Multimodal Approaches for Stress Recognition: A Comparative Study Using the StressID Dataset

Chia-Yun Lee
The Department of Data Science at Soochow University, Taiwan
jessicaleej0033@gmail.com

Matúš Pleva
Department of Electronics and Multimedia Communications, Technical University of Košice, Slovakia
matus.pleva@gmail.com

Daniel Hládek
Department of Electronics and Multimedia Communications, Technical University of Košice, Slovakia
daniel.hladek@tuke.sk

Ming-Hsiang Su
The Department of Data Science at Soochow University, Taiwan
huntfox.su@gmail.com

Abstract

Mental health concerns have garnered increasing attention, highlighting the importance of timely and accurate identification of individual stress states as a critical research domain. This study employs the multimodal StressID dataset to evaluate the contributions of three modalities—physiological signals, video, and audio—in stress recognition tasks. A set of machine learning models, including Random Forests (RF), Support Vector Machines (SVM), Multi-Layer Perceptrons (MLP), and K-Nearest Neighbors (KNN), were trained and tested with optimized parameters for each modality. In addition, the effectiveness of different multimodal fusion strategies was systematically examined. The unimodal experiments revealed that the physiological modality achieved the highest performance in the binary stress classification task (F1-score = 0.751), whereas the audio modality outperformed the others in the three-class classification task (F1-score = 0.625). In the multimodal setting, feature-level fusion yielded stable improvements in the binary classification task, while decision-level fusion achieved superior performance in the three-class classification task (F1-score = 0.65). These findings demonstrate that multimodal integration can substantially enhance the accuracy of stress recognition. Future research directions include incorporating temporal modeling and addressing data imbalance to further improve the robustness and applicability of stress recognition systems.

Keywords: Stress Detection, Multimodal Machine Learning, Audio-Visual Features, Feature-Level Fusion.

1 Introduction

As the pace of modern society accelerates and life pressures intensify, mental health is getting more attention. The World Health Organization (WHO) designates October 10th each year as World Mental Health Day, emphasizing that mental health is a fundamental human right and urging all sectors to address psychological issues and provide necessary resources. However, in high-pressure environments, many individuals struggle to recognize and manage stress, which can gradually accumulate and lead to more serious mental health challenges.

Stress is essentially a state, both mental and physical, that happens when people feel the demands of their environment are beyond their ability to cope, threatening their well-being (Lazarus, R.S. et al., 1984). It is a dynamic and interactive process that involves the individual's cognitive appraisal and coping strategies in response to stressors. Research has shown that stress has both direct and indirect effects on mental health, particularly through the regulation of psychological states via negative emotions (Moreta-Herrera, R., et al, 2023). While moderate stress can foster adaptation and motivation, prolonged and unmanaged stressors may negatively impact the nervous system, mental health, and behavior patterns (Hsu, Y. F., 2021).

Taking the campus as an example, students face multiple pressures from academic work, interpersonal relationships, and future development, which often brings their mental health issues into the news spotlight. According to statistical data released by the Ministry of Education's Campus Safety and Disaster Prevention Center in 2024（教育部校安通報中心, 2024）, suicide and self-harm incidents have ranked first in campus safety-related accidental reports for the past three years, accounting for 33% of all reported accidents. The number of deaths in higher education institutions remains high. In recent years, universities have begun implementing mental health leave, believing that it helps students with self-awareness and provides an opportunity for short-term adjustment, hoping to reduce the incidence of such incidents.

Currently, the assessment of psychological stress primarily relies on traditional questionnaire-based surveys (Scale, P.S., 1983). However, these methods are limited by their high subjectivity and lack of real-time responsiveness, which hinder the implementation of timely intervention strategies. Therefore, developing an objective and real-time stress monitoring technology has become a crucial research direction. Furthermore, existing research and datasets on stress detection face notable limitations, including small dataset sizes, a lack of diverse stress sources, and unimodal data constraints. To address these issues and advance the field of stress recognition, this study will utilize the rich resources of the StressID dataset (Chaptoukaev, H., et al., 2023). It aims to optimize and evaluate the performance of various unimodal and multimodal fusion models, with the goal of developing more objective and reliable stress identification techniques that can enhance mental health monitoring and intervention capabilities.

2 Related Literature

2.1 Stress Recognition Research

With the growing awareness of mental health, recent years have witnessed increasing research efforts dedicated to enhancing the accuracy of stress detection through a wide range of features and classification models. One notable contribution is the WESAD dataset introduced by Schmidt et al. (2018), which integrates multiple wearable sensor signals with emotion annotations and has since become a widely used benchmark for developing and evaluating multimodal stress recognition systems. Building on this resource, Abdelfattah et al. (2025) conducted a comparative analysis of machine learning and deep learning models using the WESAD dataset. Their findings suggest that deep learning methods provide superior cross-subject generalization but are computationally demanding, limiting their feasibility for real-time applications. In contrast, traditional machine learning models demonstrate greater computational efficiency and achieve high accuracy in personalized settings—reaching up to 99.8% F1 score—yet they suffer from limited generalizability. To address these shortcomings, ensemble learning has been highlighted as a promising strategy for enhancing both robustness and generalization in stress recognition. Extending this line of research, the present study explores the StressID multimodal dataset, with particular emphasis on evaluating the contributions of different modalities and investigating the impact of fusion strategies on model performance.

2.2 Classification Models for Stress Detection

To achieve this, a set of established machine learning and deep learning models is considered. Random Forest (RF) is an ensemble learning approach composed of multiple decision trees that improves classification stability and accuracy by employing a voting mechanism for both classification and regression tasks. Its performance depends on hyperparameters such as the number of estimators (n_estimators), the splitting criterion (criterion), and the maximum tree depth (max_depth), which are generally optimized through cross-validation. Support Vector Machine (SVM) is a supervised learning model that identifies the optimal hyperplane separating data points of different classes with the maximum margin, making it effective for classification tasks with well-defined decision boundaries. Its effectiveness relies on the selection of the kernel function, the regularization parameter (C), and the kernel coefficient (gamma). K-Nearest Neighbors (KNN) is a non-parametric, distance-based algorithm that classifies new data points by identifying the K nearest neighbors and applying majority voting, with hyperparameters including the number of neighbors (n_neighbors), the weighting scheme, and the neighbor computation algorithm. Multilayer Perceptron (MLP), a feedforward artificial neural network, is capable of modeling complex nonlinear relationships through an input layer, one or more hidden layers, and an output layer. Its performance is shaped by factors such as the activation function, learning rate, optimization algorithm, and hidden layer configuration. Finally, the Deep Belief Network (DBN), composed of stacked Restricted Boltzmann Machines (RBMs), is a deep generative model that performs unsupervised pretraining to capture hierarchical data representations, followed by supervised fine-tuning for classification. DBNs are particularly valued for their strong feature extraction capabilities, especially in handling structured and high-dimensional data.

3 Dataset Collection and Processing

3.1 Dataset Description

This study employs the StressID dataset, a multimodal resource integrating physiological signals, video, and audio recordings. Figure 1 illustrates the structure of the dataset. Data collection followed a rigorous and reproducible experimental protocol comprising 11 tasks organized into four main blocks: guided breathing, emotional video clips, seven interactive stress-inducing tasks, and a relaxation phase. These diverse tasks were designed to elicit varying stress responses among participants.

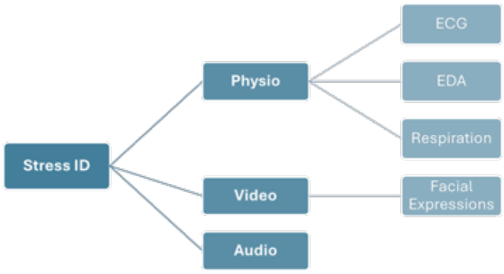

Figure 1: Multimodal Structure of the StressID Dataset

The experiment involved 65 healthy adult participants. Following each task, participants completed self-assessment questionnaires to report their perceived stress, relaxation, valence, and arousal levels. In this context, valence refers to the positive or negative emotional state experienced during a task, whereas arousal reflects emotional activation or engagement. These self-reported measures were subsequently used for supervised learning models to generate binary labels (stressed vs. not stressed) and ternary labels (relaxed, neutral, stressed).

All multimodal data were collected synchronously and processed through task segmentation and annotation procedures. The final StressID dataset comprises over 39 hours of annotated recordings, including 711 physiological signal recordings, 587 video segments, and 385 audio recordings. Its scale and diversity make the dataset one of the most extensive publicly available stress identification resources suitable for unimodal and multimodal research.

3.2 Dataset Processing

The StressID dataset provides baseline stress classification models in unimodal and multimodal settings. This section describes the feature extraction and preprocessing procedures for the three unimodal data types, inputs for subsequent machine learning models. For physiological signals, 35 features were extracted from the electrocardiogram (ECG), 23 from electrodermal activity (EDA), and 40 from respiration signals. All signals were first processed using a Butterworth filter to reduce high-frequency noise and baseline drift. Extracted features include statistical and physiological measures such as heart rate variability (HRV), skin conductance level (SCL), skin conductance response (SCR), and respiratory rate variability (RRV), all intended to quantify the participants' physiological states.

Video data were processed using the OpenFace library to extract facial features, including Action Units (AUs) and eye gaze trajectories. These features' mean and standard deviation were calculated to capture facial expressions and gaze dynamics, resulting in an 84-dimensional feature vector for each video segment. Audio recordings were down-sampled to 16 kHz, and amplitude-based Voice Activity Detection (VAD) was applied to remove non-speech segments. Handcrafted features were extracted, including Mel-frequency cepstral coefficients (MFCCs) and their derivatives, spectral centroid, and other spectral features, forming a 114-dimensional feature vector. Additionally, speech embeddings were obtained from the pre-trained Wav2Vec 2.0 (W2V) model. Embeddings were extracted every 20 milliseconds and averaged over time to generate a 513-dimensional representation per utterance, capturing variations in frequency, energy, and speech rhythm.

All features, except those extracted by Wav2Vec 2.0 (which were classified using a linear classifier), were standardized and used as inputs to machine learning models, including Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP), and K-Nearest Neighbors (KNN). These models were trained and evaluated under various parameter configurations to predict stress-related labels, including binary and three-class classification.

Multimodal integration strategies were also explored to improve classification performance. The first approach, feature-level fusion, concatenates features from each modality into a single high-dimensional vector, which is then used as input to machine learning models. The second approach, decision-level fusion, trains independent models for each modality and combines their predictions using ensemble rules such as summation, product, averaging, or maximum to generate the final decision.

A notable challenge in the StressID dataset is class imbalance, particularly in audio data, as speech tasks are often associated with elevated stress levels. SMOTE was applied to balance binary-class audio data and the multimodal training set to address this. However, in three-class audio classification, the "relaxation" category is underrepresented due to the limited presence of audio during relaxation tasks. The scarcity of relaxed audio samples and the absence of characteristic relaxed speech features limit the effectiveness of resampling in this scenario.

3.3 Model Performance Evaluation

To assess the classification performance of the model in stress detection tasks, this study evaluated the model on the test dataset using weighted F1-score and balanced accuracy. The evaluation metrics are defined as follows:

$$F1_{weighted} = \sum_{i=1}^{n} w_i \times F1_i \tag{1}$$

$$Balanced\ Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FN_i} \tag{2}$$

These metrics were used to measure the model's performance in both binary and multi-class stress classification tasks across different modalities. The weighted F1-score emphasizes classification accuracy while taking the class distribution into account. On the other hand, balanced accuracy mitigates the influence of class imbalance by averaging the recall across all classes, providing a fairer assessment of the model's ability to recognize each class equally.

4 Experimental Results and Discussion

This study implements various classification models using Python's scikit-learn library, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), and Multi-layer Perceptron (MLP), with parameter tuning for comparison. The MLP model utilizes multiple combinations of hidden layers, SVM employs the RBF kernel with C-value adjustment and a fixed gamma of 0.00714, KNN investigates the effect of different numbers of neighbors, and RF investigates the effect of different tree depths. All models were evaluated using 10 random splits to ensure robustness and reliable performance estimation.

As shown in Table 1 and Table 2, for the binary-stress classification task, the Physio modality with Random Forest achieved the best performance (F1 = 0.751) with a maximum tree depth of 20. For the three-class affect classification task, the Audio modality with SVM performed best (F1 = 0.577) with a C-value of 10. Overall, all models performed better on the binary classification task, with the Physio modality demonstrating the best binary performance, while the Audio modality outperformed others in the three-class setting.

Table 1. Comparison of unimodal baseline performances on the binary-stress classification task.

	Binary-stress	
Classifier	**F1-score**	**Accuracy**
Video. AUs + RF	0.702±0.03	0.703±0.03
Video. AUs + SVM	0.701±0.03	0.701±0.02
Video. AUs + KNN	0.706±0.03	0.705±0.03
Video. AUs + MLP	**0.708±0.04**	**0.708±0.04**
Audio. HC features + RF	0.689±0.07	0.629±0.07
Audio. HC features + SVM	0.713±0.05	0.664±0.05
Audio. HC features + KNN	0.576±0.04	0.627±0.03
Audio. HC features + MLP	**0.718±0.07**	**0.671±0.07**
W2V 2.0 classifier + MLP	**0.725±0.05**	**0.667±0.05**
Physio. HC features + RF	**0.751±0.03**	**0.744±0.03**
Physio. HC features + SVM	0.733±0.03	0.725±0.03
Physio. HC features + KNN	0.696±0.04	0.689±0.04
Physio. HC features + MLP	0.712±0.03	0.709±0.03

Table 2. Comparison of unimodal baseline performances on the affect3-class classification task.

	Affect3-class	
Classifier	**F1-score**	**Accuracy**
Video. AUs + RF	0.557±0.05	0.555±0.05
Video. AUs + SVM	**0.565±0.03**	**0.559±0.03**
Video. AUs + KNN	0.563±0.04	0.558±0.04
Video. AUs + MLP	0.564±0.03	0.557±0.04
Audio. HC features + RF	0.515±0.07	0.478±0.06
Audio. HC features + SVM	**0.577±0.04**	**0.535±0.06**
Audio. HC features + KNN	0.526±0.06	0.491±0.07
Audio. HC features + MLP	0.558±0.03	0.519±0.07
W2V 2.0 classifier	**0.625±0.05**	**0.564±0.05**
Physio. HC features + RF	0.569±0.02	0.565±0.02
Physio. HC features + SVM	**0.576±0.04**	**0.574±0.04**
Physio. HC features +KNN	0.561±0.02	0.552±0.03
Physio. HC features + MLP	0.537±0.04	0.53±0.04

In the multimodal analysis, three approaches are covered: unimodal models, feature-level fusion, and decision-level fusion. Additionally, various classifiers (SVM, RF, MLP, KNN) are compared.

According to Table 3, the best unimodal performance is achieved by Audio + SVM (F1 = 0.73), with parameters C = 10. Among the fusion strategies, feature-level fusion with MLP (1 hidden layer, 100 units) or SVM (C = 1.0, gamma = 0.00714) achieved the best performance in the binary-stress task (F1 = 0.72), slightly outperforming the decision-level fusion results.

Table 3. Comparison of multimodal baseline performances on the binary-stress classification task.

	Binary-stress	
Classifier	**F1-score**	**Accuracy**
Video. + SVM	0.7±0.04	0.64±0.05
Audio. + SVM	**0.73±0.02**	**0.68±0.02**
Physio. + RF	0.71±0.04	0.63±0.04
Feature level fusion + MLP	**0.72±0.06**	**0.66±0.07**
Feature level fusion + KNN	0.61±0.07	0.63±0.07
Feature level fusion + RF	0.67±0.05	0.57±0.03
Feature level fusion + DBN	0.63±0.05	0.57±0.04
Feature level fusion + SVM	**0.72±0.06**	**0.66±0.06**
RF + Sum level fusion	0.72±0.03	0.65±0.03
RF + Product level fusion	0.72±0.03	0.64±0.03
RF + Average level fusion	0.72±0.03	0.65±0.03
RF + Maximum level fusion	0.72±0.04	0.63±0.04

In contrast, in the affect3-class task in Table 4, the multimodal fusion strategies clearly outperform the unimodal models. Among them, the Decision-level fusion with RF (max_depth = 25, random_state = 0) + Sum/Average achieved the best performance, with F1 = 0.65. Feature-level fusion with MLP (F1 = 0.62) also showed a close performance, demonstrating practical potential.

Table 4. Comparison of multimodal baseline performances on the affect3-class classification task.

	Affect3-class	
Classifier	**F1-score**	**Accuracy**
Video. + SVM	**0.58±0.06**	**0.55±0.06**
Audio. + SVM	0.52±0.06	0.49±0.04
Physio. + RF	0.52±0.05	0.5±0.06
Feature level fusion + MLP	**0.62±0.05**	**0.61±0.04**
Feature level fusion + KNN	0.53±0.04	0.56±0.06
Feature level fusion + RF	0.54±0.06	0.49±0.06
Feature level fusion + DBN	0.34±0.11	0.35±0.04
Feature level fusion + SVM	0.57±0.05	0.51±0.04
RF + Sum level fusion	**0.65±0.06**	**0.6±0.06**
RF + Product level fusion	0.64±0.06	0.6±0.06
RF + Average level fusion	0.65±0.06	0.6±0.06
RF + Maximum level fusion	0.63±0.04	0.59±0.04

Overall, multimodal fusion strategies outperform unimodal models in both tasks. Feature-level fusion is more suitable for the binary-stress task, while Decision-level fusion shows its advantage in the affect3-class task. In comparison, KNN and DBN underperformed overall, with both accuracy and stability being relatively low.

5 Conclusion and Future Work

This study investigated unimodal and multimodal approaches for stress recognition using the StressID dataset. The results demonstrate the effectiveness of multimodal fusion, with feature-level fusion providing stable performance in binary stress classification, while decision-level fusion achieves superior performance in the three-class affective classification task. Despite these promising outcomes, challenges remain, particularly regarding class imbalance. The underrepresentation of the “relaxation” category adversely affects the performance of three-class classification models. Future research should explore strategies to mitigate these imbalances and consider the incorporation of temporal models, such as LSTM, GRU, or Transformer architectures, to better capture the dynamic nature of stress responses. Additionally, further investigation into the feasibility of these models for real-time monitoring and practical deployment is essential to enhance the timeliness, robustness, and overall accuracy of mental health interventions.

References

Lazarus, Richard S., and Susan Folkman. 1984. *Stress, appraisal and coping.* New York: Springer.

Moreta-Herrera, Ricardo, D. Zumba-Tello, J. de Frutos-Lucas, S. Llerena-Freire, A. Salinas-Palma, and A. Trucharte-Martínez. 2023. The role of negative affects as mediators in the relationship between stress and mental health in Ecuadorian adolescents. *Health Psychology Report*, 11(3), 241-251.

Hsu, Yu-Fang. 2021, June. A study of the relationships between perceived stress and the quality of life for clinical nurse preceptors (Master's thesis). Nanhua University, Chiayi County, Taiwan.

教育部校園安全暨災害防救通報處理中心資訊網. 2024, January.「教育部 111 年各級學校校園安全及災害事件分析報告」.

Scale, P. S. (1983). Perceived Stress Scale.

Chaptoukaev, H., V. Strizhkova, M. Panariello, B. D'alpaos, A. Reka, V. Manera, S. Thümmler, E. Ismailova, N. Evans, F. F. Bremond, M. Todisco, M. Zuluaga, and L. M. Ferrari. 2023. StressID: A multimodal dataset for stress identification. In *NeurIPS 2023 - 37th Conference on Neural Information Processing Systems*, NIST, New Orleans, United States.

Philip Schmidt, Attila Reiss, Robert Dürichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. *In Proceedings of the 2018 International Conference on Multimodal Interaction (ICMI '18)*, pages 400–408. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3242969.3242985

Eman Abdelfattah, Shreehar Joshi, and Shreekar Tiwari. 2025. Machine and deep learning models for stress detection using multimodal physiological data. *IEEE Access,* 13:2154–2166. https://doi.org/10.1109/ACCESS.2024.3525459

StressID: A Multimodal Dataset for Stress Identification, Access date: 2025/02/07. https://github.com/robustml-eurecom/stressID?tab=readme-ov-file