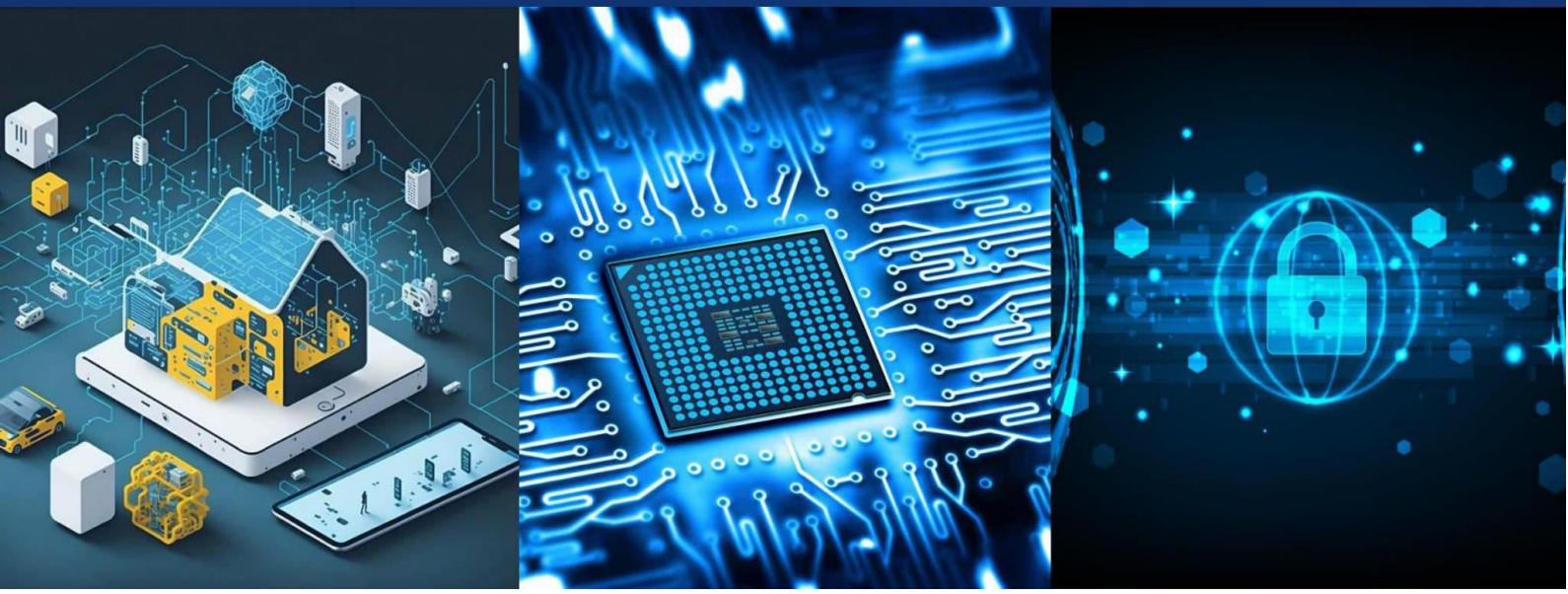
IJIRCCE

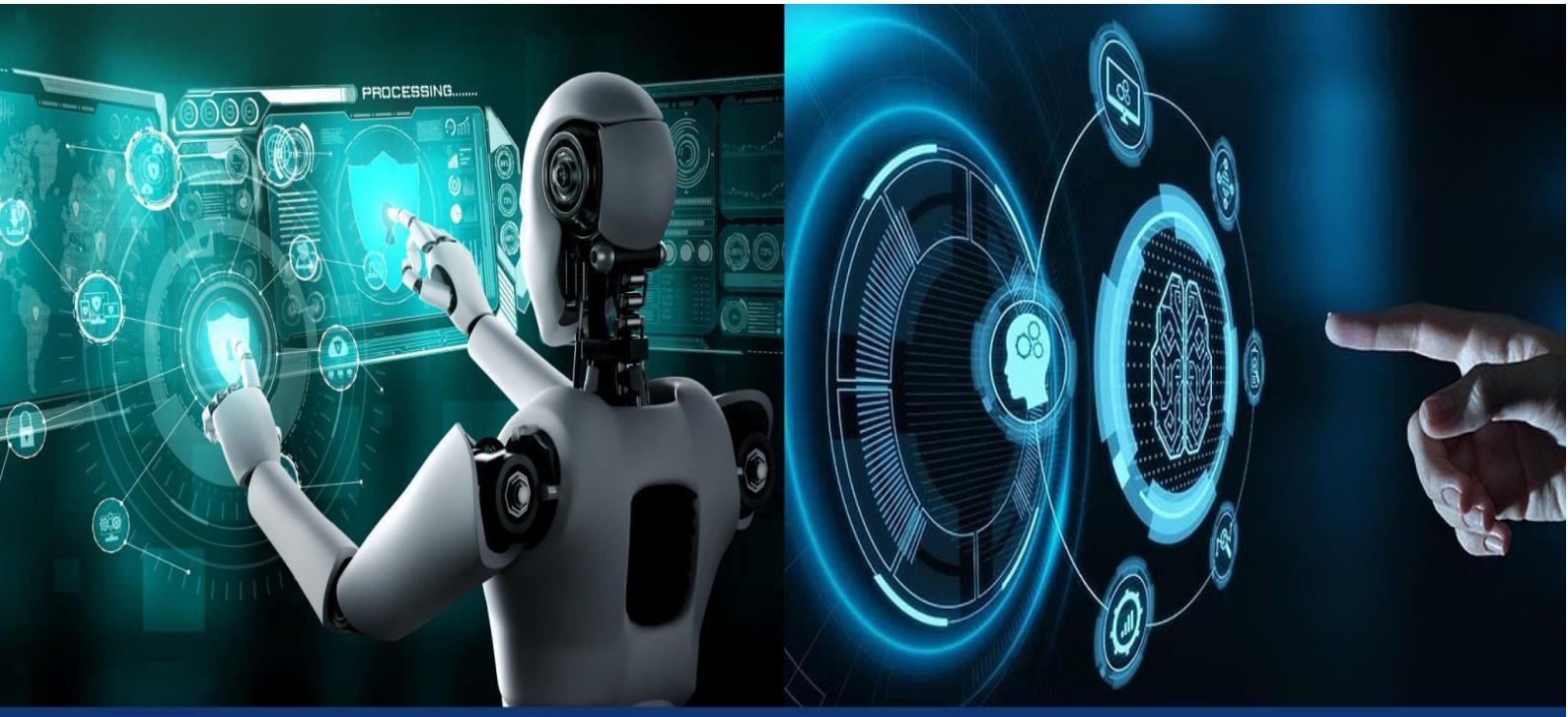
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 14, Issue 2, February 2026

© 2026 IJIRCCE | Volume 14, Issue 2, February 2026| DOI:10.15680/IJIRCCE.2026.1402016

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A Survey on Multimodal Mental State Detection using Artificial Intelligence

Prof. Vaishnavi Sonawane, Masood Madki

Department of Artificial Intelligence & Machine Learning, AISSMS Polytechnic, Pune, India

Diploma Student, Department of Artificial Intelligence & Machine Learning, AISSMS Polytechnic, Pune, India

ABSTRACT: Stress and mental health disorders have become increasingly prevalent, demanding reliable and continuous assessment methods beyond traditional clinical evaluation. Recent advances in artificial intelligence have enabled automated stress detection using behavioral and physiological signals; however, unimodal approaches often suffer from limited robustness and subjectivity. This paper presents a structured survey of AI-based stress and mental state detection systems with an emphasis on multimodal frameworks. Various data modalities, including facial expressions, speech, survey responses, galvanic skin response, heart rate, and motion signals, are reviewed along with commonly used datasets and fusion strategies. Furthermore, a stress-relevant class mapping and variance-weighted fusion space compression strategy is discussed to address the exponential growth in multimodal decision complexity. The proposed approach reduces the outcome space from 35,000 combinations to 45 clinically meaningful stress states while preserving interpretability. The survey highlights current challenges, applications, and future research directions toward scalable, real-time, and ethical multimodal stress monitoring systems.

KEYWORDS: Stress Detection, Multimodal Machine Learning, Physiological Signals, Mental Health, Artificial Intelligence

I. INTRODUCTION

Mental health disorders such as stress, anxiety, and depression have become increasingly prevalent across all age groups and social backgrounds. Prolonged stress not only affects emotional well-being but also leads to physical health issues such as cardiovascular diseases, sleep disorders, and reduced cognitive performance. Despite this growing concern, access to timely and continuous mental health assessment remains limited, especially in resource-constrained environments.

Traditional mental health assessment methods rely heavily on self-reported questionnaires and clinical observation. These methods are subjective, episodic, and incapable of capturing real-time mental states. In recent years, artificial intelligence (AI) and machine learning (ML) techniques have been explored to automate stress and mental state detection using observable indicators such as facial expressions, speech patterns, and textual responses.

However, systems based on a single modality suffer from limited reliability, as individuals can consciously mask emotions in speech or facial expressions. Human stress responses are complex and multidimensional, motivating the shift toward multimodal stress detection, where behavioral cues are combined with physiological signals. Physiological responses such as galvanic skin response (GSR), heart rate (HR), and body motion are governed by the autonomic nervous system and are difficult to manipulate intentionally, making them strong indicators of stress.

This survey reviews existing AI-based stress detection systems with a focus on multimodal approaches, datasets, fusion strategies, and comparative performance. Additionally, it presents a structured compression strategy to reduce multimodal decision complexity while preserving clinical relevance.

II. MOTIVATION AND OBJECTIVES

Most existing studies focus on single modalities or limited combinations of signals, offering incomplete representations of human stress. Moreover, the rapid growth of multimodal research has resulted in fragmented literature with inconsistent evaluation strategies.

© 2026 IJIRCCE | Volume 14, Issue 2, February 2026| DOI:10.15680/IJIRCCE.2026.1402016

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Objectives of this survey are:

- To analyze AI and ML techniques used in stress and mental state detection
- To classify approaches based on modalities and fusion strategies
- To review commonly used datasets and evaluation metrics
- To identify limitations and open research challenges
- To highlight efficient multimodal fusion and compression strategies

III. BACKGROUND AND FUNDAMENTAL CONCEPTS

Stress is a psychological and physiological response triggered when individuals perceive demands beyond their coping capacity. From a biological perspective, stress responses are regulated by the autonomic nervous system (ANS), which controls involuntary processes such as heart rate, skin conductance, and motor activity.

In computational mental health research, stress detection refers to the automated identification or estimation of stress levels using data-driven methods. A modality represents a specific type of data source, such as text, speech, facial images, or physiological sensor data. Systems that utilize more than one modality are known as multimodal systems.

Multimodal information can be integrated using different fusion strategies:

- Early fusion: Feature-level integration before model training
- Late fusion: Decision-level combination of independent models
- Hybrid fusion: Combination of feature-level and decision-level fusion

Physiological signals such as GSR, HR, and motion are particularly valuable because they provide objective indicators of stress that are difficult to consciously control.

IV. MODALITIES AND DATASETS USED

Table 1 summarizes the primary modalities and datasets commonly used in stress and mental state detection systems.

Table 1: Modalities and Datasets Overview

Modality	Dataset	Classes	Description
Facial Emotion	FER2013	7	Facial expressions mapped to stress-related emotions
Voice Emotion	RAVDESS	8	Speech emotion recognition
Survey Response	Custom	5	Likert-scale stress responses
GSR	DEAP / WESAD	Continuous	Skin conductance-based arousal
Heart Rate	MIT-BIH	Continuous	Cardiovascular stress indicators
Motion	MPU6050	Continuous	Activity and tremor measurement

Facial and voice modalities capture external emotional expressions, while physiological signals provide internal stress indicators. Survey-based inputs reflect subjective perception and are often used to complement sensor-based measurements.

© 2026 IJIRCCE | Volume 14, Issue 2, February 2026| DOI:10.15680/IJIRCCE.2026.1402016

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. TAXONOMY OF STRESS DETECTION APPROACHES

5.1 Unimodal Approaches

Early research focused on unimodal systems using text, speech, facial expressions, or physiological signals independently. While unimodal methods are computationally simple, they suffer from limited robustness in real-world conditions due to noise, subjectivity, and intentional emotion masking.

5.2 Multimodal Approaches

Multimodal systems integrate multiple data streams to improve reliability and accuracy. By combining behavioral and physiological indicators, these systems capture complementary aspects of stress. Recent approaches increasingly rely on layered architectures that allow flexible modality integration.

5.3 Fusion Strategies

Early fusion enables joint feature learning but requires precise synchronization. Late fusion offers robustness to missing modalities but limits cross-modal interactions. Hybrid fusion balances both approaches and is widely adopted in modern stress detection systems.

VI. PROPOSED SYSTEM ARCHITECTURE

The proposed multimodal stress detection system follows a layered architecture designed to integrate behavioral, physiological, and digital interaction data in a structured and scalable manner. The architecture consists of four main layers: data acquisition, preprocessing and feature extraction, multimodal fusion and stress modeling, and decision support.

In the data acquisition layer, inputs are collected from multiple sources including facial images, speech signals, survey responses, and physiological sensors such as galvanic skin response, heart rate, and motion sensors. These heterogeneous data streams capture both voluntary behavioral cues and involuntary physiological responses related to stress.

The preprocessing and feature extraction layer performs modality-specific operations such as facial emotion recognition, speech emotion analysis, survey encoding, and signal normalization for physiological data. Extracted features are mapped into stress-relevant representations to enable unified analysis.

In the multimodal fusion layer, behavioral and physiological features are integrated using a variance-weighted fusion strategy. Physiological signals are assigned higher importance due to their stronger correlation with autonomic nervous system activity. Finally, the decision support layer generates a compressed set of clinically meaningful stress states and corresponding stress scores, which can be used for monitoring and intervention.

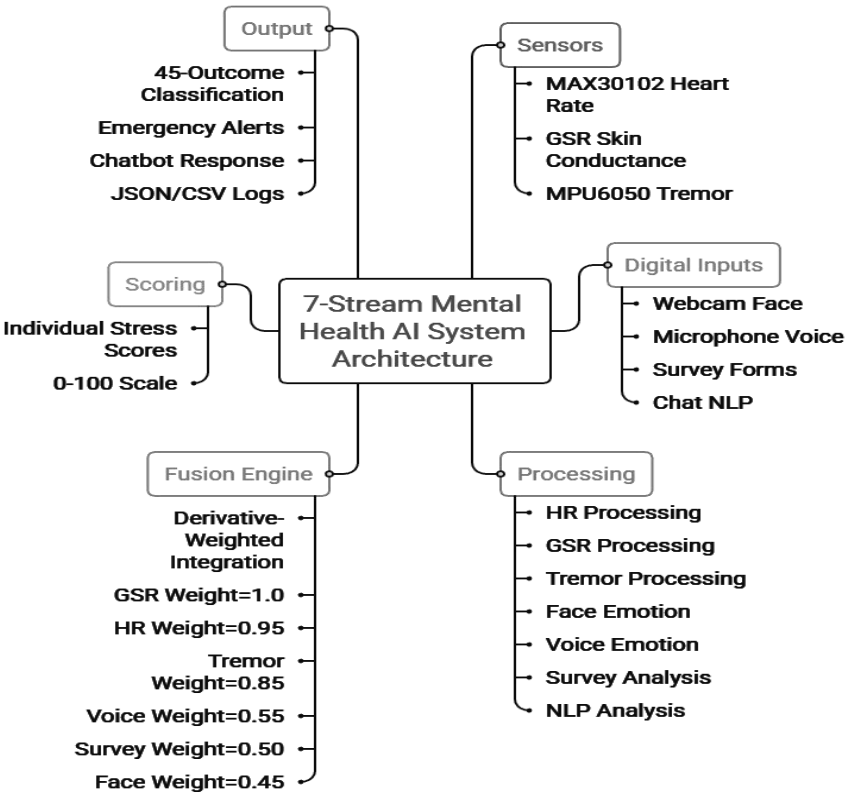
© 2026 IJIRCCE | Volume 14, Issue 2, February 2026| DOI:10.15680/IJIRCCE.2026.1402016

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

7-Stream Mental Health AI System Architecture



VII. MULTIMODAL FUSION SPACE COMPRESSION STRATEGY

Directly combining all original class labels across modalities leads to an impractically large decision space. For example:

7 (Face) × 8 (Voice) × 5 (Survey) × 5 (GSR) × 5 (HR) × 5 (Motion) = 35,000 combinations

7.1 Stress-Relevant Class Mapping

To reduce complexity, original labels are mapped to three stress-relevant classes:

- Behavioral signals: Happy, Neutral, Sad
- Survey responses: Low, Medium, High stress
- Physiological signals: Low, Medium, High activation

This reduces the combination space to:

$3^6 = 729$ combinations

7.2 Variance-Based Weighting

Physiological signals contribute significantly more to stress variance than behavioral signals due to their involuntary nature. Analysis shows approximately:

- 78% variance from physiological signals
- 22% variance from behavioral signals

7.3 Asymmetric Compression

Behavioral modalities are compressed into three presentation states (Calm, Mixed, Distressed), while physiological signals are grouped into fifteen clinically meaningful sensor patterns. This results in:
Final outcome space: $15 \times 3 = 45$ stress states
This compression achieves a 99.87% reduction in complexity while preserving clinical interpretability.

Comparative Analysis
Multimodal systems consistently outperform unimodal approaches across accuracy, robustness, and real-time applicability. Systems integrating physiological signals achieve higher reliability, especially in uncontrolled environments. Decision-level and weighted fusion strategies provide better adaptability to sensor noise and missing data.

VIII. APPLICATIONS, CHALLENGES, AND FUTURE WORK

Applications
- Wearable mental health monitoring
- Workplace stress and burnout detection
- Adaptive human–computer interaction
- Smart healthcare and telemedicine

Challenges
- Limited large-scale multimodal datasets
- Privacy and ethical concerns
- Inter-subject variability
- Sensor noise and synchronization

Future Directions
- Personalized stress modeling
- Continuous stress scoring
- Explainable AI for mental health
- Privacy-preserving learning techniques

IX. CONCLUSION

This survey reviewed AI-based approaches for stress and mental state detection, highlighting the transition from unimodal to multimodal systems. Comparative analysis confirms the superiority of physiological signals in capturing objective stress indicators. A structured fusion space compression strategy was presented, reducing decision complexity from 35,000 to 45 clinically meaningful outcomes. Future research should focus on scalable, ethical, and personalized multimodal systems for real-world mental health monitoring.

REFERENCES

[1] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions," *PLoS ONE*, vol. 13, no. 5, e0196391, 2018.
[2] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, 2015. (FER2013 Dataset)
[3] S. Koelstra *et al.*, "DEAP: A database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.
[4] P. Schmidt *et al.*, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, pp. 400–408, 2018.
[5] G. Giannakakis *et al.*, "Review on psychological stress detection using biosignals," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 440–460, 2022.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details