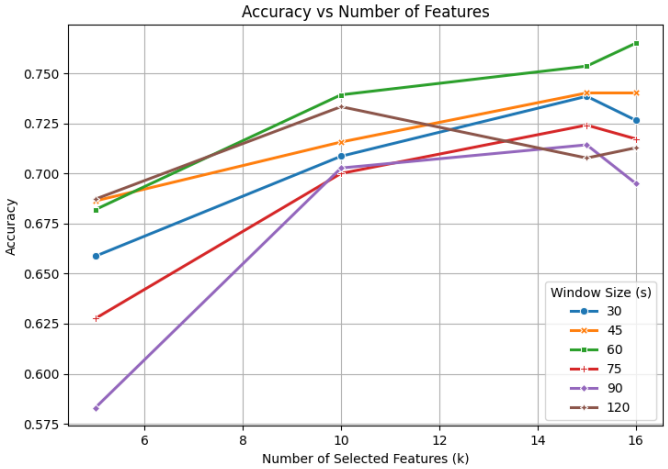

Fig. 6: Random Forest feature vs accuracy plot; comparing the accuracy on different window sizes on k number of features

The 1D CNN architecture further advanced performance by modeling local temporal dependencies in the features. It used two convolutional layers followed by pooling, dropout, and fully connected layers [47], enabling the network to capture nonlinear relationships that the RF could not. Feature selection, as detailed in Table VII, was applied only for the Random Forest model. The 1D CNN model used the full set of 16 features to leverage its inherent ability to learn relevant features directly from the data. Compared to the Decision Tree, the CNN yielded higher stress and baseline F1-scores across most windows, showing the advantage of deeper representation learning.

Overall, while each model showed distinct characteristics, combining EDA and IBI features with a non-overlapping 60-second window consistently resulted in the best performance across all models. Therefore, 60 sec window size was selected for the UQ.

For UQ, both the Random Forest and the CNN models were evaluated using predictive entropy [44]. A threshold of 0.45 was selected as the operating point, balancing reliable classification of stress and baseline with the system's ability to handle uncertain cases gracefully. This selection was guided

TABLE VIII: 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

by the entropy–accuracy plot, which illustrates a clear trade-off between prediction confidence and accuracy: lower entropy thresholds yield fewer, but more accurate, predictions. The plot also indicates an "elbow point" around 0.4 - 0.6, beyond which accuracy declines sharply, suggesting an optimal operating range for triggering dialogue responses in a trust-aware system. Beyond a threshold of ~ 0.7, the accuracy curve becomes flat at around 84% because nearly all predictions are now included, regardless of uncertainty.

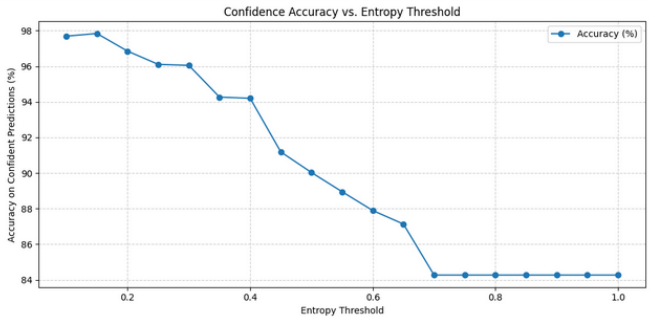

Fig. 7: Entropy–accuracy trade-off for uncertainty quantification. The plot shows prediction accuracy as a function of entropy threshold, highlighting the trade-off between confidence and reliability.

TABLE IX: Uncertainty quantification results for the RF model

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

According to Table IX and Table X, it can be seen that the RF model produced more confident correct predictions (124 correct & certain) compared to the CNN (35 correct & certain), but also yielded some confident errors (13 incorrect & certain). In contrast, the CNN was more conservative, assigning most samples to the uncertain category (226 correct & uncertain). This suggests that RF favors decisiveness with moderate risk, while the CNN favors caution, leading to more frequent but safer deferrals in a trust-aware system.

TABLE X: Uncertainty quantification results for the 1D CNN model

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

To test real-time integration, all trained models were saved after offline evaluation and then reused in ROS2 nodes. A stress inference node was built to subscribe to physiological features, apply the trained model, and publish classification results along with dialogue responses. The dialogue manager made use of UQ values, adapting its responses depending on whether the entropy was below or above the chosen threshold. Synthetic EDA and BVP data generated with NeuroKit2 library [48]; at the similar sampling rate as the WESAD dataset. Fig. (8) provides the demonstration of the ROS2 node with adaptive dialogue; which gets executed when the user_feedback is published. Currently, the ROS2 node processes preprocessed features from CSV files rather than acquiring real-time raw physiological data from sensors. While this approach allows testing and validation of the stress inference pipeline, a key improvement for future work is to integrate real-time signal acquisition and preprocessing directly into the ROS2 system to enable fully online stress detection.

```
[INFO] [1757454194.755443214] [stress_inference_node]: Stress inference + dialogue manager n
ode started.
[INFO] [1757454213.288520685] [stress_inference_node]: Predicted: 0, Entropy: 0.6414, Agent
says: Okay, let's continue for now.
[INFO] [1757454214.268895949] [stress_inference_node]: Predicted: 0, Entropy: 0.6924, Agent
says: Okay, let's continue for now.
[INFO] [1757454215.268709723] [stress_inference_node]: Predicted: 0, Entropy: 0.6238, Agent
says: Okay, let's continue for now.
[INFO] [1757454216.268883848] [stress_inference_node]: Predicted: 0, Entropy: 0.6135, Agent
says: I am uncertain about your stress level. Can you please tell me how you feel?
[INFO] [1757454217.293197248] [stress_inference_node]: Predicted: 0, Entropy: 0.6049, Agent
says: Waiting for your feedback...
[INFO] [1757454217.533078393] [stress_inference_node]: Received user feedback: stressed
[INFO] [1757454218.296780567] [stress_inference_node]: Predicted: 0, Entropy: 0.5976, Agent
says: Thanks for telling me. Let's try a short breathing exercise together.
```

Fig. 8: Example output from the stress inference and dialogue manager ROS2 node. The log shows predicted stress labels, associated entropy values (uncertainty), and corresponding agent responses, the agent asks for the query if it is uncertain for 3; it requests for user_feedback when uncertainty is high. Once the user_feedback is published; intervention takes place

VI. Conclusions

The study confirms that physiological signals from wearable sensors can be leveraged effectively to distinguish between baseline and stress conditions, even with a relatively small dataset. Classical machine learning models like Random Forest offer interpretability and feature importance insights, while deep learning approaches, such as the 1D CNN can capture subtle temporal dynamics in the data. Incorporating uncertainty quantification proved essential for determining when to engage users via the dialogue system, reducing the risk of misleading predictions. The observed accuracy trends suggest that careful selection of window sizes and features plays a critical role in optimizing model performance. Overall, this work provides a strong foundation for personalized stress monitoring systems that are capable of real-time interaction and adaptive intervention, demonstrating both feasibility and practical value for future real-world applications.

A. Summary

The results highlighted that a 60-second non-overlapping window provided the best trade-off between temporal resolution and model performance, achieving approximately 76% accuracy for Random Forest and 75% for the 1D CNN. The framework not only demonstrated effective stress classification but also enabled interactive, user-aware responses based on model confidence, combining machine learning with real-time user feedback.

B. Contributions

The key contributions of this work include the development of a structured methodology that combines feature selection, classical machine learning, and deep learning for physiological stress detection, as well as the implementation of subject-independent evaluation using a LOSO strategy to ensure robustness across participants. UQ techniques were integrated differently for each classification model allowed the system to provide confidence-aware predictions, while the design of a dialogue manager enabled adaptive user interaction based on model uncertainty. This work laid the groundwork for combining predictive modeling with human-centered adaptive interventions.

C. Future Work

Future work should focus on enhancing the system's adaptability and robustness through active learning with context awareness, allowing the model to selectively query the user for labels when uncertainty is high and thereby improve performance over time [49]. Incorporating multimodal data, such as combining physiological signals with behavioral, emotion or contextual information, could provide a richer representation of stress states and improve classification accuracy [14]. Additionally, extending the framework to work with real-time data from wearable sensors in everyday environments will be crucial for evaluating the system's practical applicability and reliability, enabling more personalized and responsive stress monitoring solutions [29], [50].

ACKNOWLEDGMENT

I would like to express my sincere gratitude to Prof. Dr. Teena Hassan and Jordan Schneider for their invaluable guidance, support, and encouragement throughout this project. Their support helped me stay motivated and focused at every stage. I am also thankful to Hochschule Bonn-Rhein-Sieg (HBRS) for providing the resources, facilities, and a supportive research environment that made this study possible.

REFERENCES

[1] Stress effects on the body. Accessed: 2025-01-20. [Online]. Available: https://www.apa.org/topics/stress/body

[2] Chronic stress puts your health at risk. Accessed: 2025-01-20. [Online]. Available: https://www.mayoclinic.org/healthy-lifestyle/stress-management/in-depth/stress/art-20046037

[3] Calm your mind. change your life. Accessed: 2024-11-01. [Online]. Available: https://www.calm.com/

[4] Everything your mind needs. Accessed: 2024-11-01. [Online]. Available: https://www.headspace.com/

[5] R. Sioni and L. Chittaro, "Stress detection using physiological sensors," *Computer*, vol. 48, no. 10, pp. 26–33, 2015.

[6] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ser. ICMI '18. ACM, Oct. 2018, pp. 400–408.

[7] P. Bobade and M. Vani, "Stress detection with machine learning and deep learning using multimodal physiological data," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2020, pp. 51–57.

[8] S. Elzeiny and M. Qaraqe, "Automatic and intelligent stressor identification based on photoplethysmography analysis," *IEEE Access*, vol. 9, pp. 68 498–68 510, 2021.

[9] K. Klęczek, A. Rice, and M. Alimardani, "Robots as mental health coaches: A study of emotional responses to technology-assisted stress management tasks using physiological signals," *Sensors*, vol. 24, no. 13, p. 4032, Jun. 2024.

[10] K. Matheus, M. Vázquez, and B. Scassellati, "A social robot for anxiety reduction via deep breathing," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2022, pp. 89–94.

[11] N. L. Robinson, J. Connolly, L. Hides, and D. J. Kavanagh, "Social robots as treatment agents: Pilot randomized controlled trial to deliver a behavior change intervention," *Internet Interventions*, vol. 21, p. 100320, Sep. 2020.

[12] V. Vasco, C. Willemse, P. Chevalier, D. De Tommaso, V. Gower, F. Gramatica, V. Tikhanoff, U. Pattacini, G. Metta, and A. Wykowska, *Train with Me: A Study Comparing a Socially Assistive Robot and a Virtual Agent for a Rehabilitation Task*. Springer International Publishing, 2019, pp. 453–463.

[13] M. Abd Al-Alim, R. Mubarak, N. M. Salem, and I. Sadek, "A machine-learning approach for stress detection using wearable sensors in free-living environments," *Computers in Biology and Medicine*, vol. 179, p. 108918, Sep. 2024.

[14] G. K. Verma and U. S. Tiwary, "Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals," *NeuroImage*, vol. 102, pp. 162–172, Nov. 2014.

[15] J. Zhai, A. Barreto, C. Chin, and C. Li, "Realization of stress detection using psychophysiological signals for improvement of human-computer interaction," in *Proceedings. IEEE SoutheastCon, 2005.* IEEE, pp. 415–420.

[16] V. Ashwin, R. Jegan, and P. Rajalakshmy, "Stress detection using wearable physiological sensors and machine learning algorithm," in *2022 6th International Conference on Electronics, Communication and Aerospace Technology*. IEEE, Dec. 2022, pp. 972–977.

[17] A. Ghaderi, J. Frounchi, and A. Farnam, "Machine learning-based signal processing using physiological signals for stress detection," in *2015 22nd Iranian Conference on Biomedical Engineering (ICBME)*. IEEE, Nov. 2015, pp. 93–98.

[18] S. A. M. Mane and A. Shinde, "Stressnet: Hybrid model of lstm and cnn for stress detection from electroencephalogram signal (eeg)," *Results in Control and Optimization*, vol. 11, p. 100231, Jun. 2023.

[19] A. O. de Berker, R. B. Rutledge, C. Mathys, L. Marshall, G. F. Cross, R. J. Dolan, and S. Bestmann, "Computations of uncertainty mediate acute stress responses in humans," *Nature Communications*, vol. 7, no. 1, Mar. 2016.

[20] J. O. Alotaibi and A. S. Alshahre, "The role of conversational ai agents in providing support and social care for isolated individuals," *Alexandria Engineering Journal*, vol. 108, pp. 273–284, Dec. 2024.

[21] G. Dosovitsky and E. L. Bunge, "Bonding with bot: User feedback on a chatbot for social isolation," *Frontiers in Digital Health*, vol. 3, Oct. 2021.

[22] O. A. Pub. Physiological signals. [Online]. Available: https://openaccesspub.org/international-physiology-journal/physiological-signals

[23] B. Wang, H. Zhou, X. Li, G. Yang, P. Zheng, C. Song, Y. Yuan, T. Wuest, H. Yang, and L. Wang, "Human digital twin in the context of industry 5.0," *Robotics and Computer-Integrated Manufacturing*, vol. 85, p. 102626, Feb. 2024.

[24] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations," *Information Fusion*, vol. 102, p. 102019, Feb. 2024.

[25] W. Fang, D. Wu, P. E. Love, L. Ding, and H. Luo, "Physiological computing for occupational health and safety in construction: Review, challenges and implications for future research," *Advanced Engineering Informatics*, vol. 54, p. 101729, Oct. 2022.

[26] V. Nemani, L. Biggio, X. Huan, Z. Hu, O. Fink, A. Tran, Y. Wang, X. Zhang, and C. Hu, "Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial," *Mechanical Systems and Signal Processing*, vol. 205, p. 110796, Dec. 2023.

[27] N. Durasov. (2020) Masksembles for uncertainty estimation. [Online]. Available: https://www.neuralconcept.com/post/the-importance-of-uncertainty-quantification-for-deep-learning-models

[28] A. Fernandes, R. Helawar, R. Lokesh, T. Tari, and A. V. Shahapurkar, "Determination of stress using blood pressure and galvanic skin response," in *2014 International Conference on Communication and Network Technologies*, 2014, pp. 165–168.

[29] E. Hosseini, R. Fang, R. Zhang, A. Parenteau, S. Hang, S. Rafatirad, C. Hostinar, M. Orooji, and H. Homayoun, "A low cost eda-based stress detection using machine learning," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2022, pp. 2619–2623.

[30] Kubios. How to analyze empatica e4 measurements. [Online]. Available: https://www.kubios.com/downloads/How-to-analyze-Empatica-E4-measurements.pdf

[31] K. Kyriakou, B. Resch, G. Sagl, A. Petutschnig, C. Werner, D. Niederseer, M. Liedlgruber, F. Wilhelm, T. Osborne, and J. Pykett, "Detecting moments of stress from measurements of wearable physiological sensors," *Sensors*, vol. 19, no. 17, p. 3805, Sep. 2019.

[32] A. Aygun, H. Ghasemzadeh, and R. Jafari, "Robust interbeat interval and heart rate variability estimation method from various morphological features using wearable sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 8, pp. 2238–2250, Aug. 2020.

[33] N. Milstein and I. Gordon, "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*, vol. 14, Aug. 2020.

[34] E. S.R.L. (2014) Empatica e4. [Online]. Available: https://box.empatica.com/documentation/20141119_E4_TechSpecs.pdf

[35] D. Jia, X. Zhang, J. T. Zhou, P. Lai, and Y. Wei, "Dynamic thresholding for video anomaly detection," *IET Image Processing*, vol. 16, no. 11, pp. 2973–2982, May 2022.

[36] S. Campanella, A. Altaleb, A. Belli, P. Pierleoni, and L. Palma, "A method for stress detection using empatica e4 bracelet and machine-learning techniques," *Sensors*, vol. 23, no. 7, p. 3565, Mar. 2023.

[37] N. El Haouij, J.-M. Poggi, R. Ghozi, S. Sevestre-Ghalila, and M. Jaïdane, "Random forest-based approach for physiological functional variable selection for driver's stress level classification," *Statistical Methods & Applications*, vol. 28, no. 1, pp. 157–185, Feb. 2018.

[38] H. S. Deshpande and L. Ragha, "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*, vol. 15, no. 1, p. 84, 2023.

[39] R. Sánchez-Reolid, F. López de la Rosa, M. T. López, and A. Fernández-Caballero, "One-dimensional convolutional neural networks for low/high arousal classification from electrodermal activity," *Biomedical Signal Processing and Control*, vol. 71, p. 103203, Jan. 2022.

[40] D. S. Benita, A. S. Ebenezer, L. Susmitha, M. Subathra, and S. J. Priya, "Stress detection using cnn on the wesad dataset," in *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, 2024, pp. 308–313.

[41] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep ppg: Large-scale heart rate estimation with convolutional neural networks," *Sensors*, vol. 19, no. 14, p. 3079, Jul. 2019.

[42] A. Shtrauss. (2022) Pytorch | cnn binary image classification. [Online]. Available: https://www.kaggle.com/code/shtrausslearning/pytorch-cnn-binary-image-classification/notebook

[43] M. H. Shaker and E. Hüllermeier, "Aleatoric and epistemic uncertainty with random forests," 2020.

[44] F. Porcher. (2023) Entropy based uncertainty prediction. [Online]. Available: https://towardsdatascience.com/entropy-based-uncertainty-prediction-812cca769d7a/
[45] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," 2015.
[46] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, Dec. 2021.
[47] D. J. Zamora-Mora. (2022) Cats vs dogs: Binary classifier with pytorch cnn. [Online]. Available: https://www.doczamora.com/cats-vs-dogs-binary-classifier-with-pytorch-cnn
[48] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, feb 2021. [Online]. Available: https://doi.org/10.3758%2Fs13428-020-01516-y
[49] R. Monarch and R. Munro, *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-centered AI*. Manning, 2021. [Online]. Available: https://books.google.de/books?id=LCh0zQEACAAJ
[50] A. Tazarv, S. Labbaf, A. Rahmani, N. Dutt, and M. Levorato, "Active reinforcement learning for personalized stress monitoring in everyday settings," 2023.

STATEMENT OF ORIGINALITY

I, the undersigned below, declare that this work has not previously been submitted to this or any other university and that it is, unless otherwise stated, entirely my own work. The report was, in part, written with the help of the AI assistant Grammerly as described in the appendix. I am aware that content generated by AI systems is no substitute for careful scientific work, which is why all AI-generated content has been critically reviewed by me, and I take full responsibility for it.

Date Signature

APPENDIX

TABLE XI: Summary of Software and Tools

Category	Tool/Software / Version
Synthetic Data Generation	NeuroKit2 / 0.2.12
Machine Learning Framework	PyTorch 2.0+
Dataset	WESAD (Wearable Stress and Affect Detection)
IDE	VS Code / 1.102.3
Version Control	Git / 2.51.0
Middleware	ROS2 (Humble)
Operating System	Ubuntu 22.04 LTS
Source Code	https://github.com/prachi0711/Stress-Management-using-Physiological-Signals-and-conversational-agents