ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 2, February 2026

Multimodal Stress Detection Using Deep Learning

[1] Satish Singh, [2] Abhijit Sharma, [3] Nandini Singh, [4] Pradeep Chauhan, [5] Sameer Awasthi

[1] [2] [3] [4] [5] Department of Computer Science and Engineering (AIML), Bansal Institute of Engineering and Technology, Lucknow, Uttar Pradesh, India

Corresponding Author Email: [1] satishsinghg07@gmail.com, [2] abhijitsharma.ab@gmail.com, [3] nandinisingh57182@gmail.com, [4] pradeepchauhanp20@gmail.com, [5]sameer.awasthi@gmail.com

***Abstract*— Stress, which can cause physical and mental disorders, is one of the most significant health problems of the twentieth century. Traditional approaches to stress detection are based on self-reported questionnaires or one-dimensional measurements (such as heart rate or face analysis), which sometimes produce inaccurate or incorrect results. To overcome this constraint, this study proposes FusionNet, a multimodal deep learning system designed to combine different physiological inputs, facial expressions, and audio for stress detection. By combining data from the DEAP and WESAD datasets, the proposed approach uses the complementing characteristics of many modalities to increase the precision of classification. Our outcomes demonstrate the reliability and resilience of our approach by demonstrating that multimodal fusion performs noticeably better than unimodal models, with accuracy rising from 70% to 83%.**

***Index Terms*— Deep Learning, DEAP, Emotion, Multi modal Learning, Physiological Signals, Recognition, Stress Detection, WESAD.**

I. INTRODUCTION

Stress represents a common human experience, with wide-ranging implications in terms of a person's long-term physical health, emotional balance, professional efficiency, and cognitive ability. Stress-related disorders, including anxiety, depression, and cardiovascular disease, have been described by WHO as among the leading causes of disability worldwide [1]. Several studies have demonstrated that long-term stress alters a person's decision-making, memory, creativity, and communication skills in an academic, corporate, and healthcare setting. Against this background, the early detection of stress has become fundamental in developing preventive care plans, personalized treatments, and real-time monitoring systems.

Traditional stress evaluation methods have been based on self-reporting tools, which include psychological questionnaires, stress inventories, and subjective interviews. These techniques, though widely used, are constrained by user prejudice, uneven self-awareness, variation in emotional states, and recollection errors. This gulf has thus driven the usage of technology in stress detection techniques using wearable sensors and devices capable of recording physiological data in real time based on IoT devices, including heart rate variability (HRV), electro-dermal activity (EDA), electromyography (EMG), and respiration rate. These physiological systems have been very effective, but they sometimes have problems when the signals are too weak, noisy, or affected by environmental factors such as motion or temperature.

Recent analyzes reveal that stress is a physiological but multimodal response that appears almost simultaneously in biometric signals, changes in voice pitch, and facial expressions. For example, subjects under stress commonly exhibit microex pressions, reduced blinking frequencies, trembling voices, prolonged speech pauses, and changed skin conductance. Since the variance in each modality reflects only part of the response to stress, predictions are incomplete or inconsistent when only one modality is used to make the prediction, such as voice, face, or sensors [2]. This encourages researchers to investigate multimodal learning, where inputs obtained from different sources are fused to improve accuracy and robustness. Multimodal fusion overcomes challenges related to noisy signals, missing data, and physiological differences between individuals. Continuous, real-time stress detection research has also been substantially facilitated by access to reputable datasets, including DEAP and WESAD [3]. We propose a FusionNet-based multimodal deep learning architecture that fuses physiological, auditory, and facial sensor input in one model to overcome the drawbacks of unimodal systems. The proposed architecture will collect complementing stress indicators across modalities using 1D-CNNs, lightweight face encoders, and audio feature extraction modules to generate a more reliable noise-resistant and person-independent stress classification mechanism.

II. PROBLEM STATEMENT

As the name suggests, unimodal stress detection systems rely on any one input: physiological signals, facial expressions, or audio features. Thus, performance under real conditions cannot be reliably established within any of these systems. Each modality reflects certain patterns of stress, but since this is a multidimensional phenomenon, no single signal generalizes well between individuals or between environments.

Physiological signals such as EDA, ECG, and respiration are easily influenced by the non-stress factors of movement,

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 2, February 2026

temperature, or exercise, which might cause potential false alarms. Poor lighting, occlusion, and low-quality cameras limit the analyzes on the face; again, detecting stress relying on an audio-based approach is not reliable in an acoustically noisy and complex environment. All the modality-specific issues lead to the lack of robustness that will make these unimodal systems mentioned unsuitable for practical deployment. Therefore, it is essential to develop a multimodal framework that incorporates face, audio, and sensor data. A fusion-based approach integrates these complementary signals to overcome the weaknesses of each of the individual modalities, offering better reliability and generalization, and hence higher accuracy in real-world scenarios.

III. RELATED WORK

Numerous research is being carried out to find people who are sad or under stress. Table 1 displays some prior relevant research on the stress detection scheme, with some researchers using publicly available datasets and others gathering their own dataset [3] – [12] (Related Work continues on the next page with Table I)

IV. DATASETS

A. WESAD Dataset

The physiological data from the WESAD (Wearable Stress and Affect Detection) data set recorded using chest and wrist sensors. The modalities it includes are ECG, EDA, and EMG, body temperature, and respiration. The data set labels stress as, amusement and neutral states.

- Subjects: 15 participants
- Modalities: ECG, EDA, EMG, respiration, temperature
- Sampling rate: 700 Hz (chest), 32 Hz (wrist)

B. DEEP Dataset

The DEAP data set is widely used for the recognition of emotions and stress. Include EEG signals, peripheral physiological data, and self-assessment labels for arousal, valence, and dominance.

- Subjects: 32 participants
- Modalities: EEG (32 channels), EDA, HR, respiration
- Sampling rate: 512 Hz

C. Preprocessing

After extraction by MTCNN, the facial frames are resized to 224×224, normalized, and then sampled at fixed intervals.

It is done by noise reduction, framing, and Mel-spectrogram computation with 128 Mel-filter. ECG, EDA and EMG signals are filtered, then normalized and divided into 10-second windows. First, all modalities are aligned on one timeline and temporally synchronized before fusion.

V. METHODOLOGY

A. Model Architecture (FusionNet)

FusionNet uses a late-fusion multi-modal setup with three separate encoders. One has a ResNet18 facial encoder, another's a 2D-CNN audio encoder that works with Melspectrograms, and the last is a 1D-CNN encoder for physiological data such as ECG and EDA signals. Each encoder pulls out its own set of high-level features—Ff for the face, Fa for the audio, and Fs for the signals. Then, it just stacks these features together into one big multi-modal embedding. From there, the model passes this fused representation through dense layers that predict stress levels. The real strength here is how these different types of features cover each other. So, even if one sensor gets noisy, blocked, or acts up, the whole system stays solid.

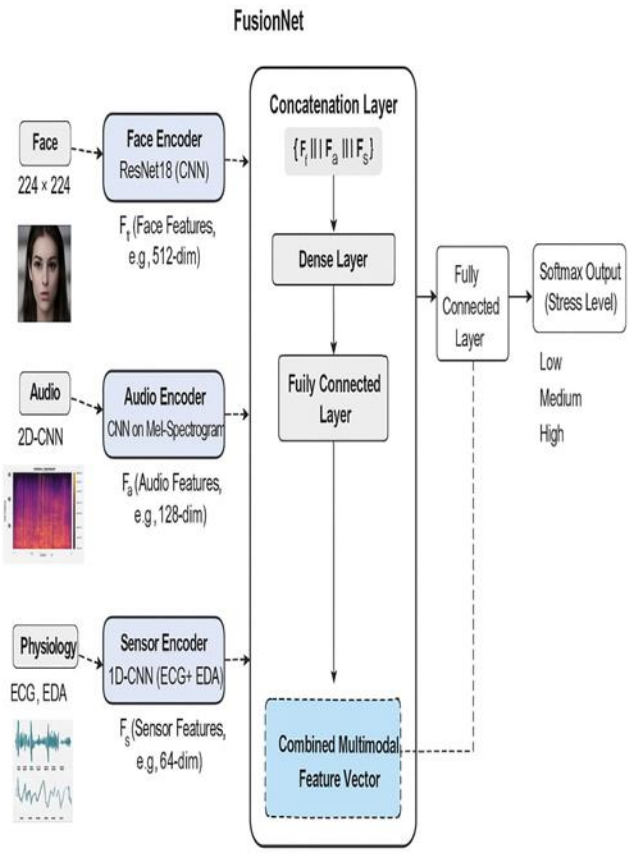

Fig. 1. Proposed FusionNet architecture combining facial, audio, and sensor modalities.

Equation: $h = f_{fusion}\,(f_{face}\,(x_f), f_{audio}(x_a), f_{sensor}(s))$

B. Training Setup

- Optimizer: AdamW
- Learning rate: 1e-3
- Loss: CrossEntropy
- Batch size: 8
- Epochs: 30
- Framework: Pytorch

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 2, February 2026

Table 1. Previous Related Research Works

[3]	Presenting WESAD, a Multimodal Dataset for Wearable Affect and Stress Identification	WESAD dataset is available to the public	RF, DT, AdaBoost, KNN, LDA, and SVM were used to obtain accuracy of 80% for three class and 93% for two classes
[4]	Machine Learning based Speech Analysis for Stress Identification	Publicly Available Data Ryerson Audio-Visual Database of emotional Speech and Song (RAVDESS) dataset	CNN- 94.26%-94.3% was attained
[5]	Machine Learning and Deep Learning for Stress identification Using Multimodal Physiological Data	WESAD public dataset Accomplished	84.32% and 95.21% accuracy were attained with RF, DT, AdaBoost, KNN, LDA, SVM and DL
[6]	Categorization of Physiological Signals for IOT Based Emotions Recognition	SAID Dataset: Private Dataset used the ECG and GSR SAID datasets to compile my own dataset	ANN- 75.8% mean accuracy and 11.38% accuracy standard deviation were attained
[7]	Using a Physical Activity Tracker and Machine Learning to Detect Stress	Confidential Data Using a FIT BIT gadget, I collected my own dataset and used ANOVA for analysis	AIC- 782.8842 (Logit model) AIC- 786.8999 (Complementary Log- Log model) AIC- 781.6256 (Probit model) A lower AIC indicates the better of model
[8]	Recognition of Emotion Using Multichannel Physiological Signals and Comprehensive Nonlinear Processing	Confidential Data gathered a dataset using the ECG, GSR, EMG	KPCA reduced the features and GBDT for classifier; 93.42% accuracy was attained
[9]	Stress detection combining physiological markers and machine learning-based signal processing	Confidential Data gathered me own dataset using heart rate, EMG, GSR, hand and foot data, respiration, and WEKA classification	KNN classifier- 92.06% accuracy was attained SVM- 96.82% accuracy was attained
[10]	IOT and Machine Learning for stress detection and prediction	Confidential Data gathered me own dataset and used Python for classification	SVM-68% LR-66%
[11]	Wearable physiological sensors for Stress detection	Confidential data gathered personal data from BN PPGED	SVM- 82% accuracy attained
[12]	Heart Rate Variability - based Emotion Recognition	The MAHNOB dataset is publicly available	SVM – 48.5% accuracy attained

Abbreviations:

RF = Random Forest, SVM = Support Vector Machine, KNN = k-Nearest Neighbors, DT = Decision Tree, AdaBoost = Adaptive Boosting, LDA = Linear Discriminant Analysis, DL = Deep Learning, LR = Logistic Regression, CNN = Convolutional Neural Network, AIC = Akaike Information Criterion, ANN = Artificial Neural Network, KPCA = Kernel Principal Component Analysis, GBDT = Gradient Boosted Decision Tree

C. Evaluation Metrics

We evaluate models using:

Accuracy: The fraction of correctly classified instances relative to all instances is known as accuracy.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- Use in this work: Indicates the overall effectiveness of FusionNet in distinguishing stress levels across the data set. However, in imbalanced datasets, accuracy alone may be misleading.

Precision: Precision quantifies the proportion of samples that are correctly predicted to belong to a particular class.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

- Use in this work: High precision ensures that when the model predicts “High Stress”, it is rarely wrong. This is particularly useful in healthcare applications where

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 2, February 2026

false alarms could cause unnecessary concern.

Recall (Sensitivity): Recall measures how many of the actual positive samples are correctly detected.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

- Use in this work: High recall ensures that most stress instances are detected. Missing actual stressed individuals (false negatives) can be dangerous in real-world monitoring systems.

F1-Score: The harmonic mean of precision and recall, balancing both metrics is known as F1-Score.

$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (4)$$

- Use in this work: Since stress detection needs both low false alarms (precision) and low missed cases (recall), the F1-score offers a fair assessment of performance.

ROC-AUC (Receiver Operating Characteristic – Area Under Curve): The trade-off between the true positive rate (recall) and the false positive rate is measured by ROC-AUC.

- Use in this work: A high ROC-AUC score indicates that FusionNet can effectively separate stressed and non-stressed states across different thresholds, providing robustness for deployment.

VI. EXPERIMENTAL SETUP

- Hardware: Intel i5 CPU, 16 GB RAM, NVIDIA RTX 2050 GPU
- Software: Python 3.10, PyTorch 2.0, Streamlit for interface
- Cross-validation: 5-fold subject-independent split

VII. RESULTS

A. Unimodal Baselines

Table 2 Shows the performance of individual modalities.

Table 2. Unimodal Stress Detection Performance

Modality	Accuracy (%)	F1-score
Face (CNN)	72	0.71
Audio (CNN+LSTM)	60	0.58
Sensors (1D CNN)	74	0.72

B. FusionNet Performance

Table III compares the proposed FusionNet with unimodal baselines.

Table 3. Multimodal Fusion Vs. Unimodal Baselines

Model	Accuracy (%)	F1-score
Face only	72	0.71
Audio only	60	0.59
Sensors only	74	0.72
FusionNet (All)	83	0.82

C. Confusion Matrix insights

- FusionNet reduces misclassification between medium and high stress.
- Sensors dominate unimodal accuracy, but fusion boosts robustness.

D. Visualizations

- With a loss of 0.526, the model obtained an accuracy of 83%, precision of 0.689, recall of 0.83, and F1-score of 0.753. The confusion matrix reveals that the model performs well in identifying the Low stress class but struggles to differentiate between Medium and High stress levels. This highlights the need for more balanced and diverse training samples across stress categories.
- The model achieved an accuracy of 80%, precision of 0.64, recall of 0.80, and F1-score of 0.71 with a loss of 0.547. From the confusion matrix, it is observed that most samples are predicted as Class 0, indicating a class imbalance or overfitting towards the dominant class. Future work could include class-weight adjustment or data augmentation to improve multi-class performance.

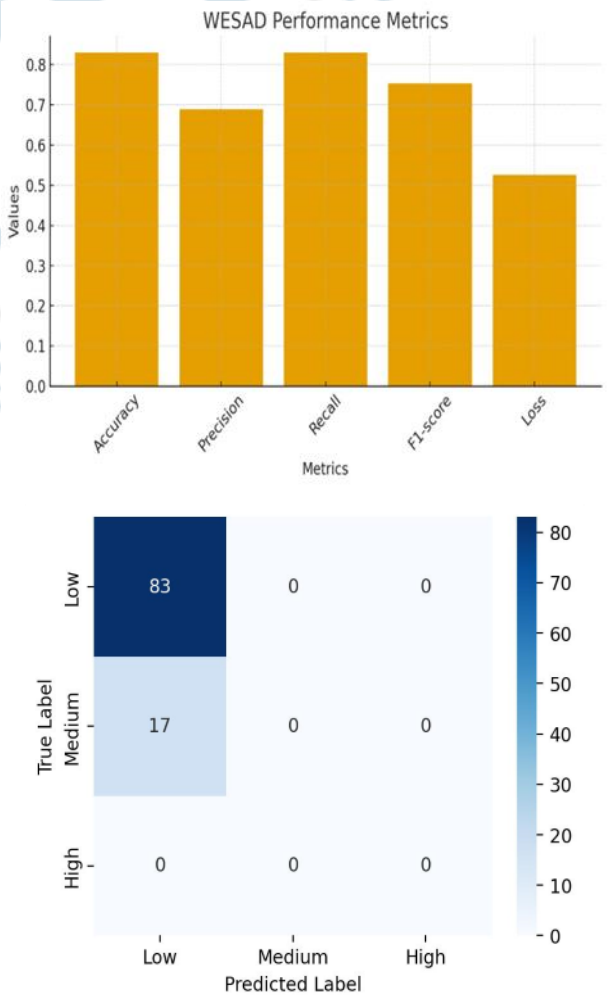

Fig. 2. Confusion matrix Performance metrics of the model trained on the WESAD dataset.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 2, February 2026

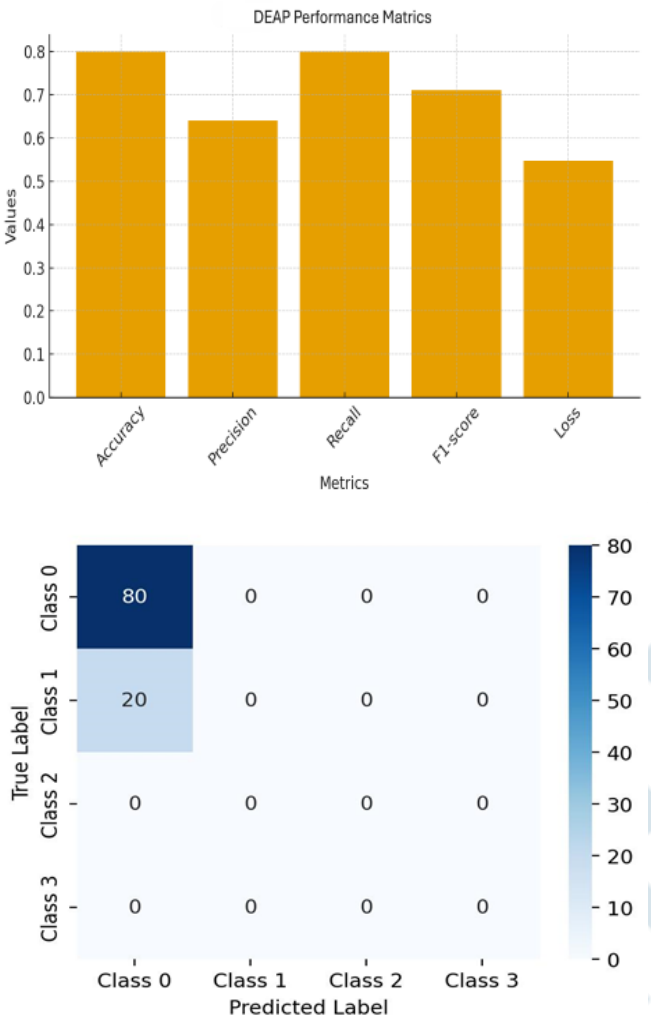

Fig. 3. Confusion matrix Performance metrics of the model trained on the DEAP dataset

VIII. DISCUSSION

These results clearly indicate the advantages of multi-modal fusion. While physiological sensors provide strong signals, integrating facial and audio features helps the system tackle ambiguous cases with greater robustness.

For example, physical activity misclassified as stress by sensors alone are corrected when facial expressions and speech cues indicate a neutral state.

In addition, FusionNet exhibits scalability and generalization across the datasets and could be used in healthcare and workplace monitoring.

IX. CONCLUSION AND FUTURE WORK

This work proposes FusionNet, a multimodal deep learning architecture that performs stress detection using facial, audio, and physiological signals. Evaluation of WESAD and DEAP datasets confirm that multimodal fusion significantly outperforms unimodal approaches, achieving an accuracy of 83%.

Future Work

- Real-time deployment in wearable devices and smartphones.
- Integration with telemedicine and workplace wellness platforms.
- Explainable AI methods to improve interpretability of stress predictions.
- Expansion to multimodal datasets including text and contextual information.

REFERENCES

[1] R. Ahuja and A. Banga, “Mental stress detection in university students using machine learning algorithms,” Procedia Computer Science, vol.152, pp. 349–353, 2019.

[2] W. Zhang, C. Li, and Y. Huang, “Multimodal emotion recognition with audio, video, and physiological signals,” in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 3457–3461

[3] Schmidt P Reiss A Duerichen R Marberger C and Laerhoven K V 2018 Proceedings of the 20th ACM International Conference on Multimodal Interaction 400-408

[4] Vaikole S Mulajkar S More A Jayaswal P and Dhas S 2020 International Journal of Creative Research Thoughts (IJCRT) 8 5 2239-44

[5] Bobade P and Vani M 2020 Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020) 51-58

[6] Tiwari S Agarwal S Muhammad Syafrullah and Adiyarta K 2019 Proc. EECSI 2019 - Bandung, Indonesia 6 107-111

[7] Ferdinando H Ye L Seppnen T and Alasaarela E 2014 Australian Journal of Basic and Applied Sciences. 8 14 50-55

[8] Padmaja B Rama Prasad V V and Sunitha K V N 2018 International Journal of Machine Learning and Computing 8 1 33-38

[9] Zhang X Xu C Xue W Hu J He Y and Gao M 2018 Sensors 2018 18 11 3886

[10] Ghaderi A Frounchi J and Farnam A 2015 22nd Iranian Conference on Biomedical Engineering (ICBME) 93-98

[11] Pandey P S 2017 2017 17th International Conference on Computational Science and Its Applications (ICCSA)

[12] Sandulescu V Andrews S Ellis D Bellotto N and Mozos O M 2015 International Work-Conference on the Interplay Between Natural and Artificial Computation 526-532