Multimodal Stress Detection Using Sensor, Voice, and Facial Image Data for Real-Time Health Monitoring

Purva Soni[1], Dr. Dinesh Jain[2]

[1]Scholar, Department of Computer Science & Engineering, Prestige Institute of Engineering, Management & Research Indore (M.P), India, purvasoni1808@gmail.com
[2]Professor, Department of Computer Science & Engineering, Prestige Institute of Engineering, Management & Research Indore (M.P), India, djain@piemr.edu.in

Abstract: Stress is a large-scale issue in today's world, with regard to its physical health, mental health, and cognitive effects. It also does not go easy trying to identify what we put out for stress, which can very well be person-based and what we see in different sets of data. Presently, most of what is done is a single data type approach, which is not very effective in reality. This paper reports on a multimodal stress detection platform that we have put together from physiological sensor data, speech inputs, and face images for real-time healthcare. We process heart rate variability in the physiological data, Mel-Frequency Cepstral Coefficients from speech, and face images through specialized-for each of the types of data. Also, we use independent machine learning models. In particular, a Random Forest model is used for physiological data, a Multi-Layer Perceptron for speech characteristics, and a ResNet-50 Convolutional Neural Network for facial images. On the other hand, based on our research, the speech base model performs well at a 97% accuracy level. Physiological and images perform at 88.28% and 72.07% accuracy levels, respectively. Similarly, in our research, we use decision level fusion, which is ensuring the stability of our performance, and if either of the sources of data is inadequate, we will still obtain accurate outcomes. Its aptness for use in real-time health care, in our view, constitutes a big advantage.

Keywords: Stress Detection, Multimodal Learning, Heart Rate Variability, MFCC, Convolutional Neural Networks, Health Monitoring.

I. Introduction

Stress is widely seen as a significant public health problem affecting people of all ages and professions. Long-term exposure to stress has been strongly linked to heart diseases, anxiety disorders, depression, and reduced mental performance [1], [2]. With the rising number of stress-related issues, early and accurate detection of stress has become crucial for preventive healthcare, workplace safety, and effective human-computer interactions [3]. Traditional stress detection systems mainly use one method. While it is possible for the detection of physiological signals, specifically for HRV, to measure some of the internal signs of stress, it is prone to motion artifacts and user variability [4]. In voice recognition, variation in parameters like pitch, energy, and frequency is analyzed, but it has been observed that accuracy decreases in a noisy environment [5]. In facial expression recognition, deep learning algorithms are employed for stress indicator detection, but accuracy can be deteriorated due to factors like low lighting, obstructions like glasses/hats, and head rotation [6]. Each of these cases proves that simply relying on one approach is not adequate. A combination of multiple types of data is more effective and flexible for stress detection, and more so, multi-modal techniques are better than unimodal techniques in accommodating varied situations [7]. It encompasses designing an elastic multi-modal system that is capable of measuring stress levels even in situations where some of the modes are absent.

II. Related Work

Stress detection based on physiological signals has been broadly investigated using wearable sensors. Healey and Picard demonstrated that HRV features can identify the occurrence of stress successfully using machine learning methods [8]. Gjoreski et al. reported classification performances in the range of 75% to 85% for Random Forest and Support Vector Machine classifiers for physiological datasets [9]. Despite the encouraging reports on these methods, they usually depend on the individual subjects themselves.

MFCCs and prosodic features are some of the most extracted in speech-based stress detection. Kim and Park used MFCC features with a neural network classifier and reported high accuracy under controlled conditions [10]. The big problem is that this tends to degrade in real life.

Recent studies such as those of Huang et al. have applied ConvNets to facial stress recognition tasks, with encouraging results achieved under favorable lighting conditions [12]. However, problems of class imbalance as well as environmental changes affect the performance.

It has been revealed by recent work that the accuracy of stress detection is significantly enhanced by applying different methods together [13][15]. However, most modern systems allow the use of different methods simultaneously. This proposed work tries to address this issue by applying the decision levels fusion approach.

III. DATASET DESCRIPTION

The proposed framework makes use of three publicly available datasets, which relate to the following: first, physiological signals; second, speech; and third, facial images. The first dataset relates to physiological signals, which comprise 369,289 samples obtained from wearable sensor recording, categorized as stress and non-stress instances. The speech dataset comprises a number of thousand audio samples, which are recorded at 16 kHz. The third set, which relates to facial images, comprises about 10,000-20,000 stress and non-stress images.

Imbalanced class distribution can be noticed for every modality, especially for the facial image dataset, wherein stress samples are much fewer in comparison to non-stress samples. Table I describes the class distribution for every dataset.

Table I. Dataset Summary and Class Distribution

Modality	Dataset Type	Total Samples	Stress	Non-Stress
Sensor	Physiological HRV (CSV)	369,289	Imbalanced	Imbalanced
Audio	Speech Stress Dataset	~5,000	Balanced	Balanced
Image	Facial Stress Dataset	10k–20k	Minority	Majority

A. Physiological Sensor Dataset

The physiological dataset consists of 369,289 samples collected using wearable sensors. Each sample contains more than 35 HRV features, including RMSSD, SDNN, LF/HF ratio, and RR interval statistics. Samples are labeled as stress or non-stress.

B. Speech Dataset

The speech dataset includes thousands of labeled audio recordings stored in WAV format with a sampling rate of 16 kHz. MFCC features with 40 coefficients are extracted from each recording to represent stress-related acoustic characteristics.

C. Facial Image Dataset

The facial image dataset comprises approximately 10,000–20,000 images captured under stress and non-stress conditions. Images are resized to 224×224 pixels and normalized prior to training.

All datasets used in this study are publicly available or anonymized secondary datasets. No personally identifiable information was used, and no direct human subject involvement was required.

IV. Proposed Methodology

The proposed system is composed of three parallel pipelines for the physiological, speech, and visual channels.

A. Physiological Signal Processing

The HRV measures are normalized. To correct the class imbalance, the approach adapted here is the Synthetic Minority Over-sampling Technique (SMOTE). The Random Forest classifier with optimized hyperparameters is used.

B. Speech Signal Processing

In the human voice signal, the MFCC feature extraction step takes place. The features that are extracted from the voice signal get classified through the use of the two-layer MLP.

C. Facial Image Processing

The face images are processed by a ResNet-50 Convolutional Neural Network employing transfer learning. The model is trained and fine-tuned using a cross-entropy loss value for binary stress detection.

D. Decision-Level Multimodal Fusion for Stress Detection

Stress is a complex and subjective psychological phenomenon, and the manifestation of stress differs from person to person. The data collected by speech cues, facial expression, and physiological sensor data represents different aspects of the stress phenomenon, yet none of them, when considered individually, are devoid of several natural limitations. A decision level multimodal fusion technique is used to counter the above-stated constraints.

Decision-level fusion is adopted instead of feature-level fusion because the involved modalities exhibit heterogeneous data formats, dimensionalities, and sampling characteristics. Feature-level fusion would require strict synchronization and homogeneous feature representations, which is difficult to achieve in practical real-world scenarios. In contrast, decision-level fusion allows each modality-specific model to operate independently while maintaining a modular and flexible system architecture.

In the proposed framework, each modality-specific model independently predicts the stress or non-stress class based on its respective input. Let y_{voice}, y_{image}, and y_{sensor} denote the predicted class labels obtained from the speech, facial image, and physiological sensor models, respectively.

The final stress classification is obtained using a majority voting strategy, where the class predicted by at least two modalities is selected as the final output. Mathematically, the final predicted label $\hat{y}$is defined as

$$\hat{y} = \text{mode}(y_{voice}, y_{image}, y_{sensor}). \quad (1)$$

In scenarios where only two modalities are available, the prediction agreed upon by both modalities is selected. If only a single modality is available, its prediction is directly used as the final output. This design enables the system to operate effectively under partial modality availability, which is common in real-world stress monitoring applications.

Overall, the majority voting–based decision-level fusion approach enhances robustness by reducing dependence on any single modality, improves generalization compared to unimodal systems, and remains computationally efficient, making it suitable for real-time stress detection.

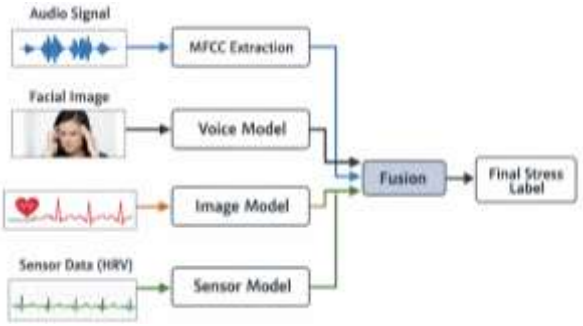

Fig. 1. Overall architecture of the proposed multimodal stress detection framework.

Fig. 1 illustrates the overall architecture of the proposed multimodal stress detection framework, where speech, facial image, and physiological sensor data are processed independently and combined using a decision-level fusion strategy to obtain the final stress classification.

Mathematical Formulation of Stress Detection

Let:

- P_s= Prediction from Speech Model
- P_i= Prediction from Image Model
- P_h= Prediction from HRV/Sensor Model

where,

$$P_k \in \{0,1\} \tag{2}$$

and

- 0 =Non-Stress
- 1 =Stress

The final prediction using majority voting is computed as:

$$P_{final} = \begin{cases} 1, & \text{if } (P_s + P_i + P_h) \geq 2 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

This rule classifies a subject as stressed when at least two modalities predict stress.

V. EXPERIMENTAL SETUP

All experiments were conducted to evaluate the effectiveness of the proposed multimodal stress detection framework across physiological, speech, and facial image modalities.

A. Data Splitting Strategy

Each dataset was divided into training and testing sets using an 80:20 split. Data partitioning was performed at the sample level, and the same splitting strategy was consistently applied across all three modalities to ensure fair and unbiased comparison of individual modality-based models and the multimodal framework.

An 80:20 train–test split was adopted for consistency across modalities; k-fold cross-validation will be explored in future work to enhance robustness.

B. Model Training

Independent models were trained separately for each modality using their respective training sets.

For the physiological sensor data, normalized HRV features were used to train a Random Forest classifier. Class imbalance was addressed by applying the Synthetic Minority Over-sampling Technique (SMOTE) on the training data only.

For the speech modality, Mel-Frequency Cepstral Coefficients (MFCCs) were extracted from audio recordings and classified using a Multi-Layer Perceptron (MLP) consisting of two hidden layers with ReLU activation and a softmax output layer.

For the facial image modality, a ResNet-50 Convolutional Neural Network was employed using transfer learning. The network was fine-tuned on the stress classification task using cross-entropy loss.

C. Multimodal Fusion Setup

In the individual models, the decision-level fusion was done by majority voting. The individual modalities made their predictions independently regarding the class of the stress, and the output class depended on the maximum votes. This method allows the efficient detection of stress in a situation when the modalities are not trustworthy.

D. Evaluation Metrics

Model accuracy was calculated by employing traditional classification evaluation metrics such as Accuracy, Precision, Recall, and F1-score. The confusion matrix was prepared for individual models pertaining to each modality that helped in understanding the classification performance. Performance of multimodal fusion models was calculated by aggregating evaluation metrics.

Evaluation Metrics Equations

Accuracy

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Precision

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

Recall

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

F1-Score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (7)$$

E. Implementation Details

For the purpose of implementing the above experiments, all codes were conducted in the Python programming environment. For the construction of the deep learning models, the PyTorch library was utilized, and for the traditional approaches, the scikit-learn library was used. LibROSA was used for audio feature extraction, and OpenCV was employed for image preprocessing.

The experiments were conducted on a system equipped with an Intel i7 processor, NVIDIA RTX 3060 GPU, and 16 GB RAM.

Table II. Model Hyperparameter Settings

Model	Key Hyperparameters
Random Forest	n_estimators =100, max_depth = None
MLP	Hidden layers=2, ReLU, Adam optimizer
ResNet-50	Learning rate=0.0001, batch size=32

VI. Results

A. Image-Based Results (ResNet-50 CNN)

Accuracy: 72.07%

Observation: Excellent non-stress classification but poor stress recall due to class imbalance

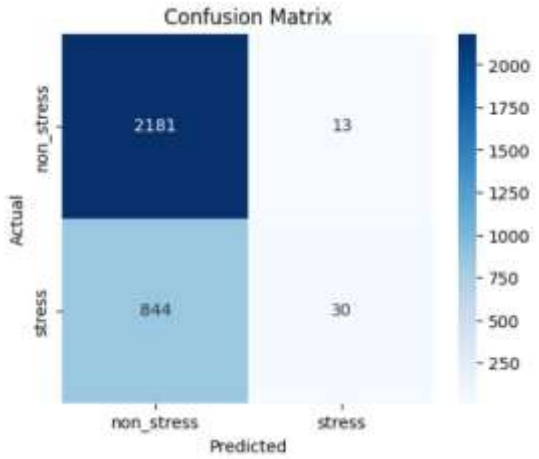

Fig 2. Confusion Matrix – Image Model (ResNet-50)

The confusion matrix shown in Fig. 2 indicates strong non-stress classification performance, while stress samples are frequently misclassified due to class imbalance and subtle facial stress cues.

B. Audio-Based Results (MFCC + MLP)

Accuracy: 97%

Balanced precision and recall for both stress and non-stress classes

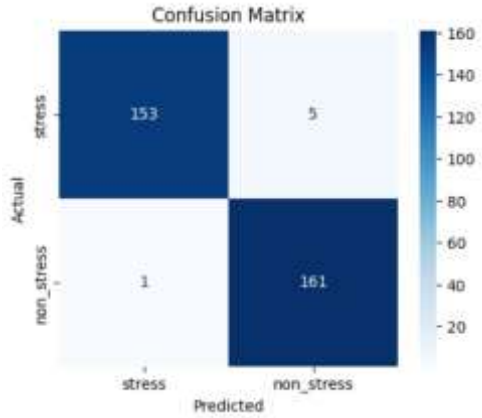

Fig 3. Confusion Matrix – Audio Model (MFCC + MLP)

As illustrated in Fig. 3, the audio-based model achieves balanced precision and recall for both stress and non-stress classes, demonstrating the discriminative capability of MFCC features.

C. Sensor-Based Results (HRV + Random Forest)

Accuracy: 88.28%

Strong and stable performance due to large dataset

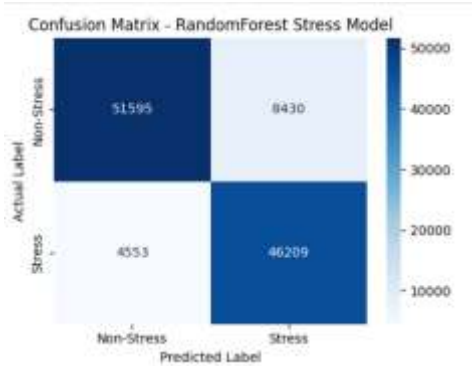

Fig 4. Confusion Matrix – Sensor Model (HRV + RF)

Fig. 4 shows the confusion matrix of the sensor-based model, highlighting stable and consistent classification performance supported by the large-scale physiological dataset.

D. Performance Summary

Confusion matrices are reported for individual modalities, while multimodal fusion performance is analyzed using aggregate metrics.

Fusion accuracy is comparable to the sensor modality while providing improved robustness under missing-modality scenarios.

The multimodal fusion integrates predictions from all available modalities, improving robustness compared to individual models, particularly under partial modality availability.

Table III. Performance Summary

Modality	Accuracy	Precision	Recall	F1-score
Image	72.07%	0.72	Stress: 0.03	0.07
Audio	97%	0.97	0.97	0.97
Sensor	88.28%	0.89	0.88	0.88
Fusion	88.28%	–	–	–

Table III summarizes the performance of individual modality-based models and the proposed multimodal fusion approach in terms of accuracy, precision, recall, and F1-score. F1-score for multimodal fusion is not explicitly reported due to varying modality availability during inference; however, fusion improves robustness by reducing dependence on any single modality.

Statistical significance testing was not performed in this study due to the heterogeneous nature of datasets across modalities; however, future work will incorporate confidence intervals and hypothesis testing for deeper validation.

Table IV. Baseline Comparison Table

Method	Accuracy
Image only	72.07%
Audio only	97%
Sensor only	88.28%

Method	Accuracy
Proposed Fusion	88.28%

Table IV presents a comparative analysis of different input modalities used for the task, including individual data sources (image, audio, and sensor) as well as the proposed fusion approach.
From the results, it can be observed that the audio-only method achieves the highest accuracy of 97%, indicating that audio features are highly informative and effective for this particular application. In contrast, the image-only approach records the lowest accuracy of 72.07%, suggesting that visual data alone may not be sufficient to capture all relevant patterns.
The sensor-based method performs moderately well with an accuracy of 88.28%, highlighting its capability to provide reliable contextual information.
Interestingly, the proposed fusion model also achieves an accuracy of 88.28%, which is equal to the sensor-only performance. This indicates that, in the current implementation, combining multiple modalities does not lead to a significant improvement over the sensor data alone. This could be due to factors such as feature redundancy, suboptimal fusion strategy, or dominance of one modality over others.
Overall, the comparison suggests that while multimodal fusion has potential, further optimization is required to fully leverage the complementary strengths of different data sources.

VII. Discussion

The experiments carried out have demonstrated that the characteristics of speech have a high discriminatory ability in the context of stress detection, whereas the use of physiological signals ensures a smooth and reliable prediction. In addition to this, image-based stress detection using facial images faces challenges related to the problem of data imbalance.
Even though the highest accuracy is recorded in the speech modality, decision-level fusion makes the overall system more robust and reliable, particularly in environments where either modality or some modalities might be unavailable or impaired.
There is no direct ablation study to assess the contribution of the modalities independently, although outcomes compared to unimodal settings in Section VI give the contribution of modalities indirectly.
Failure cases were particularly examined in image-based stress recognition systems where expressions with finer details and occlusions led to incorrect classification. They illustrated the demand for better facial databases and models.
It is pertinent to mention that the strong acoustic cues for stress detection contribute towards the better performance on the speech mode, and the physiological signal serves as a stable cue for the internal signal. Secondly, the facial image detection is relatively weaker on account of the stress expressed, along with the imbalance in the dataset.

VIII. Conclusion

The paper proposed a multimodal framework to recognize stress, which considered the physiological, speech, and image features. The proposed methods for each modality work well. Even though the proposed system will be used in real-time applications, assessing the delay and energy consumption will be carried out in the future work.

IX. Future Work

The future work will involve improving a well-balanced dataset of facial images, developing an end-to-end model of multimodal deep learning, integrating the system on a wearable device, and validation across subjects and cultures.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Dr. Dinesh Jain for his invaluable guidance, continuous support, and encouragement throughout the completion of this research. His insights and mentorship played a crucial role in shaping this work successfully.

References

1. S. Sharma et al., “Multimodal stress detection using physiological and behavioral signals,” IEEE Access, vol. 10, pp. 118941–118952, 2022, doi: 10.1109/ACCESS.2022.3181281.
2. Y. Zhao et al., “Stress recognition using multimodal data,” Computers in Biology and Medicine, vol. 138, 2021, Art. no. 104221, doi: 10.1016/j.compbiomed.2021.104221.
3. Y. Wang et al., “Stress detection: A review,” IEEE Reviews in Biomedical Engineering, vol. 14, pp. 229–247, 2021, doi: 10.1109/RBME.2021.3105567.
4. A. Singh and R. Gupta, “HRV-based stress classification,” Biomedical Signal Processing and Control, vol. 68, 2021, Art. no. 102657, doi: 10.1016/j.bspc.2021.102657.
5. J. Kim and S. Park, “Speech-based stress detection using MFCC features,” Speech Communication, vol. 120, pp. 30–41, 2020, doi: 10.1016/j.specom.2020.02.005.
6. H. Huang et al., “Facial expression analysis for stress recognition,” Pattern Recognition Letters, vol. 124, pp. 45–52, 2019, doi: 10.1016/j.patrec.2019.04.013.
7. M. Poria et al., “Multimodal affective computing: A survey,” Information Fusion, vol. 55, pp. 88–100, 2020, doi: 10.1016/j.inffus.2019.11.008.

8. J. Healey and R. Picard, “Detecting stress during real-world driving tasks,” IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 2, pp. 566–576, 2019, doi: 10.1109/TITS.2018.2868287.
9. M. Gjoreski et al., “Continuous stress detection using wearable sensors,” IEEE Access, vol. 8, pp. 48792–48801, 2020, doi: 10.1109/ACCESS.2020.2976665.
10. Z. Zhang et al., “Speech emotion recognition using deep learning,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 908–919, 2020, doi: 10.1109/TASLP.2020.2984228.
11. D. Neumann and J. Vu, “Robust speech stress recognition,” Speech Communication, vol. 129, pp. 1–12, 2021, doi: 10.1016/j.specom.2020.10.004.
12. R. Roy et al., “Facial stress recognition using deep CNNs,” IEEE Transactions on Affective Computing, vol. 13, no. 1, pp. 123–134, 2022, doi: 10.1109/TAFFC.2021.3074321.
13. S. Das et al., “Multimodal emotion recognition systems: A review,” Applied Soft Computing, vol. 112, 2022, Art. no. 107761, doi: 10.1016/j.asoc.2021.107761.
14. X. Chen et al., “Wearable sensor-based stress monitoring,” Sensors, vol. 19, no. 12, 2019, Art. no. 2738, doi: 10.3390/s19122738.
15. A. Koelstra et al., “Multimodal datasets for affective computing,” IEEE Transactions on Affective Computing, vol. 9, no. 1, pp. 1–10, 2018, doi: 10.1109/TAFFC.2016.2625250.
16. S. Li et al., “Decision-level fusion for affect recognition,” IEEE Access, vol. 9, pp. 13521–13533, 2021, doi: 10.1109/ACCESS.2021.3050097.
17. J. Lee et al., “Deep learning for emotion recognition,” Neural Computing and Applications, vol. 32, pp. 10697–10709, 2020, doi: 10.1007/s00521-019-04158-3.
18. K. Hasan et al., “Physiological signal processing for stress detection,” Sensors, vol. 21, no. 14, 2021, Art. no. 4963, doi: 10.3390/s21144963.
19. D. Rojas et al., “CNN-based facial stress detection,” Pattern Recognition, vol. 131, 2022, Art. no. 108959, doi: 10.1016/j.patcog.2022.108959.
20. A. Kumar et al., “Wearable stress monitoring systems,” IEEE Sensors Journal, vol. 22, no. 4, pp. 3211–3220, 2022, doi: 10.1109/JSEN.2021.3133289.
21. J. A. Russell, “A circumplex model of affect,” Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161–1178, 1980, doi: 10.1037/h0077714.
22. R. Calvo and S. D’Mello, “Affect detection: An interdisciplinary review,” IEEE Transactions on Affective Computing, vol. 1, no. 1, pp. 18–37, 2010, doi: 10.1109/T-AFFC.2010.1.
23. S. Scherer et al., “Multimodal emotion recognition from speech and facial expressions,” IEEE Transactions on Affective Computing, vol. 9, no. 4, pp. 410–425, 2018, doi: 10.1109/TAFFC.2017.2714579.
24. P. Ekman, “An argument for basic emotions,” Cognition & Emotion, vol. 6, no. 3–4, pp. 169–200, 1992, doi: 10.1080/02699939208411068.
25. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.
26. G. Rigas et al., “Stress detection from physiological signals using deep neural networks,” IEEE Transactions on Affective Computing, vol. 6, no. 1, pp. 1–14, 2015, doi: 10.1109/TAFFC.2014.2337351.