

A Multimodal Framework for Real-time Cognitive Load and Stress Detection Using Facial Expressions and GSR Signals

Sakthivel V.¹, Suresh Joseph K.²

¹Assistant Professor, Computer Science and Engineering, Chennai Institute of Technology, Chennai, India.

²Professor, Computer Science and Engineering, Pondicherry University, Pondicherry, India.

E-mail: ¹shakthivel0309@gmail.com, ²ksjoseph.csc@pondiuni.ac.in

Abstract

Stress detection in real-time requires the combination of both physiological and behavioral modalities because the use of a single modality provides limited information on the psychophysiological state of individuals. In this study, a multimodal model is proposed, wherein the physiological modality uses EDA, while the behavioral modality uses facial expression analysis. SCL and SCR features are extracted from GSR signals recorded with a sensing module developed using the ESP32 board. Simultaneously, EAR, blink rate, gaze, and head pose features are extracted from webcam video using the MediaPipe library. The features from the two modalities are time-stamped for alignment and extracted within 30 seconds' window. Then, the features are combined in the feature-level fusion process. The classification of stress levels is done using RF and XGBoost classifiers as low, medium, and high levels of stress. The results of this experiment show that multimodal fusion gives an accuracy of 91.6% compared to features from either one of the modalities.

Keywords: Electrodermal Activity, Facial Behavioural Analysis, MediaPipe Facial Landmarks, Temporal Feature Synchronization, Multimodal Feature Fusion, Psychophysiological Stress Classification, Real-time Stress Assessment

1. Introduction

Psychological stress and cognitive load are the two different dynamic human states, which greatly impact attention, decision-making processes, task execution, and behavior.

measurement. These factors remain crucial for a variety of applications including intelligent education, driver surveillance, health care, and human computer interaction. Traditional physiological methods, especially Electroencephalography (EEG) yield useful information on cognitive and affective states of people but they require special equipment and special conditions of data collection, making them not suitable for continuous measurements in real-life scenarios.

Facial based techniques offer a natural way of estimating the mental state using the facial features. Facial features such as EAR, blink frequency, gaze direction, and head position can provide insights into the attention level, tiredness, and behavioral reaction of the person. The Deep Learning (DL) technique, especially the CNNs, can learn discriminative spatial representation from the facial images. Nevertheless, vision-based approaches still suffer from the problem of varying light conditions, camera quality, facial obstruction, and inter-subject variability. More importantly, facial behavior is not always indicative of the actual physiological response of the subject.

Physiological sensing provides additional data that could enhance the evaluation of the internal human condition. Galvanic Skin Response (GSR) or Electrodermal Activity (EDA) measures changes in skin conductance caused by the sympathetic nervous system and is often used to evaluate physiological arousal and stress. Skin Conductance Level (SCL), Skin Conductance Response (SCR), response magnitude, onset latency, and recovery latency are some of the features used for measuring physiological stress. However, GSR does not provide much context concerning the behavioral reaction observed.

Therefore, combining information from facial behavior with that from GSR physiological measures results in a more holistic view of cognitive and stress states in humans. In such a framework, CNNs will be able to extract spatial features from facial images, while LSTMs will learn temporal features from the physiological data in the form of GSR measures.

Objective

The key goal of the current study is the development of a multimodal system for cognitive load and stress recognition from facial behavioural cues and GSR signals in real time.

Key contributions of the current study are as follows:

- Extraction of facial behaviour cues such as EAR, blink rate, gaze direction, and head pose,

as well as GSR cues relevant to stress state, from synchronized data.

- Timestamp-based synchronization and temporal fusion of complementary cues over 30-second intervals.
- Evaluation of the multimodal representation by means of Random Forest and XGBoost classifiers for stress state recognition.

2. Literature Review

Previous research on affective and stress analysis using physiological signals has been found to be very effective using benchmark data sets and physiological measurements [1], [11], [12]. Analysis of facial behavior has been done extensively using features like facial landmarks, eye movements and expressions to measure affective and behavioral states [3], [8], [9]. Deep Learning techniques have been found very efficient in learning discriminative facial representations automatically, especially Convolutional Neural Networks (CNN) [5], [6], [10].

Physiological signals are an additional source of complementary information which might not be available through the study of facial behavior. Multimodal research works have focused on the fusion of visual and physiological data to enhance the capabilities of affect recognition and stress assessment [2], [4], [13]. LSTM is a good choice for the modeling of time-dependent physiological data [7] while attention-based methods can help with information fusion across multiple modalities [14].

Despite these developments, individual modes continue to be constrained by factors like environmental changes in facial analysis and inadequate behavioral context in physiological analysis. This means that the integration of facial behavior and GSR data is one viable way forward for detecting stress and cognitive load.

3. Proposed Methodology

The proposed system is designed to be a multimodal real-time stress detection system that uses both physiological and facial behavioural data. As can be seen from Figure 1, the proposed system collects GSR data via a GSR sensor which is connected to an ESP32, and at the same time, collects the facial video data through the webcam. Physiological data is used to extract the physiological features based on GSR data, and facial video is analyzed using

MediaPipe facial landmarks to extract the behavioral features like EAR, blink frequency, gaze, and head pose. Features extracted are synchronized via timestamps and collected into 30-second intervals. Synchronized features are combined and fed to the RF and XGBoost classifiers to detect the corresponding stress level.

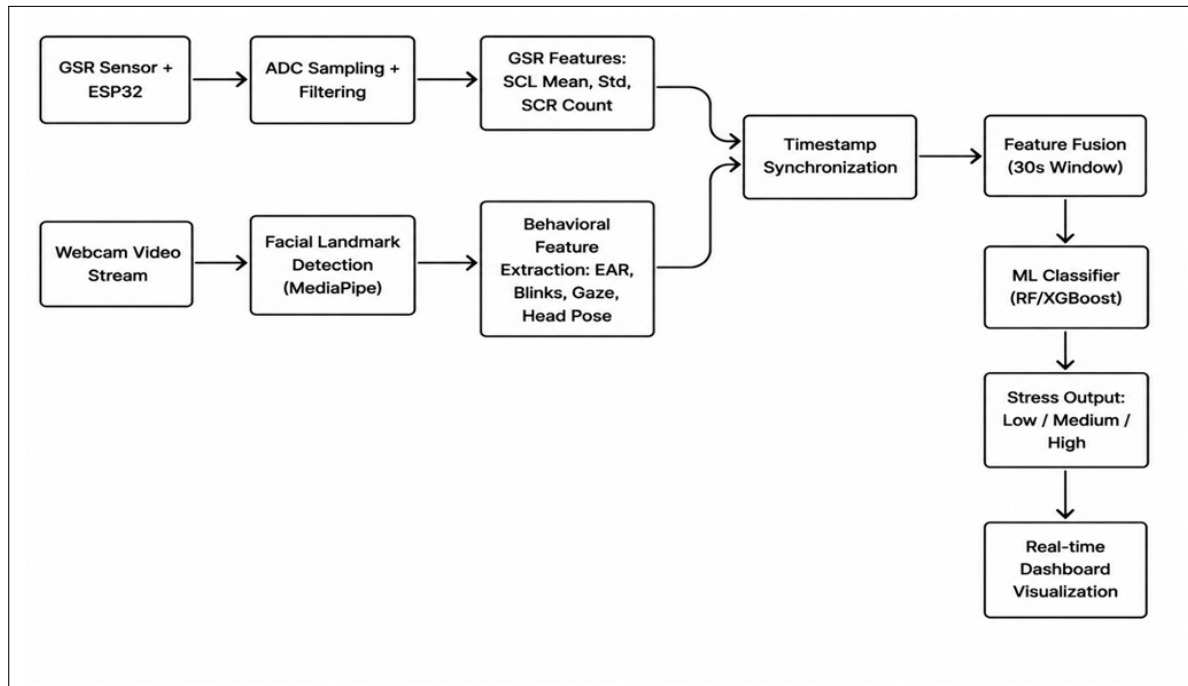


Figure 1. Proposed multimodal GSR and facial-based stress detection framework

3.1 Dataset Description

The proposed framework is assessed through the use of two publicly accessible datasets, namely UBFC-Phys [15] and WESAD [16], which give complementary behavioural and physiological data. UBFC-Phys consists of videos of faces and physiological measurements from 56 participants in rest, speaking, and arithmetic under the stress-induced conditions. The facial video and EDA recordings are especially significant for the proposed facial and GSR-based features extraction [15]. WESAD includes the physiological and motion recordings of 15 subjects, such as EDA, BVP, ECG, EMG, respiration, temperature, and acceleration, in neutral, stress, and amusement states [16]. Thus, WESAD dataset is used as a complementary physiological benchmark, whereas UBFC-Phys helps with multimodal facial-physiological analysis. The main features and experimental subject-level data partitioning of both datasets are shown in Table 1.

Table 1. Dataset characteristics and experimental data split

Dataset	UBFC-Phys [15]	WESAD [16]
Participants	56	15
Modalities Used	Facial Video, EDA, BVP	EDA, BVP, ECG, EMG, Respiration, Temperature, ACC
Relevant Conditions	Rest, Speech, Arithmetic	Neutral, Stress, Amusement
Data Split	70:15:15	70:15:15
Training	39	10
Validation	8	2
Testing	9	3
Training Instances	350	350
Validation Instances	75	75
Testing Instances	75	75

3.2 GSR Signal Acquisition and Pre-processing

GSR signals were measured using the physiological sensor unit connected to the ESP32 and processed in order to decrease noise levels and extract electrodermal traits related to stress. The acquired GSR signal was filtered in order to decrease high frequency noise and transients. The signal was normalized in order to eliminate differences between subjects and increase comparability of electrodermal responses. The pre-processed signal was decomposed into two components, one being the slow-changing SCL and another one being Rapid SCRs. Statistical and response-based features such as mean SCL, SCL standard deviation and SCR count were extracted for every 30 seconds analysis window.

For a window containing N GSR samples S_i , the mean skin conductance level is calculated as:

$$SCL_{\text{mean}} = \frac{1}{N} \sum_{i=1}^N S_i \quad (1)$$

The variability of the signal within the window is quantified by:

$$SCL_{\text{std}} = \sqrt{\frac{1}{N} \sum_{i=1}^N (S_i - SCL_{\text{mean}})^2} \quad (2)$$

The SCR count is estimated by identification of phasic responses in the filtered GSR signal. The GSR features thus extracted were correlated to the respective facial behavioural features by time-stamp synchronization before multimodal fusion.

3.3 Facial Behaviour Feature Extraction

In this research, the web camera recording was analyzed to capture behavioral facial traits related to the changes in stress-related states. Using the MediaPipe Face Mesh algorithm, the facial landmarks were extracted from each video frame, and this algorithm generates geometric points near the eyes and other facial parts. Based on these landmarks, four behavioral features were computed: EAR, blink frequency, gaze direction, and head pose. EAR is a feature that denotes eye opening and is especially helpful in recognizing eye closure and blinking. The blink rate was calculated based on the eye closure events during the analysis period. The gaze features denote visual orientation changes, and the head pose was estimated using facial landmarks geometry to reflect head orientation changes.

For an eye region represented by six landmarks $p_1, \dots, p_6$, EAR was calculated as:

$$EAR = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2\|p_1 - p_4\|} \quad (3)$$

Timestamping of the extracted facial features was done before sending them for synchronization, which involved aligning the features with the respective GSR data.

3.4 Temporal Synchronization and Multimodal Feature Fusion

The extracted features of GSR and facial behaviour were temporally aligned based on their respective timestamps to ensure that features obtained from both streams were taken from the same time frame. Since the two streams have different sample rates, simple merging of the samples was avoided. The synchronized features were then averaged in a 30 second window and concatenated together to create the multimodal feature vector as follows:

$$F_{\text{fusion}} = [F_{\text{GSR}}, F_{\text{facial}}] \quad (4)$$

where F_{GSR} represents the extracted electrodermal features and F_{facial} represents the corresponding facial behavioural features. The proposed approach enables analysis of

physiological arousal data collected by the GSR technique in combination with the behavioral data captured through facial expressions. The synchronized feature vectors were then fed into the RF and XGBoost classifiers for stress level classification. Temporal fusion also diminishes the effect of instant variation.

3.5 Stress Classification

The feature vectors obtained after synchronization of multiple modalities were classified using Random Forest (RF) and XGBoost. These models were chosen due to their ability to capture non-linear relationships between various physiological and behavioral facial features without necessitating complex transformation of features. The RF creates an ensemble of decision trees by selecting random samples and features, while the XGBoost algorithm creates decision trees in sequence with minimization of classification errors through gradient boosting.

These classifiers get the fused 30 second features vectors that include GSR and facial behavior features and classify each observation into one of the three classes of stress levels: 0 or 1. Same subject-level train-validation-test split as outlined in Table 1 was ensured in the model development phase in order to prevent any data leakage. Model performances were evaluated using accuracy, precision, recall, and F1-score metrics and the confusion matrix was used to analyze class-wise behaviors of the predictions.

3.6 Performance Evaluation

RF and XGBoost classifier performance was assessed based on accuracy, precision, recall, and F1-score to measure the overall classification performance and the class-wise performance. Accuracy refers to the overall ratio of samples that were classified correctly, while precision and recall assess the reliability and completeness of the classification predictions, respectively. In addition to that, the confusion matrix was used to assess the distribution of classification errors among the Low, Medium, and High classes. The evaluation was carried out on the hold-out test dataset while retaining the subject-level separation between datasets during their division. The results obtained were utilized for comparing the individual modality performance and the proposed multimodal features.

3.7 Real-time Stress Detection

After feature fusion and classification, the stress level prediction is done from the synchronized 30-second multimodal feature window. In this regard, the system continually captures GSR data using ESP32 and video stream from the webcam and independently processes each stream to update the fused feature vector at each analysis period. After which, the trained RF and XGBoost models classify the observation as Low, Medium or High stress. The stress prediction result is shown through the real-time dashboard interface, thus giving an immediate visualization of the stress status being detected.

3.8 Experimental Setup

The proposed approach to multimodal classification was developed as a multimodal processing pipeline involving facial video analysis with a webcam and the GSR signal analysis. Facial video data was collected using a regular webcam and analyzed using the MediaPipe framework for face landmarks and behavioral features detection. The GSR data was collected using the sensor–ESP32 setup and then analyzed using numerical and signal processing methods to extract SCL- and SCR-based features. The timestamped data from both modalities was segmented into 30-seconds temporal frames followed by the feature normalization and concatenation. The multimodal classification was done using RF and XGBoost as classification models with model training and development using the training and validation partitions and final assessment on the test partition.

Table 2. Experimental configuration

Parameter	Configuration
Facial acquisition	Webcam
Facial processing	MediaPipe
Facial features	EAR, blink rate, gaze, head pose
Physiological acquisition	GSR sensor with ESP32
GSR features	SCL mean, SCL standard deviation, SCR count
Temporal window	30 seconds
Feature synchronization	Timestamp-based
Feature fusion	Feature-level concatenation

Data partitioning	70% training, 15% validation, 15% testing
Evaluation	Accuracy, Precision, Recall, F1-score, Confusion Matrix

Table 2 is presented an overview of the experimental setup used to build the multimodal stress detection system, covering all components from the sensing modality, feature extraction, temporal processing, fusion approach, classifiers and validation setup.

4. Results and Discussion

The evaluation is done by comparing the individual behavioral and GSR feature representations with their multimodal fusion. The analysis will include performance comparison, model-wise evaluation, confusion matrix analysis, and ROC-based evaluation. The obtained results will be used to evaluate the efficiency of physiological and behavioral features combination for the stress classification into three levels.

4.1 Overall Classification Performance

Table 3 summarizes the classification performance of both the individual and multimodal feature representations.

Table 3. Performance comparison of different feature representations

Feature Representation	Accuracy (%)	Precision	Recall	F1-Score
Facial Behaviour Features	82.4	0.81	0.80	0.80
GSR Features	78.9	0.77	0.76	0.76
Multimodal Fusion	91.6	0.92	0.90	0.91

The results indicate that the multimodal feature representation has better classification performance compared to all the other feature configurations considered. The facial behavioral features had an accuracy of 82.4% whereas the GSR features had an accuracy of 78.9%. However, the fused feature configuration had an accuracy of 91.6% with precision, recall, and F1 score being 0.92, 0.90, and 0.91, respectively.

The improved gained through fusion demonstrates the complementary nature of information from both types of features. The behavioral features based on the face include the

changes that take place in the activity of the eyes, gaze direction, blinks, and head orientation, while the GSR features are the physiological changes related to autonomic arousal.

4.2 Training Convergence Analysis

The learning process depicted in Figure 2 is characterized by a fast growth in predictive accuracy at the first epochs and a slower one at the latter epochs. Training accuracy grows from 0.70 at the first epoch to 0.92 at the twentieth epoch, while validation accuracy grows from 0.68 to 0.90. The concurrent improvements in the two curves show that the learned representation keeps on gaining discrimination information as training proceeds instead of getting stuck at an early plateau.

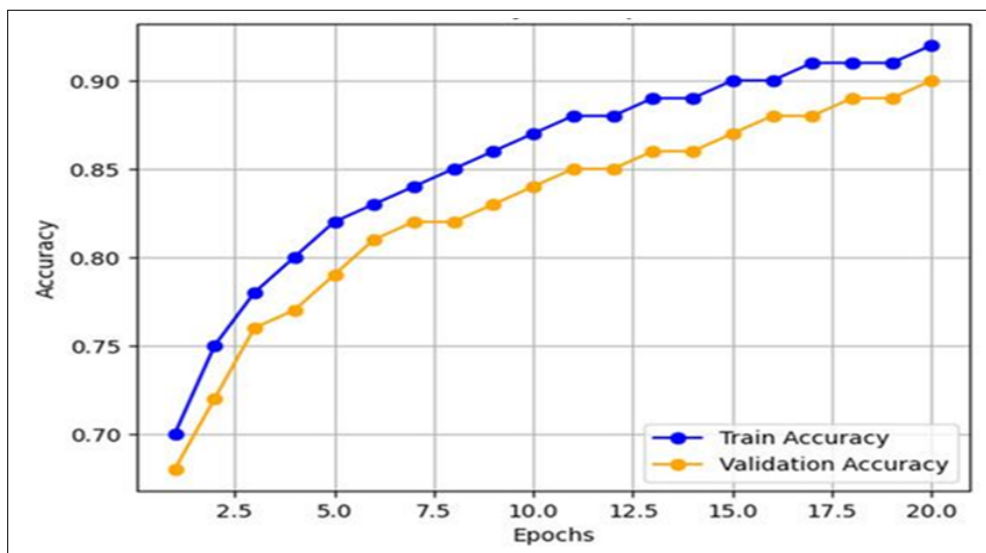


Figure 2. Training and validation accuracy convergence

At the end epoch, the difference between the two trajectories is around 0.02, meaning little discrepancy between optimization and validation performance. The lack of an increasing gap implies that the performance achieved is not affected significantly by overfitting of the training set. The steady leveling off of both the trajectories at about halfway through the training period also proposed optimization convergence.

4.3 Training Accuracy and Loss Analysis

Figure 3 shows an evaluation of both accuracy and loss for model training. The accuracy during training increases very rapidly during the first few epochs and reaches the higher performance region after about five epochs. The validation accuracy is similar in trend and maintains proximity to the training curve during the following epochs. This suggests that

the model learns the main discriminative representation during the earlier optimization phase, while later epochs contribute less to its performance.

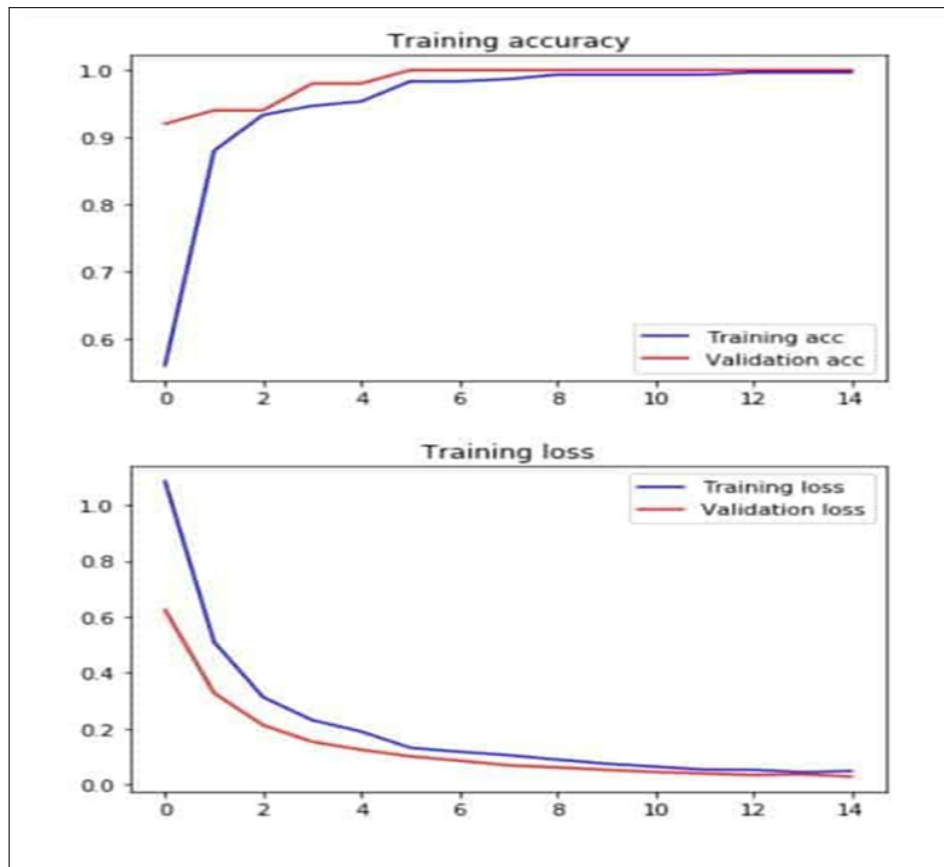


Figure 3. Training and validation accuracy-Loss convergence

The corresponding loss curves show significant drops during the early stages of training, with gradual stabilization of losses on low levels thereafter. The simultaneous drop of losses for both training and validation data indicates proper optimization of parameters, whereas the lack of continuous growth in losses for the validation data implies the absence of clear signs of classical overfitting on the presented training period.

4.4 Confusion Matrix Analysis

The behavior of the class-wise prediction is analyzed through the confusion matrix plots shown in Figure 4 and Figure 5. The multi-class confusion matrix shows that there is concentration of predictions along the diagonal, which develops that the classifier is successful at separating a large portion of the classes being evaluated. The off-diagonal elements represent the class transitions which lead to overlapping decision regions.

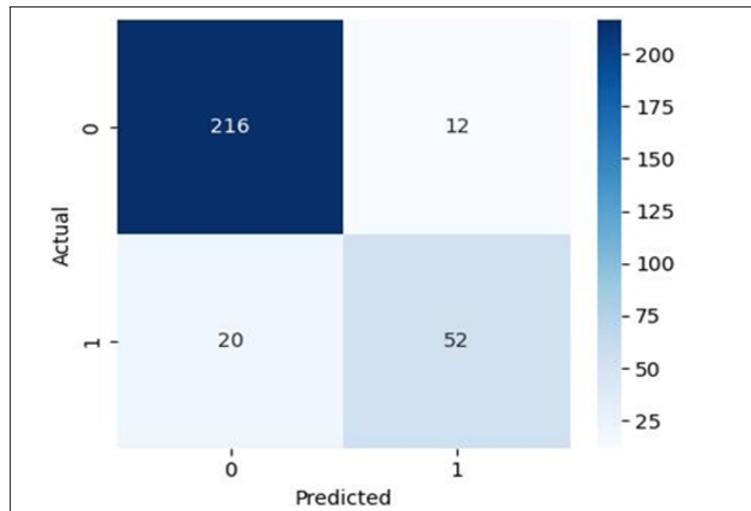


Figure 4. Multi-class classification confusion matrix

The size of the diagonal element with respect to the size of the off-diagonal element shows that there exists a considerable discriminative boundary between the tested classes. The fact that there are some off-diagonal counts also suggests that full separation between classes has not been accomplished. This sort of mistake is especially important when the classes are neighbors or physiologically related states because they have partially overlapping features.

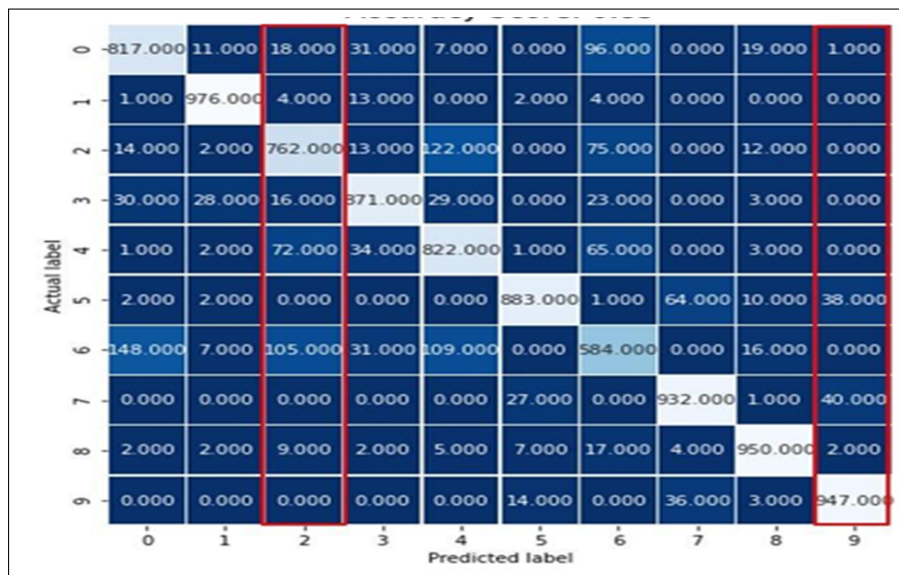


Figure 5. Binary classification confusion matrix

The confusion matrix that is shown in Figure 5 is the one that gives a more precise evaluation of the discrimination of the classes. The matrix has 216 and 52 correct instances along the main diagonal, whereas there are 12 and 20 instances in the other diagonal spots.

Hence, the total number of correct instances is 268 from the 300 instances, which corresponds to 89.3%.

Large diagonal elements clearly separate the two classes that have been evaluated, but the asymmetry of the off-diagonal elements implies that the classification errors are not equally spread out in the two classes. Such asymmetry is a clear indication that there is feature overlap in one of the classes compared to the other, which must be taken into account when evaluating sensitivity.

4.5 Integrated Discussion of Experimental Findings

The combination of the experimental findings suggests that the feature complementarity is the major factor contributing to the multimodal performance enhancement. Unimodal experiments show that facial behavioral and electrodermal features are predictive on their own, but the fusion of these features creates a much more effective representation of the classification. The increase of performance from 78.9-82.4% unimodal setups to 91.6% after the fusion implies that the predictive information of one modality can cover up the drawbacks of another.

Large diagonal elements clearly separate the two classes that have been evaluated, but the asymmetry of the off-diagonal elements implies that the classification errors are not equally spread out in the two classes. Such asymmetry is a clear indication that there is feature overlap in one of the classes compared to the other, which must be taken into account when evaluating sensitivity.

5. Conclusion and Future Work

In this research proposed a multimodal framework for detecting stress in real-time by incorporating information from physiological and facial behavioral modalities. The GSR signals measured using the ESP32-based sensing module were used to extract electrodermal features, whereas EAR, blink rate, gaze and head-pose features were extracted from the webcam video using MediaPipe facial landmarks. Timestamp-based synchronization and temporal fusion of features helped incorporate the complementary information from both modalities into one analysis window. Random Forest and XGBoost were used for the classification of stress state based on multimodal features showing the capability of multimodal feature integration for stress detection. The proposed framework offers a useful basis for stress response measurement in

real-time. The future work of this research includes adaptive subject-specific calibration, larger dataset, advanced temporal modeling, privacy-aware processing and incorporation of IoT and wearable technologies for large-scale implementation.

References

- [1] Koelstra, Sander, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. "Deap: A Database for Emotion Analysis; using Physiological Signals." *IEEE Transactions on Affective Computing*, 2011, vol. 3, no. 1: 18-31.
- [2] Soleymani, Mohammad, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic. "Analysis of EEG Signals and Facial Expressions for Continuous Emotion Detection." *IEEE Transactions on Affective Computing*, 2015, vol. 7, no. 1: 17-28.
- [3] Baltrušaitis, Tadas, Peter Robinson, and Louis-Philippe Morency. "Openface: an Open Source Facial Behavior Analysis Toolkit." In *Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2016, 1-10.
- [4] Kahou, Samira Ebrahimi, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çağlar Gülçehre, Roland Memisevic, Pascal Vincent et al. "Combining Modality Specific Deep Neural Networks for Emotion Recognition in Video." In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, 2013, 543-550.
- [5] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning." *Nature*, 2015, vol. 521, no. 7553: 436-444.
- [6] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." *Communications of the ACM*, 2017, vol. 60, no. 6: 84-90.
- [7] Graves, Alex. "Supervised Sequence Labelling." In *Supervised Sequence Labelling with Recurrent Neural Networks*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, 5-13.
- [8] McDuff, Daniel, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. "Affectiva-mit Facial Expression Dataset (am-fed): Naturalistic and Spontaneous Facial Expressions Collected." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, 881-888.

- [9] Gunes, Hatice, and Björn Schuller. "Categorical and Dimensional Affect Analysis in Continuous Input: Current Trends and Future Directions." *Image and Vision Computing*, 2013, vol. 31, no. 2: 120-136.
- [10] Li, Shan, and Weihong Deng. "Deep Facial Expression Recognition: A Survey." *IEEE Transactions on Affective Computing*, 2020, vol. 13, no. 3: 1195-1215.
- [11] Gjoreski, Martin, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. "Continuous Stress Detection using a Wrist Device: in Laboratory and Real life." In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, 2016, 1185-1193.
- [12] Sharma, Nandita, and Tom Gedeon. "Objective Measures, Sensors and Computational Techniques for Stress Recognition and Classification: A Survey." *Computer Methods and Programs in Biomedicine*, 2012, vol. 108, no. 3: 1287-1301.
- [13] Yin, Zhong, Mengyuan Zhao, Yongxiong Wang, Jingdong Yang, and Jianhua Zhang. "Recognition of Emotions using Multimodal Physiological Signals and an Ensemble Deep Learning Model." *Computer Methods and Programs in Biomedicine*, 2017, vol. 140: 93-110.
- [14] Ashish, Vaswani. "Attention is All You Need." *Advances in Neural Information Processing Systems*, 2017.
- [15] Sabour, Rita Meziati, Yannick Benezeth, Pierre De Oliveira, Julien Chappe, and Fan Yang. "Ubf-Phys: A Multimodal Database for Psychophysiological Studies of Social Stress." *IEEE Transactions on Affective Computing* 2021, vol. 14, no. 1: 622-636.
- [16] WESAD (Wearable Stress and Affect Detection) Dataset: <https://archive.ics.uci.edu/dataset/465/wesad>