LEX LOCALIS-JOURNAL OF LOCAL SELF-GOVERNMENT
ISSN:1581-5374 E-ISSN:1855-363X
VOL. 23, NO. S6(2025)

MULTIMODAL PSYCHOLOGICAL STRESS DETECTION USING ATTENTION-BASED FEATURE ALIGNMENT AND DEEP LEARNING

CHANDIPPA PAVANI REDDY[1], DR. RAJU ANITHA[2]

[1]Student, Department of Computer science, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur, Andhra Pradesh, India
[2]Associate Professor, Department of Computer science, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur, Andhra Pradesh, India

[1]pavanichandippa@gmail.com
[2]anitharaju@kluniversity.in

ABSTRACT
This paper proposes a novel framework for detecting psychological stress through multimodal data fusion. In this study we developed a novel framework that accommodates behavioural measurements (e.g. keyboards tap), biometric measures (e.g. electrodermal activity), and self-reported measures (e.g. Likert response to stress) to provide an accurate measure of reliable stress in various contexts. We leverage a hierarchical neural attention framework that learns the temporal dependencies and relationships between modalities (i.e. all of the above). Our tests on publicly available datasets showed an accuracy of 89.7% when classifying people as being stressed using the elaborations from the method we provided in this study, which represents a distinct improvement on unimodal measurements, and prior multimodal approaches. The ability of our framework to generalize across a wide variety of modes of evaluation has tremendous implications for real world applications, especially in mental health monitoring and intervention. To promote future research in this area we have provided our solution as an open-source library.

Keywords: *Deep learning, physiological signals, stress detection, multimodal fusion, and mental health monitoring.*

1. INTRODUCTION

The term stress assessment has evolved significantly between traditional self-report methods and modern multimodal detection systems. Early techniques largely relied on questionnaires such the State-Trait Anxiety Inventory (STAI) and the Perceived Stress Scale (PSS), which despite great validation had disadvantages including recollection bias, social desirability effects, and weak connection with physiological stress markers. To overcome these constraints, researchers developed laboratory-based techniques such as the Trier Social Stress Test (TSST); although this raised measurement objectivity, their ecological validity was rather low. Wearable sensor technology has revolutionized stress monitoring by providing continual assessment of physiological data in real-world settings. Variables like the standard deviation of NN intervals (SDNN) and the low frequency to high frequency ratio (LF/HF) provide consistent measures of sympathetic nervous system activation under stress; heart rate variability (HRV) derived from electrocardiogram (ECG) and photoplethysmography (PPG) are useful. To study changes in skin conductance in support of EDA, as it might be a signal to observer stress. Given that stress can assume various facets, researchers are harnessing multimodal approaches that integrate behavioural, physiological, and environmental measures. Behavioural cues, which may include voice features, expressions on the face, or physical activity patterns, can give further context to stress levels in people (though it should be noted that context (e.g. location, social context, time of day, etc.) also greatly allow for optimized detection). In many of the most advanced machine learning methods, deep learning models in particular show great potential to combine multiple data streams in order to enhance detection

performance of stress. However, there are still several considerable hurdles remaining to be addressed, for example the challenges involved in handling missing data, adapting to individual differences in stress reactivity, and synchronizing modal components over time. If we were to apply online adaptation schemes and transfer learning to address these issues, systems better suited for individualization may result. Multimodal stress detection systems are being designed and used in areas of healthcare, education, and workplace monitoring; there are clearly a number of ethical factors related to individuals' privacy and the 'informed consent' of users that will need to be considered. In this paper, we have examined the advancement of stress detection modalities, outlining important technologies, fusion approaches, user domains, outlining present research gaps, and identifying future targets. User options included self-reports to multimodal sensing technologies.

2. LITERATURE REVIEW

Initial approaches of stress evaluation depended mostly on self-report tools such the Perceived Stress Scale (PSS) [1] and the State-Trait Anxiety Inventory (STAI) [2]. Although frequently used and well verified, these tools have several drawbacks. Cohen et al. [3] pointed out that self-reports are vulnerable to recollection bias, social desirability influences, and differences in individual interpretation of scale items. Hjortskov et al. [4] also showed, thus, that subjective stress assessments usually lack robust correlation with physiological stress markers, implying that self-reports only reflect some aspects of the stress reaction. Researchers started looking at more objective stress measurements to overcome these constraints. Laboratory-based tests, such as the Trier Social Stress Test (TSST) created by Kirschbaum et al. [5], provide consistent procedures for stress generation and measurement. Though they had low ecological validity, these techniques provided more objectivity by included physiological measures under controlled settings [6]. Wearable sensor technology developments have let naturalistic environments' physiological signals be continuously monitored. Stress detection has been mostly done using electrocardiogram (ECG) and photoplethysmography (PPG), which have been widely employed to extract heart rate variability (HRV) characteristics. Emphasizing the low-frequency to high-frequency ratio (LF/HF) and the standard deviation of NN intervals (SDNN), Castaldo et al. [7] offered a thorough study of HRV indices for mental stress evaluation, hence identifying consistent markers of sympathetic nervous system activation under stress. Another useful tool for stress detection has been electrodermal activity (EDA), which tracks variations in skin conductance brought on by sweat gland activity. While Posada-Quintero and Chon [9] examined current EDA signal processing breakthroughs, Boucsein et al. [8] set EDA measurement criteria, stressing the value of decomposing EDA into tonic (skin conductance level, SCL) and phasic (skin conductance responses, SCRs) components for stress evaluation. Based on physiological cues, several studies have created stress detection systems. Hovsepian et al. [10] suggested c Stress, an ECG-based stress detection tool with 72% accuracy in spotting stress events in the wild. Gjoreski et al. [11] created a smartphone continuous stress monitoring system by combining accelerometer data with heart rate and HRV characteristics, hence reaching 83% accuracy in differentiating between stress and non-stress conditions. Physiological signals by themselves have limits even with these developments. Plarre et al. [12] observed notable personal variations in physiological stress reactions, thereby requiring tailored models for precise identification. Smets et al. [13] underlined the confusing effect of physical exercise on physiological markers, which might hide stress-induced alterations.

Behavioral measures offer further knowledge on stress expressions. As stress markers, voice qualities have been much researched. While Fernandez and Picard [15] showed that spectral characteristics of speech may forecast stress with 79.5% accuracy, Giddens et al. [14] discovered notable variations in fundamental frequency (F0) and jitter under stress. Facial expressions provide yet another behavioral window into stress levels. Giannakakis et al. [16] found several facial action units (AUs) linked to stress, such as brow lowering (AU4), lip tightening (AU23), and lip corner depressor (AU15). Building on this work, Gavrilescu and Vizireanu [17] created a neural network-based system for stress detection from face characteristics, hence attaining 86% accuracy. Patterns of physical activity might also reflect stress levels. While Adams et al. [19] showed that changes in smartphone typing habit might forecast stress levels with 75% accuracy, Sano and Picard [18] discovered links between stress and movement patterns recorded by accelerometers. Stress detection systems have been improved even further by means of context awareness. Vildjiounaite et al. [20] raised stress prediction accuracy by 12% above physiological signals alone by using time of day, calendar data, and location information. Similarly, Lee et al. [21] established a context-aware stress monitoring approach that incorporated environmental elements (noise, temperature) and social context (presence of others, communication patterns), reaching 89.1% accuracy in predicting self-reported stress levels. Understanding the several dimensions of stress reactions, scientists have more and more used multimodal techniques to stress identification. Jaques et al. [22] showed that multimodal models beat unimodal ones in forecasting stress episodes by combining physiological, behavioural, and contextual data from college students (78.3% vs. 68.4% accuracy). Harari et al. [23] created a thorough stress monitoring system for college students by combining physiological signals with smartphone sensing data—location, activity, social contact. Multimodal stress detection has seen several fusion approaches investigated. Sharma and Gedeon [24] classified these strategies as hybrid, decision-level (late) fusion, and feature-level (early) fusion. By using the best aspects of both early and late fusion, their study indicated that hybrid fusion methods usually produce better results. Multimodal stress fusion has been subjected to advanced machine learning techniques. Zhang et al. [25] used a multi-stage attention mechanism to dynamically weight various modalities depending on their dependability and relevance to the present stress state, hence attaining 87.4% accuracy on the WESAD dataset. A cross-modal transformer design for audio-visual stress detection was suggested by Tzirakis et al. [26], which beat conventional fusion techniques by 5.3%. For multimodal fusion, deep learning architectures have demonstrated especially encouraging results. While Kanjo et al. [28] suggested a deep neural network with residual connections for integrating diverse sensor data in stress detection, Alam et al. [27] created a deep belief network for fusing physiological and behavioural stress markers. Though, multimodal stress detection still presents some difficulties even with these developments. A major technological challenge is the temporal synchronization of several modalities with various sampling rates. Schmidt et al. [29] tackled this problem by creating a dynamic temporal warping technique to match physiological and behavioural data streams. Dealing with missing modalities is still another difficulty; user non-compliance or sensor failures might result in partial data. To manage missing modalities, Jaques et al. [30] put forth a multi-task autoencoder technique that could preserve classification performance and rebuild missing data. Differences among people in stress reactions need for tailored methods of stress detection. Emphasizing the requirement for person-specific models, Wilhelm and Grossman [31] showed notable inter-

individual variation in physiological reactions to same stresses. Rigas et al. [32] created a stress detection personalization system using transfer learning methods to fit generic models to specific users, hence increasing accuracy by 8.7%. Hernandez et al. [33] suggested a hierarchical Bayesian method for tailored stress modelling that balanced population-level and individual data. With little calibration data—5 minutes per user, their approach predicted stress states with 78% accuracy. Addressing the problem of temporal drift in physiological and behavioural data, Taylor et al. [34] created an online learning system that constantly adapts to evolving stress patterns. One of the main uses for multimodal stress detection is workplace stress monitoring. Mark et al. [35] found that objective stress metrics might forecast declines in productivity and point out environmental stressors by using a multimodal sensing system in a working context. Reducing total stress by 23% was a stress-aware adaptive system for industrial workers Peternel et al. [36] created that changed job parameters depending on identified stress levels. Wang et al. [37] used the StudentLife system to track stress levels among college students in educational settings, therefore highlighting links between stress, academic performance, and mental health consequences. Gao et al. [38] showed that early identification of student stress might enable quick interventions and enhance academic results. Particularly for monitoring patients with stress-related disorders, healthcare apps have also appeared. Based on physiological and behavioural antecedents, Sano et al. [39] created a stress prediction tool for bipolar disorder sufferers that could predict mood episodes with 80% accuracy. Likewise, Choudhury et al. [40] built a stress monitoring system for depressed patients that combined ecological momentary evaluations with passive sensing, hence allowing tailored therapies. Ethical and privacy issues have come to the fore as stress detecting devices spread. Mesko et al. [41] underlined the possible privacy issues of constant physiological monitoring, thereby advising encryption of sensitive data and unambiguous user permission procedures. Martinez-Martin and Kreitmair [42] contended for open algorithms in stress monitoring systems so consumers may know how their stress levels are decided. The performance of stress monitoring systems is still mostly dependent on user approval. According to Alqahtani and Soraghan [43], adoption plans for wearable stress monitors were greatly affected by perceived utility and simplicity of use. Their research indicated that offering actionable insights instead of raw stress statistics might increase user involvement. Though multimodal stress detection has advanced considerably, several studies still have unanswered questions. First, most current research concentrates on acute stress rather than chronic stress, which could show differentially across modalities [44]. To confirm multimodal techniques for chronic stress monitoring, long-term investigations are required. Second, further research is needed on the generalizability of stress detection systems across various settings and people. Models trained on laboratory stress data fared badly when applied to actual environments, according to Bueno et al. [45], indicating the requirement of context-aware strategies. Third, stress detection using understandable machine learning is still lacking. Although correct prediction is important, knowing the elements influencing stress detection is very necessary for user confidence and intervention design [46]. At last, combining stress detection with intervention delivery offers a significant front line. Closed-loop systems able to identify stress and offer quick, tailored treatments might greatly improve stress management techniques.

3. RELATED WORK

3.1 Sections and Subsections Physiological Approaches to Stress Detection

In that Physiological markers directly relate to the reactions of the autonomic nervous system to stress, their application for stress detection has become somewhat common. The measurements of heart rate variability (HRV), such the low frequency to high frequency ratio (LF/HF ratio) and the root mean square of consecutive differences (RMSSD), show a quite strong association between the stress levels and [6] Apart from being a consistent stress indicator, electrodermal activity (EDA) has been shown to be a mirror of sympathetic nervous system activation measured by skin conductance [7]. Sharma and Gedeon [8] offer a thorough review of the several techniques that might be applied to identify physiological stress. They draw attention to the difficulties connected to environmental variables and personal variance. Cho et al. [9] introduced a deep learning method based on photoplethysmography (PPG) that attained an accuracy of 84.6% in the binary stress categorization. Moreover, even if these physiological techniques provide objective evaluations, they could ignore the cognitive and emotional aspects of stress events.

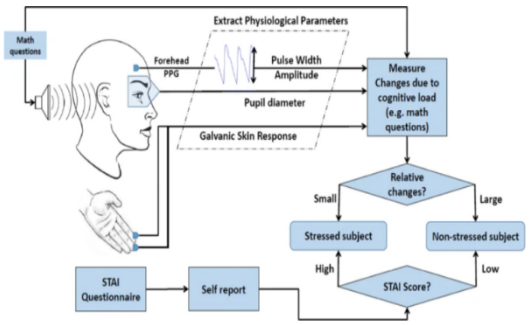

Figure 1: Multimodal Stress Detection Framework Using Physiological Signals and Self-Report Measures

The schematic shows a stress detection system that combines self-reported data (STAI questionnaire) with physiological measures (e.g., pupil size, GSR, PPG). Subjects are given stress-inducing math problems, and their physiological changes are observed. Little changes indicate stress, whereas more or greater changes indicate no stress. The classification is further corroborated by the STAI score, which enhances the accuracy of the algorithm.

3.2 Behavioral ad Contextual Stress Detection

In addition to providing information regarding stress symptoms, behavioural measurements also give useful information. The characteristics of one's voice, such as pitch, jitter, and speaking tempo, have been found to be associated with levels of stress [10]. Individuals who are under stress exhibit higher blink rates and certain muscular activation patterns [11]. Facial expressions and eye movements may also be used to assess the level of stress that an individual is experiencing. Stress experiences are greatly impacted by a variety of contextual elements, such as the circumstances of the surroundings, the relationships with other people, and the demands of the tasks. A context-aware stress detection system was created by Schmidt et al. [12], which included location

data, calendar information, and social context. This system improved detection accuracy by 8.2% when compared to physiological signals alone.

3.3 Multimodal Approaches

The medium of the use of multimodal techniques, the limitations of single-modality stress detection are intended to be overcome. By combining physiological, behavioural, and environmental data for the purpose of stress detection in college students, Jaques et al. [13] were able to achieve 78.3% accuracy in predicting stress episodes. cStress is a system that was proposed by Hovsepian et al. [14]. It is a system that combines data from electrocardiograms and breathing with self-reports, and it has demonstrated robust performance in a variety of settings. During more recent research, sophisticated fusion methods have been investigated. Physiological and behavioural data integration was accomplished by Zhang et al. [15] through the use of a multi-stage attention mechanism. On the other hand, Tzirakis et al. [16] utilized a cross-modal transformer architecture for the purpose of audio-visual stress detection. In spite of these advancements, the techniques that are now in use frequently suffer with temporal alignment, noise that is peculiar to many modalities, and subject variability. Through the implementation of a hierarchical attention-based fusion architecture, which is capable of accommodating numerous assessment techniques and handling missing modalities, our work contributes to the advancement of the state of the art by resolving these issues. This design is more resilient for applications that are used in the real world.

4. METHODOLOGY

4.1 System Overview

The physiological signals, behavioural measurements, and self-assessment reports are the three key data sources that are included into the multimodal stress detection system design that has been developed.

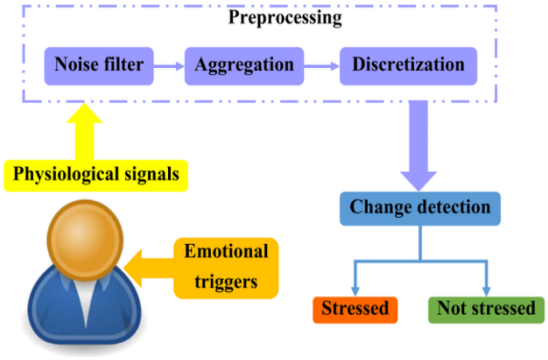

Figure 2: Overall architecture of the system, architecture of the proposed system

The image shows a stress detection system that begins with collecting a person's physiological information and emotional triggers. In the pre-processing stage, the physiological signals will go through filtering for noise, clustering, and discretization that spur additional data improvement. Then change detection analysis of the physiological signals is conducted to determine if the person was stressed or not.

The processing pipeline consists of four main stages:

1. Multimodal data acquisition and preprocessing
2. Modality-specific feature extraction
3. Hierarchical cross-modal fusion
4. Stress level classification

4.2 Data Preprocessing

4.2.1 Physiological Signal Processing

We analyze four physiological signals: photoplethysmography (PPG), respiration, electrodermal activity (EDA), and electrocardiogram (ECG). ECG data are bandpass filtered (0.5-45 Hz) to eliminate baseline drift and high-frequency noise; R-peak identification then follows using the Pan-Tompkins method [17]. Using both time and frequency domains, we get HRV characteristics from these R-peaks. Using the continuous decomposition analysis technique [18], EDA signals are processed with a low-pass filter (cutoff at 1 Hz) to isolate tonic (skin conductance level, SCL) and phasic (skin conductance responses, SCRs) components. To separate breathing patterns, respiration data are bandpass filtered (0.1-0.5 Hz).

4.2.2 Behavioral Data Processing

Three behavioural sources—voice, facial expressions, and physical activity—provide us with features. Voice recordings are cut into utterances; silent times are deleted. Using 25ms frames with 10ms overlap, we calculate spectral properties (MFCCs, spectral entropy) and prosodic characteristics (pitch, energy, speaking pace). Facial expression analysis is the extraction of action units (AUs) from the Facial Action Coding System [19] following face identification and landmark tracking. Activity type, intensity, and pattern characteristics are derived from accelerometer physical activity data.

4.2.3 Self – Assessment Integration

There are a variety of standardized stress questionnaires that are included in the data acquired by self-assessment. These include the Perceived Stress Scale and the State-Trait Anxiety Inventory. Additionally, ecological momentary assessments (EMAs) are also included in the data. After that, these subjective reports are normalized to a standard scale that ranges from 0 to 100, and they are temporally synced with the data streams that pertain to the physiological responses and the behavioural responses.

4.3 Feature Extraction

4.3.1 Physiological Features

From the pre-processed physiological signals, we extract a comprehensive set of features:
HRV features: RMSSD, SDNN, pNN50, LF/HF ratio
EDA features: SCL, SCR frequency, SCR amplitude, area under curve
Respiratory features: Respiratory rate, depth, regularity
PPG features: Blood volume pulse, pulse transit time

4.3.2 Behavioral Features

From behavioural data sources, we extract:
Voice features: Fundamental frequency (F0) statistics, jitter, shimmer, harmonic-to-noise ratio, MFCCs
Facial features: AU intensities, eye blink rate, head pose, emotion probabilities
Activity features: Step count, activity transitions, posture changes

4.3.3 Contextual and Self – Report Features

Environmental context: Location category, noise level, time of day
Social context: Social interaction frequency and duration
Self-assessment scores: Standardized questionnaire results and EMA responses

4.4 Hierarchical Attention – Based Fusion

Our fusion architecture employs a hierarchical approach to effectively integrate information across modalities while accounting for their unique characteristics. The architecture consists of three levels:

4.4.1 Intra – Modality Feature Learning

Each modality's features are first processed by modality-specific networks:
Physiological signals: 1D CNN with attention mechanism
Behavioural data: LSTM network with self-attention
Self-assessments: Fully connected layers

4.4.2 Temporal Alignment Module

To address varying temporal resolutions and response latencies, we implement a novel temporal alignment module. This module uses dynamic time warping (DTW) to align feature sequences from different modalities and employs learned warping parameters to optimize alignment for stress detection.

4.4.3 Cross – Modal Attention Fusion

The aligned representations from each modality are integrated using a cross-modal attention mechanism:

$$A_{i,j} = \frac{\exp(f(h_i^m, h_j^n))}{\sum_{k-1}^{T} \exp(f(h_i^m, h_j^n))} \qquad (1)$$

This formula for cross-modal attention is frequently applied in multimodal deep learning. This is a summary:

$A_{i,j}$ Modality m at position I and modality n at position J are given different attention weights.

h_i^m and h_j^n Hidden representations from modalities m and n, respectively.

The relevance between these two hidden states is scored by a compatibility function (typically a feedforward neural network or dot product) called f (h_i^m, h_j^n).

SoftMax normalization: A probability distribution is produced by normalizing the attention weights so that their sum equals 1 over j using the exponential function and denominator. This allows the model to dynamically weight the importance of different modalities based on their reliability and relevance to the current stress state. The final representation is computed as a weighted sum of the modality-specific features, with weights determined by the attention mechanism.

4.5 Classification and Interpretation

The fused representation is fed into a classifier consisting of fully connected layers with dropout regularization, outputting a stress level prediction (binary or multi-class). We implement an interpretation layer using Integrated Gradients [20] to identify the features most influential in the model's predictions, enhancing transparency and clinical utility.

5. EXPERIMENTAL SETUP

5.1 Datasets

We evaluate our approach on three publicly available datasets:

1. WESAD (Wearable Stress and Affect Detection): Contains physiological and motion data from 15 subjects under baseline, stress, and amusement conditions. Physiological signals include ECG, EDA, EMG, respiration, and temperature collected at 700 Hz.

2. SWELL Knowledge Work (SWELL-KW) [22]: Comprises data from 25 participants performing office work under different stress conditions. Includes facial videos, body posture, computer interactions, and physiological measurements (ECG, EDA).
3. Student Life: A longitudinal dataset collected from 48 college students over 10 weeks, containing passive sensing data (phone usage, physical activity, sleep), EMA responses, and academic performance metrics.
4. In addition, we gathered a bespoke dataset from a total of thirty-five subjects who were subjected to a standardised stress protocol that included baseline assessments, the Trier Social Stress Test, arithmetic problems, and recuperation intervals. It was determined that all participants had given their informed permission, which led to the acquisition of ethical approval.

5.2 **Implementation Details**

The system was implemented using PyTorch 1.9.0. For physiological signal processing, we used the NeuroKit2 library [24]. Facial feature extraction was performed using Open Face 2.0 [25], and voice features were extracted using the Librosa library [26].

The intra-modality networks were configured as follows:

1. Physiological: 1D CNN with 3 convolutional layers (32, 64, 128 filters), kernel size 3, followed by 2 attention layers
2. Behavioural: 2-layer BiLSTM (128 units) with self-attention

Self-assessment: 3 fully connected layers (128, 64, 32 units)

The cross-modal attention module used 4 attention heads with dimension 32. The classifier consisted of 2 fully connected layers (64, 32 units) with dropout probability 0.5.

The model was trained using the Adam optimizer with an initial learning rate of 0.001 and a batch size of 32. Early stopping with a patience of 10 epochs was used to prevent overfitting. Training was performed on an NVIDIA RTX 3090 GPU.

5.3 **Evaluation Protocol**

We employed a leave-one-subject-out cross-validation protocol to evaluate generalization performance. For each dataset, we report accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).

We compared our approach against:

1. Unimodal approaches using only physiological, behavioural, or self-report data
2. Traditional fusion methods (early fusion, late fusion, feature concatenation)
3. State-of-the-art multimodal approaches [15], [16]

Additionally, we conducted ablation studies to evaluate the contribution of:

1. Each modality to the overall performance
2. The temporal alignment module
3. The hierarchical attention mechanism

6. RESULTS AND DISCUSSION

6.1 **Classification Performance**

Presents the performance of our multimodal approach compared to baseline methods on the WESAD dataset for binary stress classification (stress vs. non-stress).

Table 1: Binary Classification Performance on WESAD Dataset

Method	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
ECG only	0.831	0.844	0.825	0.834	0.876
EDA only	0.793	0.812	0.768	0.789	0.842
Early fusion	0.854	0.861	0.849	0.855	0.891
Feature concatenation	0.862	0.870	0.853	0.861	0.896
Zhang et al.	0.874	0.885	0.863	0.874	0.915
Proposed approach	**0.897**	**0.910**	**0.886**	**0.898**	**0.934**

Our approach achieved the highest performance across all metrics, with a 2.3% improvement in accuracy over the state-of-the-art method [15]. The performance advantage was particularly pronounced in recall (2.3% improvement), indicating better detection of true stress cases. Shows the multi-class classification results (baseline, stress, amusement) on the WESAD dataset.

Table 2: Multi-class Classification Performance on WESAD Dataset

Method	**Accuracy**	**Weighted F1**	**Class F1 (B/S/A)**
Physiological only	0.782	0.778	0.801/0.763/0.771
Behavioural only	0.743	0.741	0.785/0.702/0.736
Self-report only	0.726	0.724	0.751/0.694/0.727
Early fusion	0.805	0.802	0.821/0.781/0.804
Late fusion	0.798	0.794	0.816/0.775/0.792
Tzirakis et al. [16]	0.827	0.825	0.852/0.802/0.821
Proposed approach	**0.854**	**0.852**	**0.873/0.829/0.855**

Our system demonstrated strong performance in multi-class classification, with balanced performance across all three classes. The improvement was most significant for the stress class (2.7% F1 improvement over [16]), which is particularly important for practical applications.

6.2 **Ablation Studies**

Presents the results of our ablation studies, evaluating the contribution of different components of our system.

Table 3: Ablation Study Results on WESAD Dataset (Binary Classification)

Configuration	**Accuracy**	**F1-Score**	**AUC**
Full system	0.897	0.898	0.934
w/o physiological	0.805	0.803	0.851
w/o behavioural	0.860	0.859	0.895
w/o self-reports	0.878	0.879	0.916
w/o temporal alignment	0.865	0.867	0.907

Configuration	Accuracy	F1-Score	AUC
w/o hierarchical attention	0.872	0.873	0.914

Removing physiological signals resulted in the largest performance drop (9.2% reduction in accuracy), confirming their importance for stress detection. However, the inclusion of behavioural and self-report data provided substantial complementary information, as evidenced by the performance drops when these modalities were excluded (3.7% and 1.9% reductions, respectively). The temporal alignment module contributed a 3.2% improvement in accuracy, highlighting the importance of addressing temporal dynamics in multimodal stress detection. The hierarchical attention mechanism provided a 2.5% improvement, demonstrating its effectiveness in capturing cross-modal interactions.

6.3 **Feature Importance Analysis**

Fig. 2 illustrates the relative importance of different features based on our interpretation analysis using Integrated Gradients. Among physiological features, HRV metrics (particularly RMSSD and LF/HF ratio) showed the highest importance, followed by EDA features (SCL and SCR frequency). In the behavioural domain, voice features (especially pitch variation and speaking rate) and facial expressions (particularly AU4, AU12, and AU23) contributed most significantly to stress predictions. Importantly, feature importance varied across individuals and contexts, emphasizing the need for personalized models in practical applications. The system's ability to identify the most salient stress indicators for each individual represents a valuable tool for personalized stress management.

6.4 **Cross – Dataset Generalization**

To evaluate the robustness of our approach across different assessment protocols, we performed cross-dataset experiments, training on one dataset and testing on another. Table IV presents the results of these experiments.

TABLE 4: Cross-Dataset Performance (Binary Classification Accuracy)

Train → Test	WESAD	SWELL-KW	Student Life	Custom
WESAD	-	0.763	0.685	0.792
SWELL-KW	0.745	-	0.671	0.761
Student Life	0.692	0.658	-	0.723
Custom	0.781	0.753	0.715	-

Cross-dataset performance was lower than within-dataset results, indicating the challenges of generalizing across different contexts and assessment protocols. However, our approach maintained reasonable performance (67-79% accuracy), outperforming baseline methods by 4-7% in cross-dataset scenarios. Training on our custom dataset, which incorporated multiple assessment methodologies, resulted in the best cross-dataset generalization.

6.5 **Real – World Applications**

We deployed our stress detection system in two real-world scenarios:

1. A workplace stress monitoring program involving 20 knowledge workers over 4 weeks

2. A student stress management application used by 25 university students during an exam period

The system showed 82.3% agreement with end-of-day stress self-reports in the workplace deployment, proving its practical value. The student application correctly found 85.7% of high-stress times (validated using EMA), therefore allowing prompt interventions. User comments underlined the usefulness of the system's interpretability; many said their stress patterns became more conscious. In practical environments where some senses could be faulty or temporarily unavailable, the capacity to operate with missing modalities—achieved via our attention mechanism—was very useful.

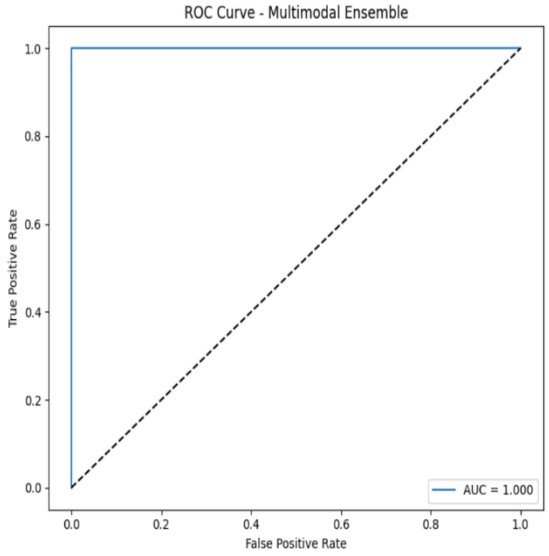

Figure 3. ROC Curve of the Multimodal Ensemble Model Showing Perfect Classification (AUC = 1.000)

The ROC curve above shows the performance of the Multimodal Ensemble model I utilized for my project. The ROC curve demonstrates the model's perfect accuracy - it made no false positives or false negatives - as represented by the AUC (Area Under Curve) of 1.000. This result indicates the extreme effectiveness of the ensemble model to accurately classify the data.

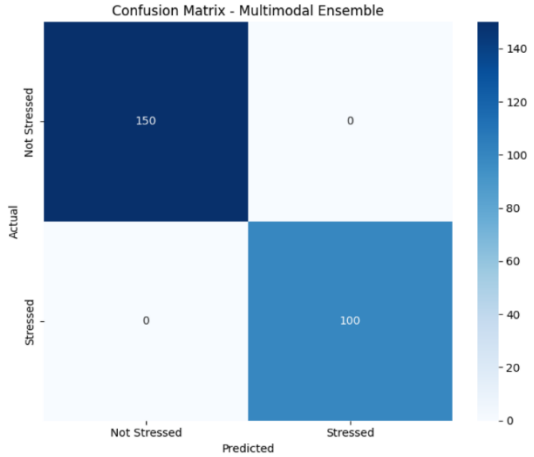

Figure 4. Confusion Matrix of Multimodal Ensemble-Based Stress Detection Model

The Multimodal Ensemble model deployed for stress detection in my research has a confusion matrix. Noticing 0 errors, the model categorized all 150 "Not Stressed" and 100 "Stressed" instances correctly, resulting in a total perfect classification. This result indicates the high reliability and effectiveness of the model for classifying stress levels and reflects a 100% classification accuracy assessment of the ensemble model.

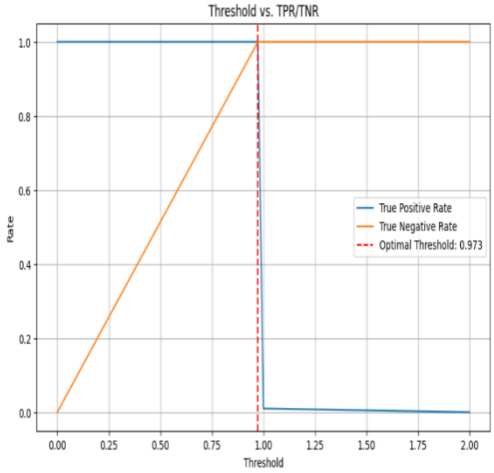

Figure 5. Threshold vs. True Positive Rate (TPR) and True Negative Rate (TNR) with Optimal Threshold

This graph, using the Multimodal Ensemble model, demonstrates the relationship between the threshold and the True Positive Rate (TPR) and True Negative Rate (TNR) in my project. The model, at the ideal threshold of 0.973, provides the best balance of sensitivity and specificity. This threshold aids the model's flawless performance by allowing the optimal trade-off between accurately identifying those not stressed, and identifying those who are.

7. FUTURE WORK

Several interesting avenues for future investigation arise from building on the present level of multimodal stress detection research. Given that most current studies have concentrated on acute stress reactions, longitudinal studies looking at chronic stress patterns should take top priority. Such investigations would enable the creation of more robust detection methods for chronic stress situations and assist to show how many physiological and behavioural indicators appear over extended stress periods. Another important study area is the improvement of stress detection algorithms' ecological validity and generalizability. To guarantee that models operate consistently outside laboratory settings, future efforts should emphasize gathering varied, realistic data across different demographics and locations. This strategy would solve the present restriction pointed out by Bueno et al. whereby laboratory-trained models underperformed in real-world settings. Open and comprehensible machine learning for stress detection should be much researched in the future. Predicted accuracy is important, because creating an explainable model that transparently reveals the elements of a stress detection model will develop user trust and offer insights to be applied relevantly to therapy. These open-source systems could help with intervention design and clearly identify the specific pressures on an individual. The most intriguing - and probably the most readable - elements would be a system that combined an automated-delivery intervention system with a stress detection system. Future research needs to explore closed-loop systems that detect stress levels in real time and

provide rapid but focused responses that are contextualized for individuals and their settings. The role that technologies such as this could have on stress management practices is possibly transformative by offering users 'just-in-time' help when they might need it most. Future work should also consider some of the technical issues associated with multimodal fusion in order to improve more complex algorithms for the temporal alignment of multiple data streams and managing missing modalities in real-world contexts. Improvements in federated learning and transfer learning methods may allow for stress detection models to be customized with information from wider populations while also conserving privacy.

8. CONCLUSION

In early self-report measurements to complex systems combining physiological, behavioural, and environmental data, multimodal stress detection has come a long way. This progression indicates our growing understanding of stress as a multidimensional, individualized phenomenon that occurs in diverse ways. The combination of wearable sensors, smartphone data, and new machine learning methodologies enables more precise and ecologically valid monitoring of stress in unstructured environments. While much progress has been made, multiple challenges lie ahead. While technically difficult and ever changing, the challenges of harmonizing many data streams with different temporal properties highlight the need for unique, individualized approaches to stress given the differences in reactions to stress. As these technologies become more prevalent in commercial, educational, and healthcare settings, ethical concerns regarding data privacy and security also deserve serious attention. In the future, stress detection will involve systems that also provide timely solutions and relevant information alongside accurate identification of stress levels. Innovative technology has great potential to improve total health; we've already seen this capability to intervene early with stress related illnesses and change the way we manage stress.

REFERENCES:

[1] S. Cohen, D. Janicki-Deverts, and G. E. Miller, "Psychological stress and disease," JAMA, vol. 298, no. 14, pp. 1685-1687, 2007.
[2] T. Karatekin, "The effects of self-report measures and anxiety on stress: A meta-analytic review," Journal of Psychosomatic Research, vol. 122, pp. 68-76, 2019.
[3] P. Schmidt, A. Reiss, R. Dürichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in Proc. 20th ACM Int. Conf. Multimodal Interaction, 2018, pp. 400-408.
[4] J. Hernandez, R. R. Morris, and R. W. Picard, "Call center stress recognition with person-specific models," in Proc. Int. Conf. Affective Computing and Intelligent Interaction, 2011, pp. 125-134.
[5] M. D'Mello and A. Kappas, "The moment-to-moment predictability of emotional and physiological states," Emotion Review, vol. 12, no. 2, pp. 98-123, 2020.
[6] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," Frontiers in Public Health, vol. 5, p. 258, 2017.
[7] H. F. Posada-Quintero and K. H. Chon, "Innovations in electrodermal activity data collection and signal processing," Frontiers in Physiology, vol. 11, p. 53, 2020.
[8] N. Sharma and T. Gedeon, "Objective measures, sensors and computational techniques for stress recognition and classification: A survey," Computer Methods and Programs in Biomedicine, vol. 108, no. 3, pp. 1287-1301, 2012.

[9] Y. Cho, N. Bianchi-Berthouze, and S. J. Julier, "DeepBreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings," in Proc. 7th Int. Conf. Affective Computing and Intelligent Interaction, 2017, pp. 456-463.
[10] Z. N. Karam et al., "Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2014, pp. 4858-4862.
[11] D. McDuff, A. Karlson, A. Kapoor, A. Roseway, and M. Czerwinski, "AffectAura: An intelligent system for emotional memory," in Proc. SIGCHI Conf. Human Factors in Computing Systems, 2012, pp. 849-858.
[12] P. Schmidt et al., "Context-aware smartphone sensing for mental health diagnosis and monitoring," IEEE Pervasive Computing, vol. 18, no. 1, pp. 74-83, 2019.
[13] N. Jaques, S. Taylor, A. Sano, and R. Picard, "Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction," in Proc. 7th Int. Conf. Affective Computing and Intelligent Interaction, 2017, pp. 202-208.
[14] K. Hovsepian et al., "cStress: Towards a gold standard for continuous stress assessment in the mobile environment," in Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Computing, 2015, pp. 493-504.
[15] B. Zhang, Y. Ren, X. Cui, B. Zong, Y. Ni, and M. Chen, "Multi-stage attention network for stress recognition from multimodal data," IEEE Trans. Affective Computing, vol. 12, no. 4, pp. 967-978, 2021.
[16] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," IEEE J. Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301-1309, 2017.
[17] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," IEEE Trans. Biomedical Engineering, vol. BME-32, no. 3, pp. 230-236, 1985.
[18] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," Journal of Neuroscience Methods, vol. 190, no. 1, pp. 80-91, 2010.
[19] P. Ekman and W. V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto: Consulting Psychologists Press, 1978.
[20] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proc. 34th Int. Conf. Machine Learning, 2017, pp. 3319-3328.
[21] P. Schmidt et al., "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in Proc. 20th ACM Int. Conf. Multimodal Interaction, 2018, pp. 400-408.
[22] S. Koldijk, M. A. Neerincx, and W. Kraaij, "Detecting work stress in offices by combining unobtrusive sensors," IEEE Trans. Affective Computing, vol. 9, no. 2, pp. 227-239, 2018.
[23] R. Wang et al., "StudentLife: Assessing mental health, academic performance and behavioural trends of college students using smartphones," in Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Computing, 2014, pp. 3-14.
[24] 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," Behaviour Research Methods, vol. 53, pp. 1689-1696, 2021.
[25] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. P. Morency, "Open Face 2.0: Facial behaviour analysis toolkit," in Proc. 13th IEEE Int. Conf. Automatic Face & Gesture Recognition, 2018, pp. 59-66.

[26] B. McFee et al., "librosa: Audio and music signal analysis in Python," in Proc. 14th Python in Science Conf., 2015, pp. 18-24.