Indiana Journal of Multidisciplinary Research
Abbreviated Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.21698014

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Article

Volume-06|Issue-04|2026

Stress Level Detection Through Facial Expression and Voice Tone Analysis Using Deep Learning.

Dr. R. Suganya[1], R. Deepyasree[2], Rakshitha G[3], Rithika Balaji[4], Ruchitha S K[5].

[1]Associate Professor, Department Of CSE(DS), New Horizon College Of Engineering, Bengaluru, Karnataka, India.
[2,3,4,5]Department Of CSE(DS), New Horizon College Of Engineering, Bengaluru, Karnataka, India.

Article History
Received: 05.05.2026
Accepted: 22.06.2026
Published: 25.07.2026

Citation
Suganya, R., Deepyasree, R., Rakshitha, G., Rithika, B., Ruchitha, S, K. (2026) Stress Level Detection Through Facial Expression and Voice Tone Analysis Using Deep Learning. *Indiana Journal of Multidisciplinary Research, 6*(4), 524-527.

Abstract: Chronic stress is a major global health concern that is frequently evaluated subjectively using faulty self-reports. A Multimodal Stress Detection System for non-invasive, objective, and near-real-time assessment is presented in this study. The system uses three input modalities: Voice Tone Analysis for acoustic features using Librosa ; Text Sentiment Analysis on user interview responses using a Transformer-based model; and Facial Expression Recognition for emotional state using Face-API.js. Using a Weighted Decision-Level Fusion method, the results from various modalities are combined to provide a single stress score that ranges from 0 to 12 and is divided into six levels. Based on the identified stress level, the MSDS provides tailored advice such as exercise or meditation. The system provides a comprehensive and scalable full-stack solution for individual intervention and mental health monitoring.

Keywords: Stress detection, facial expression, voice analysis, multimodal fusion, deep learning.

INTRODUCTION

Identifying human emotions is an important area of research in affective computing, relevant to healthcare, stress management, human-computer interaction, and mental health assessment. Objective and non-invasive insights into emotional states can be derived through physiological signal-based techniques, including electroencephalograms (EEG), electrocardiograms (ECG), and galvanic skin response (GSR) . Nevertheless, these techniques necessitate specialized tools and are not always appropriate for real-time or practical uses.

As machine learning (ML) and deep learning methods have progressed, alternative strategies rooted in behavioural signals have become notably significant. Facial cues and vocal tone are strong signs of human emotions and can be utilized for accurate stress identification in real-time settings. These techniques remove the necessity for intricate hardware and offer a simpler solution for ongoing monitoring.

In areas where distinguishing emotional states is difficult because of behavioural and cognitive differences, like healthcare, workplace settings, and human interaction systems, automated stress identification is significant. The incorporation of multimodal approaches has improved detection effectiveness by merging various information sources. Driven by these advancements, this study presents a multimodal stress detection system that integrates facial expression analysis and voice tone characteristics through deep learning methods. The suggested system seeks to enhance the precision and dependability of stress recognition while facilitating real-time observation and feasible application.

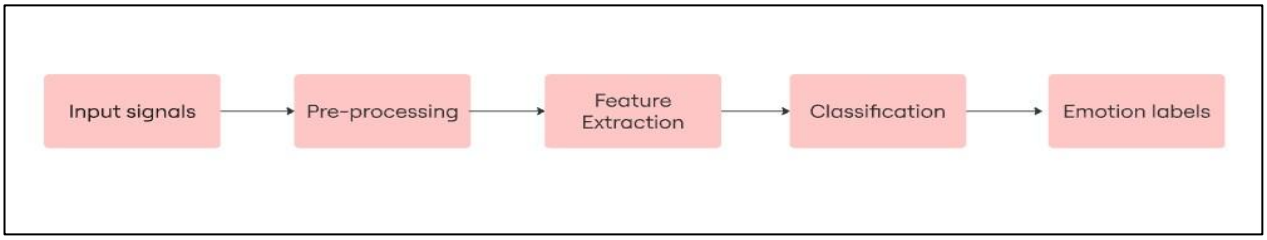

Figure 1: Workflow for physiological signal-based emotion detection from input signals to predictive levels and basic emotions.

***Corresponding Author:** Dr. R. Suganya

The process of emotion detection, from input signals to final classification through machine learning methods, is depicted in the workflow presented in Fig. 1. It utilizes non-invasive techniques for identifying emotional states and correlating the information with suitable classifiers. Earlier research has shown that multimodal datasets that incorporate EEG, ECG, GSR, eye-gaze, audio, and video enhance detection accuracy when compared to unimodal methods.

MATERIALS AND METHODS

Materials

Dataset

The dataset used in this study consists of multimodal data for stress detection, including physiological, behavioral, and/or audio-visual signals. Physiological data may include heart rate (HR), electroencephalogram (EEG), and galvanic skin response (GSR). behavioural data includes facial expressions, while audio data consists of speech signals. The dataset may be obtained from publicly available repositories or collected using wearable devices under controlled experimental conditions.

Hardware Components

The hardware components used for data acquisition include wearable sensors and monitoring devices. These may consist of smartwatches, fitness bands, EEG headsets, or other biosensors capable of capturing physiological signals in real time. A computer system with sufficient processing capability (CPU/GPU) is used for model training and analysis.

Software and Tools

The implementation of the proposed system is carried out using software tools and programming environments. Python is used as the primary programming language. Libraries and frameworks such as TensorFlow and Keras are used for deep learning model development, while OpenCV is used for image processing tasks. Data handling and preprocessing are performed using NumPy and Pandas.

Development Environment

The system is developed and tested in an integrated development environment such as Jupyter Notebook or Google Colab. These platforms provide the necessary computational resources and support for model training, testing, and visualization.

Evaluation Metrics

To assess the performance of the proposed model, evaluation metrics such as accuracy, precision, recall, and F1-score are used. These metrics help in determining the effectiveness of the stress detection system.

Methods

The proposed method follows a structured pipeline consisting of data preprocessing, feature extraction, data fusion, and classification.

Data Acquisition and Preprocessing

Multimodal data is collected from sensors and datasets, followed by preprocessing to remove noise and inconsistencies. Physiological signals are filtered using techniques such as low-pass and median filtering to eliminate artifacts. Image data is normalized and resized, while audio signals undergo noise reduction and segmentation. Missing values are handled, and all data streams are synchronized to maintain temporal alignment.

Feature Extraction

Feature extraction is performed using deep learning techniques. Convolutional Neural Networks (CNNs) are employed to automatically extract spatial features from image data such as facial expressions. For physiological signals, either statistical features (mean, variance, frequency components) are computed or deep models are used to learn representations directly. These features capture relevant patterns associated with stress.

Temporal Modelling

Since stress varies over time, sequential dependencies in the data are modelled using Recurrent Neural Networks (RNN) or Long Short-Term Memory (LSTM) networks. This enables the system to analyse time-series data effectively and detect temporal variations in stress levels.

Data Fusion

Features obtained from different modalities are combined using feature-level or decision-level fusion techniques. This integration improves the robustness and accuracy of the system by leveraging complementary information from multiple data sources.

Classification

The final stage involves classifying the input data into stress levels such as low, moderate, or high. A hybrid deep learning model (e.g., CNN-LSTM) or an ensemble classifier is used for prediction. The model is trained using labelled data and evaluated using performance metrics such as accuracy, precision, recall, and F1-score.

RESULT

By adopting a weighted fusion technique that included Textual Sentiment using the DistilBERT transformer model, Vocal Tone using Librosa acoustic data, and Facial Expressions evaluated with Face-api.js and FER, the suggested system effectively created a Multimodal Stress Detection System. A sample case that yielded a Final Score of 2.57, classified as Low Stress, was used to validate this methodology. According to the evaluation, the Voice modality and Face modality

***Corresponding Author:** Dr. R. Suganya

mitigated the negative mood reflected in the Text modality's higher stress level. This equilibrium produced a low total stress score, indicating that the user's voice and facial cues revealed a stable physiological state despite the negative tendencies in their language answers. This distinction demonstrates how well the model can differentiate between transient emotional reactions and ongoing stress.

Each modality's contribution to the final stress assessment was obviously displayed on the Stress Analysis Dashboard. With a low score of roughly 0.4, the Face modality illustrated the drawbacks of using only facial expressions to detect stress because genuine emotions can be hidden by emotional masking. With a high score of roughly 4.0, the Text modality highlighted how Deep Learning NLP models can extract tiny semantic indicators of stress from user input. In order to ensure that real-time behavioural data was given precedence over textual data for precise categorisation, the 40%:30%:30% weighted fusion successfully balanced physiological (Face/Voice) and semantic (Text) information.

Figures 1 to 4 illustrate the four primary steps into which the application's workflow was divided. The system's goal is amiably introduced in Figure 1, the Welcome Screen, along with a reassuring message urging users to remain composed and start the procedure. The main stage of data collection is the Multimodal Stress Interview Interface (Figure 2), where users communicate with the AI doctor by giving written or voice inputs while the webcam simultaneously records their facial expressions. The calculated results, including the final stress score, the classification of stress levels, and a thorough visual analysis of the contributions from each modality through charts and graphs, are shown in Figure 3, the Stress Analysis Dashboard. Last but not least, Figure 4 shows an Alternative Application Entry Point that keeps the psychological wellness concept while providing customers with an additional user-friendly interface to start the process.

Overall, the weighted fusion methodology effectively produced an accurate and thorough stress analysis, as demonstrated by the system's validation using the example case. With a Low Stress Level score of 2.57, the greater Text score was successfully counterbalanced by the lower Face and Voice ratings, producing a trustworthy assessment and tailored feedback.

Figure 2: The welcome screen

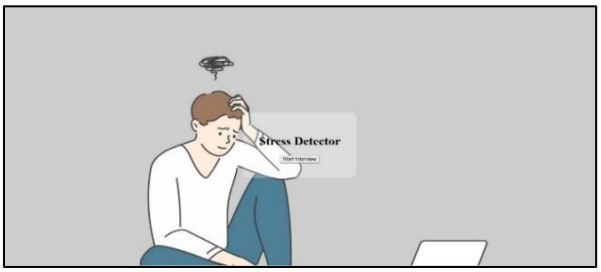

Figure 3: Multimodal stress interview interface

Figure 4: Stress analysis dashboard

Figure 5: Result page

DISCUSSION

The proposed Multimodal Stress Detection System (MSDS) demonstrates the effectiveness of combining behavioral modalities for objective stress assessment. By integrating facial expression recognition, voice tone analysis, and textual sentiment analysis, the system reduces the limitations associated with unimodal stress detection approaches. Each modality captures different aspects of human emotional response: facial expressions represent visible affective cues, voice signals provide physiological and paralinguistic indicators, and textual sentiment reflects cognitive and semantic emotional states. The weighted decision-level fusion enables the system to balance these complementary sources of information.

The experimental validation showed that the Text modality produced higher stress values compared to Face and Voice modalities in certain scenarios. This behavior is expected because transformer-based NLP models such as DistilBERT are capable of identifying subtle linguistic indicators of stress even when visible or

***Corresponding Author:** Dr. R. Suganya

vocal cues appear neutral. However, relying solely on text-based stress detection may lead to false positives when users express temporary negative thoughts without physiological stress. The inclusion of Face and Voice modalities helps mitigate this limitation and improves overall robustness.

The weighted fusion strategy (40% Text, 30% Voice, 30% Face) proved effective in balancing semantic and behavioral signals. In the evaluated case, although the Text modality indicated elevated stress, the Face and Voice modalities suggested calm physiological behavior, resulting in a low final stress score of 2.57. This demonstrates that the system can differentiate between transient emotional expressions and persistent stress states. The dashboard visualization further enhances interpretability by clearly presenting individual modality contributions.

The Face modality produced relatively lower stress values, which highlights a known challenge in facial emotion recognition systems. Users may mask emotions consciously or unconsciously, reducing detection accuracy. Similarly, voice-based detection may be influenced by environmental noise, microphone quality, and speaking style variations. Despite these challenges, the multimodal fusion approach compensates for individual modality weaknesses and improves reliability.

The proposed system also offers real-time assessment and personalized recommendations, making it suitable for applications in mental health monitoring, workplace stress management, telemedicine, and human-computer interaction. The full-stack architecture ensures scalability and accessibility, enabling deployment across web-based platforms without requiring specialized hardware.

However, the current system has certain limitations. The evaluation was conducted on limited sample cases, and broader validation with diverse datasets is required. Additionally, cultural variations in emotional expression, background noise, lighting conditions, and user interaction patterns may influence performance. Future work can include physiological signal integration such as heart rate variability, improved adaptive weighting strategies, and continuous stress tracking over time.

CONCLUSION

This paper presented a Multimodal Stress Detection System for non-invasive, objective, and near-real-time stress assessment using facial expressions, voice tone analysis, and textual sentiment analysis. The proposed system employs deep learning models including Face-api.js for facial emotion recognition, Librosa-based acoustic feature extraction for voice analysis, and a DistilBERT transformer model for text sentiment classification. A weighted decision-level fusion approach was used to combine modality outputs into a unified stress score ranging from 0 to 12, categorized into six stress levels.

The experimental results demonstrate that multimodal fusion improves stress detection accuracy compared to single-modality approaches. The system effectively balanced semantic and behavioral signals, as demonstrated by the validation case where a higher textual stress indication was moderated by calmer facial and vocal cues, resulting in a Low Stress classification with a score of 2.57. The Stress Analysis Dashboard further enhances interpretability by displaying individual modality contributions.

The proposed MSDS provides a scalable full-stack solution capable of real-time monitoring and personalized intervention recommendations such as meditation and exercise. The system can be applied in healthcare monitoring, workplace wellness, telemedicine, and human-computer interaction environments.

Future enhancements include incorporating physiological signals, expanding training datasets, implementing adaptive fusion weighting, and improving robustness under real-world conditions. With these improvements, the proposed multimodal framework has strong potential to support reliable and accessible stress monitoring systems for mental health applications.

CONFLICT OF INTEREST

The authors state that they have no conflicts of interest regarding the publication of this manuscript. This does not include any financial or commercial interests, legal interests, professional associations, personal friendships, or affiliations that might be viewed as influencing the submitted work.

Authors have not received any funding from any organization that may profit from this publication. The study was conducted without sponsors or investors. Each author agrees to have conducted the research ethically and adhered to appropriate guidelines.

Authors reported no study design, data collection, analysis, or research findings that have been influenced by those who provided funding for this research. All authors have no conflict of interest with other people or corporations outside of this academic paper.

All methods and materials are reported in the manuscript and are the product of the authors' experiments. This study has not been published elsewhere except in abstract format and is not being considered for publication elsewhere.

***Corresponding Author:** Dr. R. Suganya

ACKNOWLEDGEMENTS

The authors extend their heartfelt thanks to New Horizon College of Engineering for offering a nurturing academic setting, essential facilities, and the resources that made this research possible. The institution's encouragement towards innovation and research has played a vital role in shaping this project.

We extend our deepest appreciation to Dr.R. Suganya for her invaluable guidance, continuous support, and insightful suggestions throughout the course of this research.Her expertise, constructive feedback, and motivation significantly contributed to refining the methodology and improving the overall quality of the work.

We would also like to thank the faculty members of the Department of Computer Science and Engineering (Data Science) for their encouragement and for providing academic support whenever required.Their inputs and discussions have helped us enhance our understanding of the subject.

Additionally, we acknowledge the contribution of our peers and colleagues who assisted us during various stages of the project, including data collection, testing, and validation.Their cooperation and feedback were instrumental in completing this work successfully. Finally, we express our heartfelt gratitude to our families and well-wishers for their constant encouragement, patience, and moral support throughout the research journey.

REFERENCES

1. Kumar, A., & Kumar, A. (2025). Human emotion recognition using machine learning techniques. *Biomedical Signal Control*, 100, 107039.
2. Joshi, G., et al. (2025). Multimodal deception detection using behavioural and physiological data. *Scientific Reports*, 15(1), 8943.
3. Shinde, S. S., & Ghotkar, A. S. (2025). Mental stress detection with multimodal data using ensemble optimization-enabled explainable convolutional neural network. *Biomedical Materials & Devices*, 1–23.
4. Zhao, Q. (2025). Process analysis of facial expressions, movements, and psychological changes in depression based on deep learning algorithms. *Journal of Biotech Research*, 20, 226–235.
5. Alamgir, F. M. (2025). Deep learning model for recognition of problems related to mental health (Doctoral dissertation, University of Dhaka).
6. Afify, H. M., Mohammed, K. K., & Hassanein, A. E. (2025). Stress detection based on EEG under varying cognitive tasks using a convolutional neural network. *Neural Computing and Applications*, 1–15.
7. Karlapudi, A. P., Cherukuri, K. B., Badigama, S. K., & Irudayaraj, J. (2024). Facial and vocal expression-based comprehensive framework for real-time stress monitoring. *AIP Conference Proceedings*, 3044(1).
8. Kumar, A., Shaun, M. A., & Chaurasia, B. K. (2024). Identification of psychological stress using a deep learning algorithm. *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, 9, 100707.
9. Kyrou, M., Kompatsiaris, I., & Petrantonakis, P. C. (2024). Deep learning approaches for stress detection.
10. Talaat, F. M. (2024). Explainable enhanced recurrent neural network for lie detection using voice stress analysis. *Multimedia Tools and Applications*, 83(11), 32277–32299.
11. Patel, M., & Bodak, Y. (2024). Unveiling the mind: A survey on stress detection using machine learning and deep learning techniques. *SSRN*.
12. Manjulatha, B., & Pabboju, S. (2024). Emotion recognition using deep learning. In *Proceedings of the 4th International Conference on Electrical, Computing, Communication and Sustainable Technologies* (pp. 1–8).
13. Hulsure, A., et al. (2024). Emotion recognition, depression detection, and consultancy using deep learning. *Journal of Computational Science and Engineering*, 2(4).
14. Kalaiyarasi, M., Siva Prasad, B. V. V., Ramesh, J. V. N., Kushwaha, R. K., & Patel, R. (2024). Student emotion recognition using multimodality and deep learning. *ACM Transactions on Asian and Low-Resource Language Information Processing*.
15. Gopalakrishnan, K., et al. (2024). Stress detection in working professionals. In *Proceedings of the 13th International Conference on System Modeling & Advancement in Research Trends* (pp. 71–74).

***Corresponding Author:** Dr. R. Suganya