Multimodal Fusion for Emotion Classification based on GSR, BVP, and SKT

Original Scientific Paper

Adhitya P. Velip*
Department of Electronics and Telecommunication Engineering,
Goa College of Engineering, Affiliated to Goa University, Ponda, India
a.velip00@gmail.com

Hassanali G. Virani
Department of Electronics and Telecommunication Engineering,
Goa College of Engineering, Affiliated to Goa University, Ponda, India
virani@gec.ac.in

Amita U. Dessai
Department of Electronics and Telecommunication Engineering,
Goa College of Engineering, Affiliated to Goa University, Ponda, India
amitachari@gec.ac.in

*Corresponding author

Abstract *– Emotion recognition in next-generation healthcare is revolutionizing patient care by fostering empathetic and personalized interactions. Advanced systems that analyse facial expressions, speech, and physiological signals enhance our understanding of emotional states, ultimately improving patient experiences. However, interpreting emotions through visible cues can often lead to inaccuracies. This underscores the importance of physiological emotion recognition, which relies on involuntary responses such as heart rate and skin conductance to provide a more accurate assessment of emotion. Such insights are crucial across various fields: in healthcare for better treatment plans, in human-computer interaction for empathetic user experiences, in robotics for more intuitive communications, and in market research for gauging consumer reactions and optimizing strategies. Overall, the integration of physiological emotion recognition has the potential to revolutionize our understanding of and interactions with emotional experiences across numerous disciplines. Galvanic Skin Response (GSR), Blood Volume Pulse (BVP), and Skin Temperature (SKT) are physiological signals used for emotion classification because they are not easily manipulated by individuals. Researchers employ machine learning techniques for emotion classification using GSR, BVP, and SKT; however, more accurate models are needed. This paper proposes a novel multimodal and unimodal emotion classification system that leverages Galvanic Skin Response (GSR), Blood Volume Pulse (BVP), and Skin Temperature (SKT) parameters. The model's performance is evaluated on the PHYMER dataset, which stands for Physiological Dataset for Multimodal Emotion Recognition with Personality as Context. This paper also leverages mutual information feature selection in conjunction with a Random Forest classifier, achieving accuracies of 97% and 96% for arousal and valence, respectively, when analyzing Galvanic Skin Response (GSR). In the case of Blood Volume Pulse (BVP), notable accuracy of 96% for arousal and 90% for valence was reported. For both arousal and valence, SKT achieved 97% accuracy. Additionally, feature-level fusion resulted in arousal and valence accuracies of 97% and 98%, respectively. More compelling are the results from decision-level fusion, which achieved accuracies of 98% for arousal and 99% for valence. These findings not only demonstrate the robustness of our approach but also highlight the great potential for progress in emotion classification.*

Keywords: *Multimodal fusion, GSR, BVP, SKT, Chatterjee's correlation coefficient, Nu-SVM, PHYMER*

Received: November 25, 2025; Received in revised form: February 6, 2026; Accepted: March 28, 2026

1. INTRODUCTION

Human-Computer Interaction (HCI) is an interdisciplinary field focused on designing, evaluating, and implementing interactive computing systems to enhance the relationship between people and technology. The goal is to create user-centered designs that address human needs, taking into account physical, cognitive, and emotional interactions. Affective computing is a multidisciplinary field involving computer science, psychology, engineering, neuroscience, and other re-

lated disciplines. Its primary goal is to develop computer systems capable of understanding, recognizing, and simulating human emotions. [1]. Affective computing involves creating systems that recognize and interpret human emotions using technologies like machine learning and deep learning [2]. By analyzing inputs such as facial expressions, voice, text, and physiological signals, it aims to make human-computer interactions more natural and empathetic. Affective computing systems detect emotions, offering insights into users' emotional states. This technology is utilized in healthcare, education, entertainment, and marketing to boost user engagement. Often called artificial emotional intelligence, it aims to revolutionize healthcare by fostering more intuitive and humane connections between humans and technology [3]. The next generation of healthcare AI for emotion classification uses advanced machine learning and physiological signals to monitor emotional states in clinical settings. This innovative technology can transform patient care by providing insights into emotional well-being, leading to personalized and empathetic treatment approaches and improving patient-physician relationships.

Emotions are complex feelings that influence our thoughts, behaviour, and physical and mental state. It plays a crucial role in social interaction, communication, decision-making, and the conveyance of ideas among individuals. Emotion classification is helpful in healthcare and mental health, customer service, human-computer interaction, robotics, automotive, education, e-learning, and market research. In healthcare and mental health, applications include autism, Parkinson's disease, stress management, personalized treatment, and monitoring mental health. In automotive, it is used for analyzing drivers [4, 5]. Emotion classification also finds applications in Brain Computer Interaction (BCI) by using the brain signals, EEG, to identify emotions using machine learning [6]. The next generation of healthcare AI for emotion classification uses advanced machine learning to interpret emotional states through physiological signals in clinical settings. This understanding leads to more personalized treatment, enhancing care quality and strengthening patient-physician relationships. Clinicians gain from automated administrative tasks, objective progress tracking, and evidence-based feedback. Researchers employed images, music, videos, audiovisual clips, and virtual reality to elicit emotion [7]. Emotional assessment is conducted using self-assessment manikins and questionnaires. Emotion classification uses both observable physical and physiological signals to infer a person's emotional state. Physical signals include facial expressions, gestures, and speech, whereas physiological signals include EEG, GSR, PPG, SKT, and respiration. Physical signals can be readily collected and controlled. Physiological signals are more reliable as they are involuntary responses of the autonomic nervous system. [5, 8] comprehensively reviewed physiological signals, elicitation techniques, datasets, and classification techniques used in emotion recognition. After EEG, the most widely used method in emotion classification using physiological signals, followed by ECG [9]. Galvanic Skin Response (GSR) is a method for assessing the electrical conductance of the skin, which varies with sweat gland activity, which is governed by the sympathetic branch of the autonomic nervous system. GSR and Skin Temperature (SKT) are essential components of the integumentary system, serving as physiological and behavioral indicators of human emotions. These manifestations reflect the sympathetic nervous system's activity during emotional events, with the skin responding to emotional states through autonomic and glandular functions. Blood Volume Pulse (BVP) is a component of the cardiovascular system and is regulated by the autonomic nervous system. The interplay between the cardiovascular and integumentary systems enhances their ability to classify emotions effectively and provide a deeper understanding of emotional states. GSR indicates arousal levels, BVP provides insights into cardiovascular responses, and SKT illustrates vasomotor changes, collectively offering a comprehensive view of emotional reactions. In contrast to electrocardiography (ECG) and electroencephalography (EEG), integrating GSR, BVP, and SKT into a single wristband improves mobility, comfort, and setup ease. ECG and EEG, however, require stationary configurations, increased setup times, and costly equipment, often necessitating controlled environments for accurate data collection.

Emotions are classified using discrete and dimensional models. The Russell circumplex model of emotion classification was employed, which delineates the dimensions of arousal and valence [9]. Arousal and valence are the two dimensions in the circumplex model of emotions. Arousal refers to the intensity of excitation, while valence refers to the pleasantness of an emotion, as shown in Fig. 1. The features extracted for classification when using machine learning techniques are the time domain, frequency domain, and time-frequency domain features [9]. To improve the accuracy of machine learning models, feature selection techniques can be used to identify and retain the most relevant features from the dataset while discarding redundant ones.

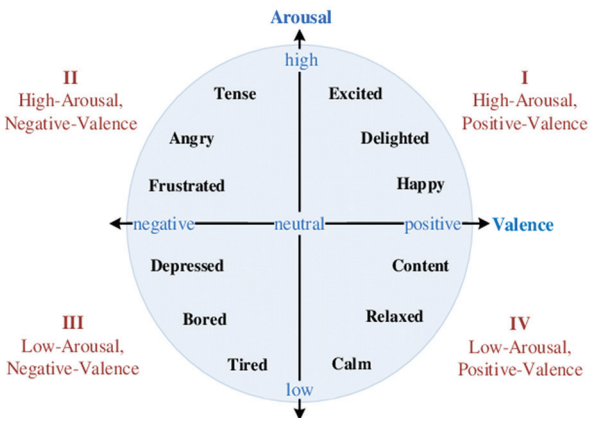


Fig. 1. Valence Arousal model (reprinted from reference [10])

The machine learning techniques used are Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Decision Tree (DT), and Random Forest (RF) [11]. In deep learning, Recurrent Neural Network, Long Short-Term Memory, Transformers, Multilayer Perceptron (MLP), Probabilistic Neural Network (PNN), and Convolutional Neural Networks (CNN) [12] are used for classification [13]. SVM is most commonly used for machine learning emotion classification. Deep learning overcomes the problem of manual feature extraction of machine learning models, but they suffer due to the complexity of the architecture [14].

Our study aims to enhance the performance of machine learning models for emotion classification using Galvanic Skin Response (GSR), Blood Volume Pulse (BVP), and Skin Temperature (SKT) physiological signals. A very few studies have explored these parameters for emotion detection. We plan to achieve this by applying relevant feature selection techniques, employing various machine learning models, and tuning their hyperparameters. Ultimately, our work advances emotion classification by integrating machine learning with physiological signals. Additionally, we aim to improve the model's performance compared to existing models in these fields. The main contributions of the study are as follows:

- This is the first study to use early fusion (feature-level fusion) for emotion classification with GSR, BVP, and SKT signals.
- This study is the first to apply late fusion (decision-level fusion) to GSR, BVP, and SKT signals for emotion classification.
- This study presents an innovative application of Chatterjee's correlation feature selection method for emotion classification.
- We aim to improve the accuracy of emotion classification by using machine learning with GSR, BVP, and SKT signals, compared with existing studies.
- We seek to advance the fields of emotional intelligence, affective computing, and healthcare by developing an emotion classification model that employs GSR, BVP, and SKT signals. The model is intended for integration into innovative healthcare systems for practical applications.

2. LITERATURE REVIEW

This section reviews existing work on emotion classification that uses GSR, BVP, and SKT as unimodal and multimodal signals and applies machine learning techniques. In emotion studies, emotion elicitation is crucial for provoking emotions to facilitate emotion classification. The stimuli used to elicit emotions include images, music, videos, audio-visual clips, and virtual reality. Physiological signals are influenced by the emotions elicited. Compared to speech and facial cues, physiological signals are more involuntary, less prone to deliberate manipulation, and better reflect a person's actual internal state. Regarding cardio-based physiology, the PPG is less commonly used than the ECG [8].

Ab. Aziz *et al*. [15] developed the Asian Affective and Emotional State Dataset, which comprises participants from diverse cultural backgrounds across multiple countries. Researchers employed audiovisual stimuli to collect data and used a range of analytical techniques, incorporating both traditional machine learning and deep learning methods. The classification process was grounded in two fundamental dimensions: binary arousal and valence. The machine learning models implemented in this study comprised Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbors (KNN), and Random Forest (RF). Notably, in the analysis of photoplethysmography (PPG) data, the Random Forest model achieved the highest accuracy for arousal (67.30%), whereas the SVM model attained the highest accuracy for valence (64.94%). Notably, traditional models outperformed deep learning methods in this evaluation. The dataset had the advantage of focusing on the Asian population, which aligns with its intended demographic. However, there were limitations, including suboptimal performance on deep learning tasks and a limited scope of application within unimodal systems; the dataset was not tested for multimodal implementations or for other modalities. Sayed Ismail *et al*. [16] developed an emotion recognition system (ERS) utilizing photoplethysmography (PPG) and electrocardiogram (ECG) signals, assessing its performance within the framework of unimodal ERS. The datasets employed for testing included DREAMER, DEAP, and the Asian affective and emotional state dataset. From these datasets, time- and frequency-domain features were extracted, and various classifiers, including Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbors (KNN), and Decision Tree (DT), were used for classification. The findings indicated that the SVM model achieved the highest accuracy, with 64.61% for arousal and 64.94% for valence using PPG. However, the study acknowledged certain limitations, particularly the absence of a large dataset and the PPG sensor's susceptibility to motion artifacts from hand movement. Gohumpu *et al*. [17] developed a multidimensional emotion recognition model and assessed its performance using various machine learning techniques. This research employed physiological signals, specifically photoplethysmography (PPG), galvanic skin response (GSR), and skin temperature (SKT). The findings indicated that the support vector machine (SVM) model yielded the highest accuracy for GSR signals, achieving 64.4% for arousal and 65.1% for valence. For skin temperature, the decision tree (DT) model achieved the highest accuracy, with 64.7% for arousal and 65.3% for valence. Furthermore, for PPG signals, the SVM again demonstrated superior performance, achieving accuracies of 64.5% for arousal and 65.4% for valence on the DEAP dataset. The study also addressed the challenges posed by a small dataset during model training and validation, as well as the generation of noise from motion artifacts. Hadeel Alharbi *et al*. [18] conducted a thorough investigation into emotion recognition within a virtual reality (VR) environment, employing advanced machine learning and deep learning techniques

using the EEG physiological signal. The machine learning methodologies used in this study included Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), Deep Neural Networks (DNN), DNN with a flattened layer, Bidirectional Long Short-Term Memory (BiLSTM), and Attention Long Short-Term Memory (Attention LSTM). The SVC achieved the highest accuracy of 88% for both binary and multi-class classification tasks. In contrast, the Attention LSTM model achieved an accuracy of 99.99% using deep learning. These results underscore the efficacy of integrating a VR environment for emotion recognition through machine learning and deep learning approaches. However, it is noteworthy that the study did not assess the model's performance across diverse demographics, cultural contexts, and real-world conditions, which may limit the generalizability of the findings.

A. Dessai *et al*. [11] employed the AMIGOS dataset to present an innovative emotion classification approach that integrates ECG and GSR signals through early fusion. The features comprised both time- and frequency-domain data, with mutual information used to guide feature selection. The efficacy of the method was assessed using classifiers including Support Vector Machines (SVM), Decision Trees (DT), k-Nearest Neighbors (KNN), and Random Forests (RF). Among these, the KNN classifier achieved the highest accuracy, 96% for valence and 94% for arousal, when using GSR data. The early fusion of GSR and ECG signals yielded comparable high accuracy values for both valence and arousal classifications.

Zhu *et al*. [19] conducted a comprehensive analysis of frequency-domain, morphological, and pulse rate variability (PRV) features derived from the photoplethysmography (PPG) signal for the classification of arousal and valence. The support vector machine (SVM) classifier demonstrated accuracies of 87.5% for arousal and 81.4% for valence on the experimental dataset. Furthermore, the integration of these features resulted in an overall accuracy of 90%. When the developed model was validated using the DEAP dataset, it achieved accuracies of 73.5% for arousal and 71.5% for valence, indicating its viability for affective state classification. A comprehensive investigation of frequency-domain features and their significant correlation with PPG was conducted. Bamonte *et al*. [20]extracted approximate entropy, Lyapunov exponent, and various Poincaré indices from the nonlinear features of photoplethysmography (PPG) and galvanic skin response (GSR) signals. The researchers employed several classification models, including Random Forest, Support Vector Machine, Gradient-Boosting Machine, K-Nearest Neighbors, Decision Tree, and Convolutional Neural Network with a single-layer perceptron. Among these models, the Random Forest classifier achieved the highest classification accuracy of 76% for both arousal and valence. The findings indicate that established machine learning models can significantly contribute to emotion classification, contingent on careful feature selection. Mattern *et al*. [21] collected physiological data from 40 participants using the Empatica wristband. The data included blood volume pulse, skin temperature, electrodermal activity, and acceleration. Using the 134 segments, accuracy was evaluated using 10-fold cross-validation. The test was conducted using different time-length segmentations, a multiple-segment approach, Relief, principal component analysis, and sequential forward feature selection techniques. The SVM classifier with the multiple-segment approach achieved an accuracy of 75%. The model showed the lowest accuracy when tested using leave-one-subject-out cross-validation (subject-independent).

Kumar *et al*. [22] utilized the AMIGOS and ASCERTAIN datasets, applying KNN, SVM, and LR classifiers for valence and arousal classification using EEG, GSR, and ECG signals. RF improved GSR accuracy by 8.145% compared with other models. Performance was evaluated using accuracy, recall, precision, F1-score, and Matthews Correlation Coefficient (MCC).

Kang *et al*.[23] employed the PPG and GSR physiological signals. The datasets utilized were DEAP and the MERTI-Apps database. The method employed a 1D convolutional autoencoder, a 1D CNN, and an LSTM. On the DEAP dataset, arousal and valence accuracies were 81.33% and 80.25%, respectively, with a window size of 1.1 seconds. The model showed the same result as the MERTI-Apps database. M. Saleem *et al*. [24] proposed a multimodal transformer framework for affective state recognition. This architecture comprises a differential multi-scale feature-extraction module, a global cross-attention encoder, and a difference-augmented weighted-fusion mechanism. The input modalities utilized in this framework include blood volume pulse (BVP), electromyography (EMG), respiration (RESP), electrocardiography (ECG), and galvanic skin response (GSR). In the context of unimodal emotion classification using GSR, the model achieved accuracies of 67.23% for valence and 65.23% for arousal. In contrast, the multimodal classification yielded accuracies of 93.91% for arousal and 94.23% for valence. The fusion strategies implemented in this study encompassed feature-level early fusion, mid-level (intermediate) fusion, and decision-level (late) fusion.

E.M. Polo *et al*. [7] developed and rigorously evaluated an emotion classification model that employed virtual reality scenarios to elicit emotional responses. This model was specifically designed for both discrete and dimensional emotion classification. The physiological signals utilized in this study included electrocardiogram (ECG), blood volume pulse (BVP), galvanic skin response (GSR), and respiration rate. The feature selection process employed a range of methods, including Friedman's test, Recursive Feature Elimination, Ensemble methods, K-Best selection, Minimum Redundancy Maximum Relevance, and square feature selection. Furthermore, various machine learning algorithms were implemented, such as Logistic Regression (LR), Gaussian Naïve Bayes, Decision Tree Classifier, Linear Discriminant Analysis, Support Vector Machine, Random Forest, and XGBoost. Among these approaches, the Logistic Regression mod-

el, combined with the square feature selection method, achieved the highest accuracies, yielding 80% for arousal, 85% for valence, and 70% for four-class emotion classification. The methodology demonstrated a distinct advantage in achieving robust classification that is independent of individual subject variability.

We found that Support Vector Machines (SVMs) are primarily used to classify discrete and dimensional emotions using machine learning. The DEAP dataset is the most commonly used dataset for this purpose. Most models are not evaluated in real-world environments that account for motion artifacts. Another drawback is the use of a combination of GSR, BVP, and SKT non-invasive physiological signals. The novelty of our work lies in developing an emotion classification system that incorporates feature selection techniques and evaluates various machine learning classifiers using unimodal signals from Galvanic Skin Response (GSR), Blood Volume Pulse (BVP), and Skin Temperature (SKT). Additionally, we aim to develop early- and late-fusion models that integrate GSR, BVP, and SKT data. The evaluation of our models is conducted using a range of metrics, including accuracy, F1 Score, and the Matthews correlation coefficient.

3. METHODOLOGY

The GSR, BVP, and SKT signals were obtained from the PHYMER dataset. The valence and arousal levels of emotions were detected using the GSR, BVP, and SKT parameters. The data underwent several steps, including pre-processing, z-score normalization, segmentation, feature extraction, feature selection, and classification. After feature extraction, the mutual information and Chatterjee correlation coefficient feature selection methods were tested, and the data were divided into training and test sets. Five classifiers—SVM, Nu-SVM, KNN, DT, and RF—were employed for the analysis. The classifiers were trained using the selected features and their respective models. To enhance model robustness, grid search and K-Fold cross-validation were employed. The model was evaluated using the test data. The block diagram for emotion classification using early and late fusion of physiological signals is shown in Fig. 2. For early fusion, the features are concatenated, and a subset of features is selected. For late fusion, the classification results for each physiological signal are obtained, and the outputs are combined via majority voting for prediction.

3.1 DATASET

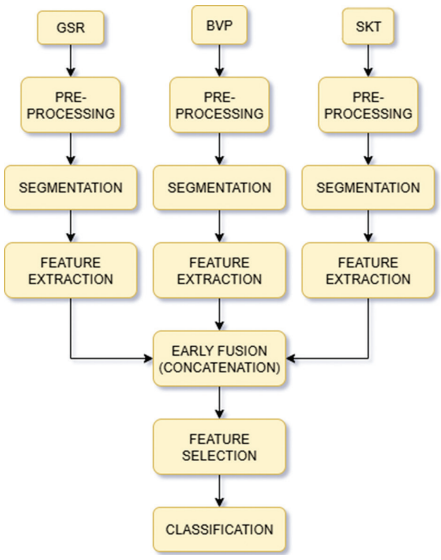
The GSR, PPG, and SKT signals from the PHYMER [25] dataset was utilized for the research. The dataset includes EEG, GSR, BVP, and SKT physiological signals, as well as personality traits. The study involved recordings from 30 participants at a Korean University. There were 15 males and 15 females, with a mean age of 23.56 years. Participants watched 23 videos to elicit emotions. The ratings for each video were based on arousal and valence levels. The binary framework was used as it enhances computational efficiency and provides a more direct correlation to arousal and valence than does the discrete emotion approach. It effectively mitigates the limitations of discrete emotion classification by enabling differentiation between emotions that may share similar valence yet differ in intensity. Additionally, multiclass classification poses greater challenges, particularly due to the prevalence of minority classes within the dataset; in this model, the data is segmented into four distinct quadrants. The selected videos were 2, 4, 5, 9, 10, and 17. Each video represents a different quadrant of Russell's model: Video 2 (HAHV), Video 5 (HALV), Video 17 (HALV), Video 9 (LALV), and Video 4 (LAHV), Video 10 (LAHV). For arousal classification, high-arousal data from videos 2, 5, 17, and low-arousal data from videos 4, 9, 10. For valence classification, high-valence data from videos 2, 4, 10, and low-valence data from videos 5, 9, 17 were used.

3.2 PREPROCESSING

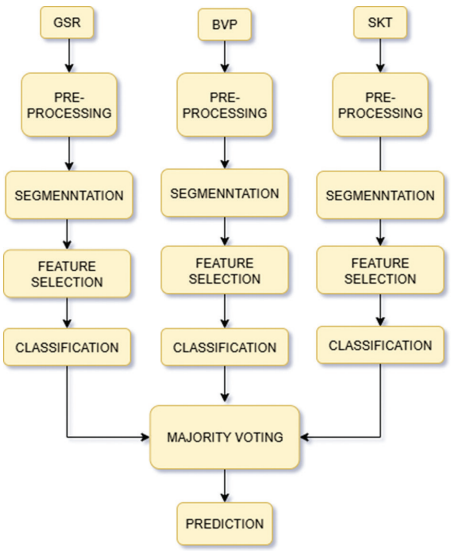
A Butterworth low-pass filter of order 5 with a cutoff frequency of 1 Hz was used to smooth the signal and reduce noise and motion artifacts. For the BVP signal, the Butterworth bandpass filter of order 4, with the lower cutoff frequency set to 1 Hz, and the upper cutoff frequency set to 4 Hz [25, 21].

3.3. SEGMENTATION

In segmentation, the signal is divided into segments of fixed duration (in seconds). In time-based segmentation, the signal is split into equal time intervals. The time-based segmentation uses a 120-point window, which corresponds to a 30-second window for GSR. Methods for PPG segmentation include fixed-window segmentation, in which the signal is divided into segments of a fixed length. Fixed-window segmentation was performed using a 1920-point window (30 seconds) at a sampling rate of 64 Hz. Segmentation of SKT signals was performed using a 120-point window, corresponding to 30 seconds. The extracted video samples were divided by arousal and valence, according to the quadrants of the valence-arousal model.



(a)



(b)

Fig. 2. Block diagram of fusion (**a**) Early fusion; (**b**) Late fusion

The total number of video samples for GSR, BVP, and SKT was 600 per modality. The data distribution for LV/HV and LA/HA is 1.4:1, with the majority class containing 350 samples and the minority class containing 250 samples. It results in a minor or moderate imbalance.

3.4. FEATURE EXTRACTION

To identify the relevant information in the GSR, BVP, and SKT signals, time domain, frequency domain, and cepstral features were extracted from the data using the BioSPPY Python library [26, 27]. The features were selected from each emotion sampled. 69 features from each sample, which include 47 time-domain, 8 frequency-domain, and 14 cepstral features. The time domain features consist of raw signals, first-order differential, and second-order differential.

3.4.1 Time domain features

The time domain features extracted from the BioSPPY library include mean, median, minimum, maximum, maximum amplitude, range, variance, standard deviation, root mean square, kurtosis, skewness, first quartile, second quartile, third quartile, sum of autocorrelation, linear regression slope, linear regression intercept, Pearson correlation, Hjorth parameters (hjorth_mobility, hjorth_complexity, hjorth_chaos, hjorth_hazard), and various first, second, and absolute differences such as mean, median, minimum, maximum, maximum amplitude, range, variance, standard deviation, root mean square, kurtosis, skewness, sum, and their respective means.

3.4.2 Frequency domain features

The frequency-domain features were derived from the Fast Fourier Transform of the emotion samples and were then used to compute the mean, median, minimum, maximum, maximum amplitude, range, variance, and standard deviation.

3.4.3 Cepstral features

Mean, median, minimum, maximum, maximum amplitude, range, variance, standard deviation, root mean square, kurtosis, skewness, first quartile, second quartile, and third quartile of Mel-frequency cepstral coefficients were extracted.

3.5. Z SCORE NORMALIZATION

After feature extraction, z-score normalization was applied. The z-score normalization equation is shown in Eqs 1.

$$z = (x - \mu)/\sigma \tag{1}$$

where:
μ=*mean*
x=*data points*
σ=*standard deviation*

3.6. FEATURE SELECTION

Feature selection is a dimensionality-reduction technique that identifies the most relevant features from the original feature set. It is important for improving the model performance, reducing computational cost, and increasing interpretability. To determine the optimal features from the GSR, BVP, and SKT feature sets for classification, the Chatterjee Correlation Coefficient and Mutual Information were used. Mutual information (MI) measures the relationship between two variables and indicates the extent to which one variable provides information about the other. Assuming the joint probability distribution of two random variables x and y is $p(x, y)$. marginal probability distribution $p(x)$ and $p(y)$ respectively as shown in Eqs 2.

$$I(X,Y) = \iint p(x,y)\, log\, (\frac{p(x,y)}{p(x)p(y)}) \tag{2}$$

When the variables are mutually independent, their mutual information is minimal and equals 0. The higher the mutual information, the stronger the correlation, indicating dependency between the variables. Features were selected based on their high mutual information with the target variables. The features with the smallest mutual information should be eliminated [28]. Chatterjee's correlation coefficient (CCC) is a rank-based measure of functional dependence between two variables, intended to identify any form of dependence, not just linear or monotonic relationships. It ranges from 0 to 1; 1 indicates a strong deterministic relationship, whereas 0 indicates independence or no relationship. It is robust to a non-monotonic relationship. Chatterjee's rank-based correlation coefficient, a nonparametric measure, makes no assumptions about the distributions of the variables, making it versatile. Top features are selected for training. The equation for Chatterjee's correlation coefficient is as shown in Eqn (3).

$$\varepsilon_n = 1 - \frac{3\sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1} \tag{3}$$

where:
r = *rank of y values*
n = *number of samples*

3.7. FUSION

Fusion is the process of combining data from multiple sources or model outputs. The types of fusion techniques used in machine learning applications include early, late, and hybrid fusion. Early fusion is also known as feature-level fusion. It combines data from different sources at the feature level prior to training via concatenation. Late fusion is also known as decision-level fusion. It trains a separate model for each data source, and the predictions are subsequently combined. In our work, we have used early fusion and late fusion-based arousal and valence-based classification [29]. Late fusion analysis is categorized into three methods: majority (hard) voting, averaging, and weighted averaging (soft) voting. In this evaluation, the model employed soft voting by selecting the label based on the average predicted probability. In contrast, hard voting assigns the label through majority-vote predictions. The application of weighted averaging is particularly advantageous when adjustments to specific models suggest parameter variations that could introduce misleading signals, especially in real-time applications. In this scenario, weights are assigned to the model with superior performance, thereby enhancing the overall reliability of the predictions. As one moves toward hybrid fusion, model complexity tends to increase due to the integration of early- and late-fusion methods. This increased complexity, resulting from the overlapping challenges of late and early fusion, has the potential to diminish overall accuracy and elevate the computational costs associated with the model. Hence, the techniques employed in this study were deliberately limited to early and late fusion.

3.8. CLASSIFICATION

The machine learning classifiers were trained on the key features extracted from GSR, BVP, and SKT. Five supervised learning models were used to develop classifiers independently and to compare their performance. Support Vector Machine (SVM) is a supervised machine learning algorithm used for both regression and classification tasks. It works by finding an optimal hyperplane that best separates the data classes, maximizing the distance between the hyperplane and the closest data points, called support vectors. The main objective of SVM is to find the hyperplane with the maximum possible margin, as a large margin usually yields better generalization and a more robust model [30]. Nu-Support Vector Machine (Nu-SVM) is a specialized variant of the Support Vector Machine that regulates the number of support vectors through the 'nu' hyperparameter, balancing the trade-off between training error and margin. Nu-SVM performs data transformations, identifies the optimal hyperplane, and sets the 'nu' parameter to control training error and model complexity, while establishing a lower bound on the number of support vectors. K-Nearest Neighbours (KNN) is a widely used supervised machine learning algorithm for classification and regression tasks. The method begins by defining the parameter k, which represents the number of neighbours. To classify a new data point, the algorithm computes distances between the point and all other points, then sorts them in ascending order to identify the k-nearest neighbours. The new data point is then assigned to the class with the highest frequency among its nearest neighbours, a process known as majority voting. A Decision Tree (*DT*) is a supervised learning algorithm utilized for both regression and classification tasks. It functions by constructing a tree-like structure that mimics human decision-making processes. The process begins with a root node, which serves as the starting point. Optimal features are identified and subsequently split, creating a child node for each selected feature. This process is repeated for each child node until the most suitable stopping criterion is reached. Leaf nodes ultimately signify the final prediction or decision, and in the context of classification, they denote the class label. [5]. The Random Forest (RF) algorithm is an ensemble learning approach, commonly referred to as bagging. Its foundational components consist of decision trees. Each tree is constructed using bootstrap sampling, in which a resampled subset of the data is used. Moreover, for each tree, only a random selection of features is considered for making splits. The growth of each tree continues until it reaches a designated stopping criterion. Ultimately, the algorithm synthesizes the predictions from multiple individual decision trees, employing majority voting to derive the final predictions [31].

3.9. EVALUATION

To evaluate the model's performance, accuracy (A), F1 Score, and Matthew's correlation coefficient (MCC) were computed. The fundamental components of binary classification performance metrics in machine learning are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The accuracy, F1 score, and Matthew's correlation coefficient (MCC) are shown in Eqs. (4)-(8).

Accuracy is defined as the percentage of correctly classified predictions of the total number of predictions.

$$Accuracy=(TP+TN)/(TP+TN+FP+FN) \quad (4)$$

Precision is defined as the percentage of correct predictions that were actually correct.

$$Precision=TP/(TP+FP) \quad (5)$$

Recall (Sensitivity) is defined as the percentage of correct predictions to the sum of the correct predictions and incorrect predictions.

$$Recall=TP/(TP+FN) \quad (6)$$

The F1 score is the harmonic mean of the precision and recall.

$$F1=(2\times(Precision\times Recall))/(Precision+Recall) \quad (7)$$

Matthew's correlation coefficient measures the correlation between the predicted and actual outcomes.

$$MCC=((TP\times TN)-(FP\times FN))/(\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}) \quad (8)$$

4. RESULT

This section delineates the results derived from feature selection techniques and the outcomes of the proposed methodology for classifying emotions based on physiological signals utilizing machine learning classifiers. The objective of this study is to evaluate and compare the efficacy of single versus multimodal signal combinations in the classification of emotions employing Galvanic Skin Response (GSR), Blood Volume Pulse (BVP), and Skin Temperature (SKT). A subject-independent assessment was conducted, in which 80% of the dataset was allocated to training and the remaining 20% to testing. The classification procedure encompasses both training and testing phases. During the training phase, a 10-fold cross-validation was implemented, partitioning the data into 10 folds: 9 for training and 1 for validation. The application of cross-validation in machine learning models is instrumental in obtaining a robust estimate of model performance and stability. All algorithms were executed utilizing the Scikit-learn machine learning library in Python. Hyperparameter optimization was conducted using grid search in the Python library to identify the optimal hyperparameter combination for each model. The accuracy of k-fold cross-validation was computed from the best_score attribute of GridSearchCV, which reflects the average performance of the optimal hyperparameters across all training folds. The analysis was conducted on a 64-bit Windows operating system with 8 GB of RAM and an 11th-generation Intel® Core™ i5-1135G7 processor running at 2.40 GHz.

4.1. CASE 1: EMOTION CLASSIFICATION USING GSR SIGNAL DATA.

The features selected for Galvanic Skin Response (GSR) analysis are depicted in Figs. 3 and 4. Fig 3 illustrates the features identified for arousal classification through mutual information, whereas Fig. 4 details those selected for valence utilizing the Chatterjee correlation coefficient. To enhance the performance of the machine learning models, feature selection based on mutual information and the Chatterjee correlation coefficient yielded twenty features. These features predominantly stemmed from the time domain and excluded any cepstral features.

For GSR arousal, the features chosen through mutual information comprised the following: FFT variance, FFT minimum, FFT median, FFT range, FFT maximum, FFT standard deviation, FFT mean, FFT maximum amplitude, first difference variance, variance, second difference variance, autocorrelation sum, first difference sum, first difference mean, range, linear regression slope, maximum amplitude, standard deviation, first quartile, and second quartile. The features identified using the Chatterjee correlation coefficient were the same, except that the range of first differences was replaced by the linear regression intercept. In terms of GSR valence, the features selected with mutual information included: FFT variance, first difference variance, FFT minimum, variance, second difference variance, FFT median, autocorrelation sum, FFT mean, standard deviation, FFT maximum, FFT range, FFT maximum amplitude, FFT standard deviation, maximum amplitude, absolute difference median, range, third quartile, root mean square, linear regression intercept, and mean. For the Chatterjee correlation coefficient analysis of valence, the first-difference range, linear regression slope, and first-difference mean were replaced by the second quartile, median, and linear regression variance. The mutual information values for both arousal and valence exceeded 0.2, with peaks exceeding 0.5, while the Chatterjee correlation coefficients were above 0.5, reaching a maximum of 0.9. This indicates a strong correlation between the selected features and the target variables, thereby providing pertinent information for the machine learning models.

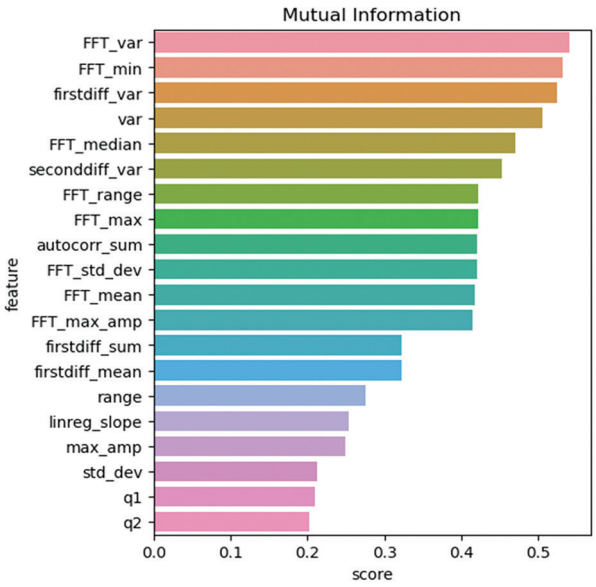


Fig. 3. Performance of Mutual Information feature selection techniques for GSR arousal

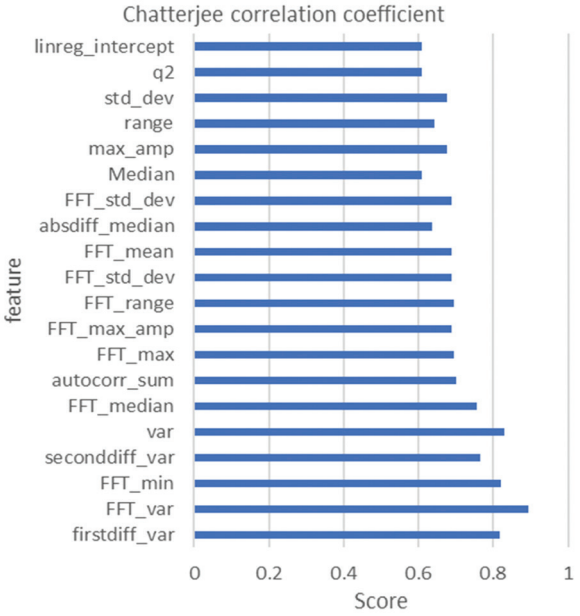


Fig. 4. Performance of Chatterjee correlation coefficient feature selection techniques for GSR valence

Fig. 5 presents the results of K-fold cross-validation accuracy, indicating that accuracy exceeded 90% for both arousal and valence across the feature selection techniques and classifiers employed. The combined findings from K-fold cross-validation with grid search reflect the average performance of the optimal hyper-parameter combinations across the designated K folds. These results suggest that the models exhibited robust performance in classifying both arousal and valence, irrespective of the feature selection method applied. In the context of mutual information feature selection, the K-Nearest Neighbors (KNN) classifier achieved the lowest accuracy of 90% for both arousal and valence classification. Conversely, the Random Forest (RF) classifier achieved the highest accuracy, reaching 97% for arousal and 96% for valence. These findings substantiate the models' efficacy during the training phase.

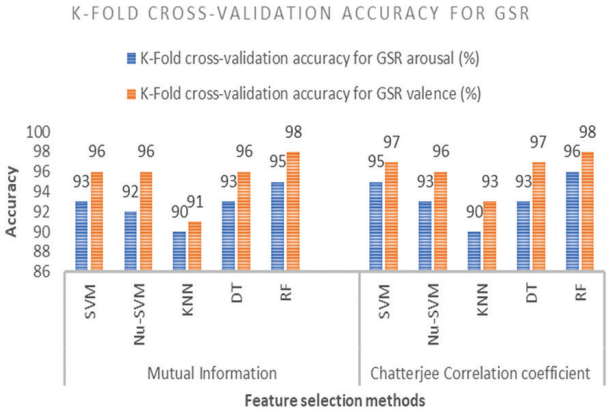


Fig. 5. K-fold cross-validation accuracy for GSR arousal and valence classification

Table 1 and 2 shows the performance of GSR arousal and valence classification, respectively. It reflects the performance evaluation of the fully tuned model for the unseen data. The use of feature selection techniques provided insights into their impact on model performance. For GSR arousal and mutual information, accuracy ranged from 93% to 97%. RF achieved the highest F1 score (0.96) and 0.94 for F1 score, and Mathew's correlation coefficient (0.94), resulting in the highest accuracy (95%). The mutual information accuracy increased by 2%.

Table 1. Performance evaluation for GSR arousal classification

Feature selection	Classifier	F1 Score	MCC	Accuracy (%)
Mutual Information	Nu-SVM	0.89	0.85	93
	SVM	0.90	0.85	93
	KNN	0.91	0.87	94
	DT	0.86	0.79	90
	RF	0.96	0.94	97
Chatterjee Correlation coefficient	Nu-SVM	0.85	0.76	87
	SVM	0.83	0.73	89
	KNN	0.82	0.72	86
	DT	0.89	0.83	92
	RF	0.94	0.91	95

Table 2. Performance evaluation for GSR valence classification

Feature selection	Classifier	F1 Score	MCC	Accuracy (%)
Mutual Information	SVM	0.95	0.91	95
	Nu-SVM	0.93	0.86	93
	KNN	0.88	0.74	86
	DT	0.96	0.93	96
	RF	0.96	0.93	96
Chatterjee Correlation coefficient	SVM	0.96	0.91	95
	Nu-SVM	0.96	0.91	95
	KNN	0.93	0.84	92
	DT	0.97	0.92	96
	RF	0.98	0.96	98

Fig. 6 illustrates the confusion matrix for the SVM, Nu-SVM, KNN, DT, and RF models applied to the test data for GSR arousal and valence, utilizing mutual information and the Chatterjee correlation coefficient.

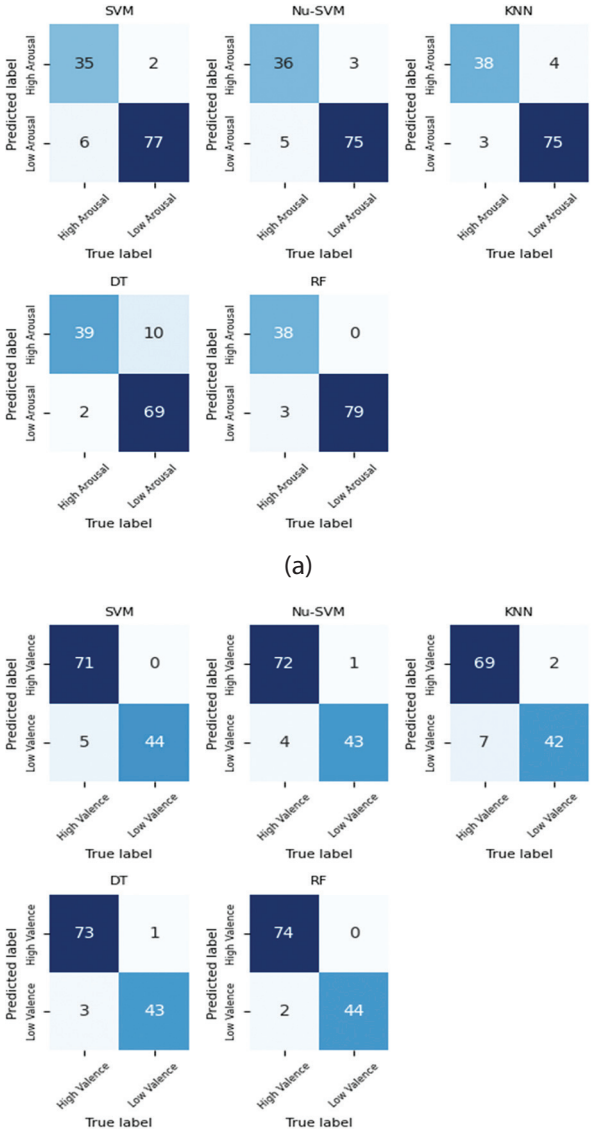


Fig 6. Confusion matrix; (**a**) GSR arousal using Mutual Information; (**b**) GSR valence using Chatterjee correlation coefficient

For GSR valence and the Chatterjee correlation coefficient, the accuracy achieved was 98%, which was higher than that for mutual information. The F1 score and MCC were 0.98 and 0.96, respectively. The random forest classifier achieved exceptional performance in classifying arousal and valence. The k-fold cross-validation and evaluation results on the test data were consistent, indicating that the trained model generalizes well to unseen GSR data for arousal and valence.

4.2. CASE 2: EMOTION CLASSIFICATION USING BVP SIGNAL DATA

A total of 20 features were selected for model training utilizing the Chatterjee correlation coefficient and mutual information.

In the arousal analysis, mutual information identified 11 time-domain features, 8 frequency-domain features, and 1 cepstral feature. The selected features included Hjorth complexity, FFT variance, FFT minimum, Hjorth mobility, FFT median, auto-correlation sum, FFT mean, variance, Hjorth hazard, FFT standard deviation, FFT maximum amplitude, FFT range, FFT maximum, kurtosis, third quartile, MFCC standard deviation, first difference variance, first difference median, second difference variance, and first difference kurtosis. Concurrently, Chatterjee's correlation coefficient identified 3 cepstral features, 8 frequency-domain features, and 10 time-domain features, demonstrating substantial overlap with those identified for arousal. Regarding valence, mutual information selected 12 time-domain features, 7 frequency-domain features, and 1 cepstral feature. These included Hjorth complexity, FFT variance, auto-correlation sum, FFT minimum, second difference variance, Hjorth mobility, FFT median, MFCC maximum amplitude, FFT maximum amplitude, FFT maximum, FFT range, first difference variance, FFT mean, overall variance, maximum amplitude, second difference standard deviation, first difference kurtosis, first difference skewness, minimum, and absolute difference mean. Chatterjee's correlation coefficient, by contrast, identified 3 cepstral features, 4 frequency-domain features, and 13 time-domain features, of which 9 differed from those identified by mutual information. The features uniquely selected through mutual information included FFT maximum amplitude, FFT maximum, FFT range, first difference variance, maximum amplitude, second difference standard deviation, first difference kurtosis, minimum, and absolute difference mean. The features selected using Chatterjee's correlation included first-difference minimum, MFCC range, MFCC maximum, first-difference range, first-difference maximum amplitude, first-difference median, kurtosis, RMS, and standard deviation. The mutual information values exceeded 0.05, reaching a maximum of 0.4, indicating a moderate relationship with the target variable. Similarly, Chatterjee's coefficients exceeded 0.4, with the highest value exceeding 0.7. Fig. 7 illustrates the K-fold cross-validation accuracy for blood volume pulse (BVP), showing that it ranges from 74% to 92% for arousal and from 62% to 91% for valence across both feature selection techniques and all classifiers. Consequently, the model demonstrated robust performance for both arousal and valence across the selected feature extraction methods. In terms of classifier performance for mutual information feature selection, the K-nearest neighbors (KNN) classifier achieved the lowest accuracy, 79% for arousal and 70% for valence. In contrast, the Random Forest (RF) classifier achieved the highest accuracy of 93% for arousal and 90% for valence. When using Chatterjee's correlation coefficient for feature selection, the KNN classifier again achieved the lowest accuracy, with 74% and 62% for arousal and valence, respectively. The RF classifier, however, achieved the highest accuracy of 92% and 91% for arousal and valence, respectively.

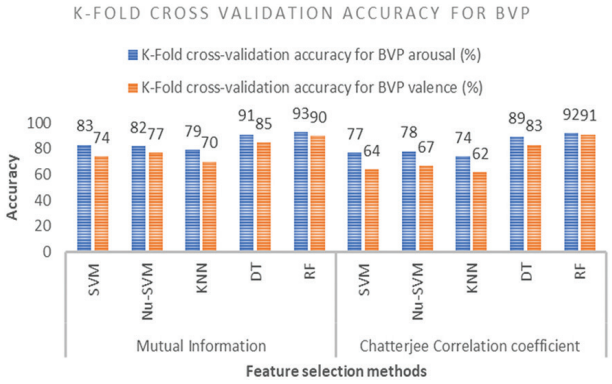


Fig. 7. K-fold cross-validation for BVP arousal and valence classification

Comparing mutual information and Chatterjee correlation coefficient for the best result, the mutual information demonstrated an increase in accuracy of 1% for arousal. In comparison, the Chatterjee correlation coefficient showed an increase of 1% for valence for the RF classifier. The performance of BVP arousal and valence classification is illustrated in Tables 3 and 4. Our analysis revealed an accuracy range of 90%-96% for BVP arousal. The Random Forest (RF) classifier performed best, achieving an F1 score of 0.95 and a Matthews Correlation Coefficient (MCC) of 0.93, both exceeding the Chatterjee correlation coefficient, with an overall accuracy of 96%. For BVP valence, accuracy reached 90%, once again exceeding the Chatterjee correlation coefficient.

Table 3. Performance evaluation for BVP arousal classification

Feature selection	Classifier	F1 Score	MCC	Accuracy (%)
Mutual Information	SVM	0.89	0.81	90
	Nu-SVM	0.91	0.84	92
	KNN	0.85	0.76	88
	DT	0.91	0.86	93
	RF	0.95	0.93	96
Chatterjee Correlation coefficient	SVM	0.69	0.54	79
	Nu-SVM	0.72	0.58	80
	KNN	0.63	0.44	75
	DT	0.90	0.85	93
	RF	0.91	0.87	94

Table 4. Performance evaluation for BVP valence classification

Feature selection	Classifier	F1 Score	MCC	Accuracy (%)
Mutual Information	SVM	0.76	0.45	73
	Nu-SVM	0.76	0.48	74
	KNN	0.71	0.26	65
	DT	0.84	0.61	81
	RF	0.91	0.79	90
Chatterjee Correlation coefficient	SVM	0.73	0.39	70
	Nu-SVM	0.76	0.43	72
	KNN	0.72	0.37	69
	DT	0.82	0.61	83
	RF	0.89	0.72	86

The F1 score and MCC for valence were 0.91 and 0.79, respectively, underscoring the classifier's effectiveness. Overall, the RF classifier showed exceptional ability to differentiate between arousal and valence. Its performance remained consistent through k-fold cross-validation, indicating reliability with unseen data. However, the accuracy for valence was slightly lower, revealing some challenges in generalization. These results highlight the capabilities of advanced classification techniques in understanding emotional states.

4.3. CASE 3: EMOTION CLASSIFICATION USING SKT SIGNAL DATA.

In our analysis, we selected 20 features based on the Chatterjee correlation coefficient and mutual information (MI) for emotion classification. For arousal, the MI includes first difference minimum, second difference minimum, second difference range, first difference maximum, range, first difference range, second difference sum, second difference mean, second difference maximum amplitude, first difference mean, first difference sum, second difference skew, FFT maximum, FFT range, FFT maximum amplitude, first difference maximum amplitude, second difference maximum, FFT standard deviation, RMS, and mean. The Chatterjee correlation selected similar features, with only the second quartile and median replacing the RMS and mean of MI. For valence, the MI consists of first difference minimum, first difference maximum, second difference minimum, second difference range, second difference maximum amplitude, second difference mean, second difference sum, range, first difference range, first difference sum, first difference mean, FFT max, FFT range, FFT standard deviation, FFT maximum amplitude, second difference skew, RMS, mean, Q3, and Q2. The selected Chatterjee correlation features were similar, with only the autocorrelation sum and the maximum of the second difference replacing the second- and third-quartile features.

Most of these features for arousal and valence come from the time domain, with no cepstral features included. Our results show that the mutual information values for the selected features exceed 0.3, with the highest value reaching above 0.6. Similarly, Chatterjee's correlation coefficients exceed 0.6, with a peak exceeding 0.9. These findings demonstrate a strong relationship between the selected features and the target emotional states, improving the relevance and robustness of our machine learning model for emotion classification. Fig. 8 illustrates the K-fold cross-validation accuracy concerning Skin Temperature (SKT). The findings reveal that the model performed effectively in classifying both arousal and valence, employing two distinct feature-selection methods. When using mutual information for feature selection, the Random Forest (RF) classifier achieved an accuracy of 97% for arousal and 99% for valence. Conversely, applying Chatterjee's correlation coefficient for feature selection yielded 98% accuracy for both arousal and valence. A comparative analysis of the two methods indicates that mutual information yielded a 1% increase in valence accuracy, whereas the Chatterjee correlation coefficient yielded a 1% increase in arousal accuracy.

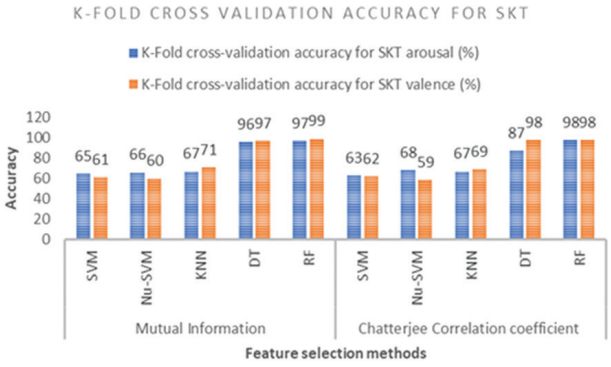


Fig. 8. K-fold cross-validation for SKT arousal and valence classification

Table 5. Performance evaluation for SKT arousal classification

Feature selection	Classifier	F1 Score	MCC	Accuracy (%)
Mutual Information	SVM	0.64	0.31	65
	Nu-SVM	0.64	0.31	65
	KNN	0.68	0.41	70
	DT	0.95	0.90	95
	RF	0.97	0.95	97
Chatterjee Correlation coefficient	SVM	0.54	0.14	57
	Nu-SVM	0.65	0.39	70
	KNN	0.66	0.35	67
	DT	0.87	0.79	89
	RF	0.96	0.93	96

Table 6. Performance evaluation for SKT valence classification

Feature selection	Classifier	F1 Score	MCC	Accuracy (%)
Mutual Information	SVM	0.64	0.31	65
	Nu-SVM	0.64	0.31	65
	KNN	0.68	0.41	70
	DT	0.95	0.90	95
	RF	0.97	0.95	97
Chatterjee Correlation coefficient	SVM	0.54	0.14	57
	Nu-SVM	0.65	0.39	70
	KNN	0.66	0.35	67
	DT	0.87	0.79	89
	RF	0.96	0.93	96

Further details regarding SKT classification for arousal and valence are available in Tables 5 and 6. For arousal, the accuracy achieved using mutual information ranged from 65% to 97%, with the RF classifier achieving an F1 score of 0.97 and a Matthews correlation coefficient (MCC) of 0.95, thereby exceeding the Chatterjee correlation coefficient, which achieved an accuracy of 96%. In the context of valence classification, SKT achieved 97% accuracy with both methodologies, with corresponding F1 scores and MCCs of 0.97 and 0.95. The Random Forest classifier demonstrated exceptional efficacy in classifying both arousal and valence. The results from K-fold cross-validation and evaluations on the test data corroborate the model's strong performance on unseen data for arousal and valence in SKT.

4.4. CASE 4: EMOTION CLASSIFICATION USING FUSION OF GSR, BVP, AND SKT.

The Random Forest (RF) classifier achieved higher accuracy than other classification models in unimodal classification. Therefore, it was selected as the classifier for early fusion and late fusion of Galvanic Skin Response (GSR), Blood Volume Pulse (BVP), and Skin Temperature (SKT).

4.4.1. Early fusion

Tables 7 and 8 show the performance of early fusion for arousal and valence classification, respectively. For arousal and mutual information, the Random Forest (RF) classifier achieved an F1 score of 0.96 and a Matthews correlation coefficient (MCC) of 0.94, outperforming the Chatterjee correlation coefficient, which yielded an accuracy of 96%. For valence and mutual information, accuracy reached 98%, which was 3% higher than that of the Chatterjee correlation coefficient. The F1 score and MCC were 0.98 and 0.96, respectively. Overall, the Random Forest classifier demonstrated exceptional performance in classifying both arousal and valence. The results indicate that the model trained with early fusion of physiological signals for arousal and valence generalizes effectively to unseen data.

Table 7. Performance evaluation for early fusion arousal classification

Feature selection	F1 Score	MCC	Accuracy (%)
Mutual Information	0.96	0.94	97
Chatterjee Correlation coefficient	0.95	0.93	96

Table 8. Performance evaluation for early fusion valence classification

Feature selection	F1 Score	MCC	Accuracy (%)
Mutual Information	0.96	0.94	97
Chatterjee Correlation coefficient	0.95	0.93	96

4.4.2. Late Fusion

Tables 9 and 10 present the outcomes of late fusion techniques applied to arousal and valence classification, respectively.

The predictions generated by multiple classifiers were synthesized through the method of majority voting to produce the final output. The overall result was derived by aggregating the class labels assigned by each classifier. In the context of arousal classification, the model attained an accuracy of 98%, accompanied by an F1 score of 0.98 and a Matthews correlation coefficient (MCC) of 0.96. These results represent the highest performance among the Chatterjee correlation coefficients, which achieved an accuracy of 97%. Additionally, the accuracy associated with mutual information improved by 1%. Regarding valence classification, the mutual information technique achieved an accuracy of 99%, which is 3% higher than that of the Chatterjee correlation coefficient. The F1 score and MCC for valence classification were reported as 0.99 and 0.98, respectively.

Table 9. Performance evaluation for late fusion arousal classification

Feature Selection	Majority voting	F1 Score	MCC	Accuracy (%)
Mutual Information	hard	0.98	0.96	98
	soft	0.97	0.95	97
Chatterjee Correlation coefficient	hard	0.96	0.94	97
	soft	0.95	0.93	96

Table 10. Performance evaluation for late fusion valence classification

Feature Selection	Majority voting	F1 Score	MCC	Accuracy (%)
Mutual Information	hard	0.99	0.98	99
	soft	0.99	0.98	99
Chatterjee Correlation coefficient	hard	0.96	0.93	96
	soft	0.96	0.93	96

During the feature selection process, utilizing mutual information and Chatterjee's correlation coefficient, it was observed that the majority of the features selected were derived from the time domain. Mutual information has proven to be an effective method, as it is capable of identifying both linear and non-linear relationships. This technique primarily assesses the extent to which uncertainty regarding one feature is diminished by considering another feature, making it particularly valuable for feature selection within complex datasets. The approach facilitates the identification of the most informative features, the elimination of non-informative ones, and the reduction of redundancy. Furthermore, Chatterjee's correlation coefficient exhibits superior performance by accommodating various forms of dependence, including monotonic, non-monotonic, and linear relationships. Its robustness against outliers and noise further enhances its utility. The findings highlight the significant impact of feature selection models, demonstrating that diverse classification algorithms can effectively address complex classification challenges. The performance of each model exhibited considerable variation depending on the features selected through these feature selec-

tion techniques, underscoring the critical importance of selecting appropriate features that contain essential information for classification. This facilitates a more efficient and resource-conserving model training process, ultimately achieving not only high accuracy but also optimal performance in emotion classification. The integration of signals also yielded an increase in accuracy, from a maximum of 97% using skin temperature (SKT) alone to 99% through multimodal late fusion with blood volume pulse (BVP), galvanic skin response (GSR), and SKT. An ensemble model, such as a Random Forest, outperformed other methods, indicating its ability to handle complex, high-dimensional data. The success of amalgamating multiple physiological signals contributes to enhanced emotion classification accuracy, resulting in a more reliable classification system.

4.5. STATISTICAL SIGNIFICANCE TEST

A McNemar's test was conducted to assess whether the error proportions differed significantly between two classifiers. The comparison between the decision tree and random forest classifiers revealed that both algorithms yielded identical error rates, failing to reject the null hypothesis (H0). Consequently, this outcome indicates that the classifiers are equally effective in terms of overall error rates, with any minor differences deemed statistically insignificant. In subsequent analyses involving the random forest classifier against K-nearest neighbors (KNN), support vector machine (SVM), and NU-SVM, the random forest exhibited a significantly superior performance relative to the other models.

4.6. COMPUTATION TIME

The computation time for each classifier was meticulously recorded. The training process exhibits considerable variability, as it entails optimizing various combinations of hyperparameters and performs an exhaustive search across parameter combinations and cross-validation folds to ascertain the most effective model due to the use of GridSearchCV. Inference time is defined as the duration required for a trained model to generate predictions on unseen data. Inference time is typically measured in milliseconds.

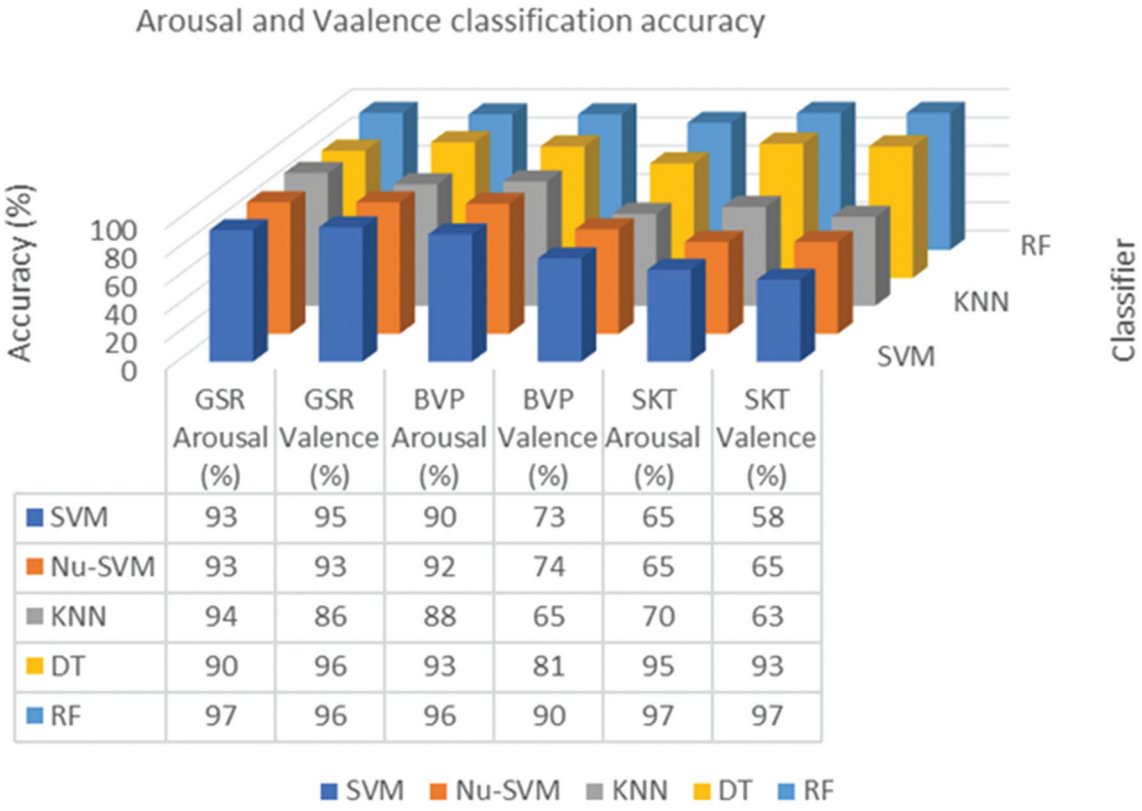
5. DISCUSSION

The main aim of this study was to predict specific arousal and valence values using Galvanic Skin Response (GSR), Blood Volume Pulse (BVP), and Skin Temperature (SKT). Fast Fourier Transform (FFT) features are widely used in Galvanic Skin Response (GSR) analysis because they decompose non-stationary physiological signals into interpretable frequency components. These components represent varying levels of arousal and emotional intensity. This technique facilitates differentiation between low-frequency tonic and high-frequency phasic components, which are directly correlated with emotional states. Furthermore, first- and second-derivative features are frequently employed, as they effectively capture dynamic and temporal changes (phasic components) in skin conductance associated with rapid emotional arousal. This methodology thus enhances the ability to detect phasic responses and assess sensitivity to emotional arousal effectively. The application of the Fast Fourier Transform (FFT) to photoplethysmography (PPG) facilitates the conversion of time-varying, complex signals into interpretable frequency components that exhibit strong correlations with emotional changes. These frequency components are particularly significant as variations in the power spectral density of the heart rate signal serve as direct indicators of both arousal and valence.

The first derivative features denote the velocity of blood volume changes, while the second derivative illustrates the acceleration of blood flow. Collectively, these derivatives enhance the signal's representation, rendering it more informative, robust, and capable of effectively discerning physiological changes associated with emotional states.

Fig. 9 shows the classification accuracies for valence and arousal emotion classification using unimodal physiological signals. It shows that decision tree and random forest classifiers demonstrated superior performance. Decision trees are capable of handling complex, high-dimensional data. Random forests, an ensemble method, combine multiple decision trees to improve accuracy, reduce overfitting, and handle high-dimensional and imbalanced data. Both approaches excel at capturing non-linear and intricate relationships within the data. For GSR and BVP, KNN showed the lowest accuracy, which could potentially be due to the curse of dimensionality caused by the numerous features and the distance between the data points in high-dimensional space, as well as imbalanced classes. The Random Forest algorithm consistently outperformed other classification methods. Skin temperature shows excellent performance in unimodal emotion classification, particularly in controlled environments. Unimodal systems are sensitive to specific sensor noise and are at higher risk of failure. For real-world applications, the primary challenge with unimodal skin temperature is isolating temperature changes attributable to weather or movement (e.g., outdoor-to-indoor transitions) from those attributable to emotional responses. Conversely, multimodal systems offer complementary data that enhances performance and compensates for missing information, thereby improving robustness and accuracy. Fig. 10 illustrates the classification accuracy achieved through the fusion of Galvanic Skin Response (GSR), Blood Volume Pulse (BVP), and Skin Temperature (SKT).

The late fusion method attained the highest accuracy rates of 98% for arousal and 99% for valence, surpassing the performance of the early fusion multimodal approach. The integration of mutual information with the Random Forest algorithm, coupled with majority voting, contributed to achieving these superior accuracy levels, while also enhancing the F1-score and Matthew's correlation coefficient (MCC).



	GSR Arousal (%)	GSR Valence (%)	BVP Arousal (%)	BVP Valence (%)	SKT Arousal (%)	SKT Valence (%)
SVM	93	95	90	73	65	58
Nu-SVM	93	93	92	74	65	65
KNN	94	86	88	65	70	63
DT	90	96	93	81	95	93
RF	97	96	96	90	97	97

Fig 9. Classification accuracy for unimodal GSR, BVP, and SKT

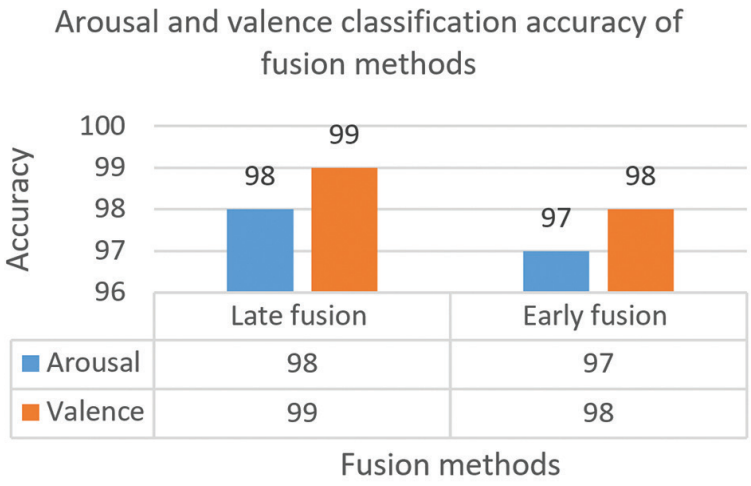


Fig 10. Classification accuracy for multimodal emotion classification

A comparison between unimodal and multimodal signals underscores the advantages of integrating diverse physiological data to enhance emotion classification. The Random Forest algorithm consistently outperformed other classification techniques in both unimodal and multimodal settings. This assertion is supported by the F1-score, MCC, and accuracy metrics, which indicate the model's efficacy relative to Support Vector Machines (SVM), Nu-SVM, K-Nearest Neighbors (KNN), and Decision Tree (DT) methods. The model is proficient in detecting valence and arousal levels within emotions. For instance, when distinguishing between emotions such as anger and calm, the valence classifier assigns low valence to anger and high valence to calm. Similarly, the arousal classifier recognizes high arousal in anger and low arousal in calmness. The methodology developed in this research has yielded substantial advances in the classification of arousal and valence, demonstrating significant improvements over models used by prior researchers.The best models for classifying physiological signals using valence-arousal were compared with recent approaches using the PHYMER dataset. Since there are currently no studies that have used this dataset for the classification of GSR, BVP, and SKT, this represents a gap in the existing comparisons. To address this issue, the results were compared

with existing research that used different datasets but employed similar or related classification methods, as shown in Tables 11, 12, and 13. Tables 11, 12, and 13 show that mutual information feature selection, combined with the Random Forest classifier, improves emotion classification accuracy for unimodal GSR, BVP, and SKT, compared with other studies.

This study contributes by developing a more accurate machine learning model for emotion classification. It is also the first study to apply the Chatterjee correlation coefficient feature selection technique for emotion classification and to combine GSR, BVP, and SKT multimodally via decision and feature-level fusion.

Table 11. Emotion classification accuracy for the GSR signal

Sr. No.	Reference	Dataset	Feature selection	Classifier	Accuracy
1	[17]	DEAP	-	SVM	Arousal = 64.4% Valence = 65%
2	[27]	AMIGOS	Mutual information	SVM	Arousal = 85.75% Valence = 83.9%
3	[11]	AMIGOS	Mutual information	KNN	Arousal = 94% Valence = 96%
4	[32]	DEAP	-	RF	Arousal = 71.53% Valence = 71.04%
5	Present work	Phymer	Mutual information	RF	Arousal = 97% Valence = 96%

Table 12. Classification for BVP (PPG) signal

Sr. No.	Reference	Dataset	Feature selection	Classifier	Accuracy
1	[19]	Experiment, DEAP	-	SVM	Experiment Arousal = 85.9% Valence = 91% DEAP Arousal = 79.3% Valence = 75.9%
2	[16]	A2ES	-	SVM	Arousal = 64.61% Valence = 64.94%
3	[32]	DEAP	-	RF	Arousal = 70.92% Valence = 70.76%
4	[33]	DEAP	Minimum redundancy maximum relevance	RF	Arousal = 68.59% Valence = 70.23%
5	Present work	Phymer	Mutual information	RF	Arousal = 96% Valence = 90%

Table 13. Classification accuracy for the SKT signal

Sr. No.	Reference	Dataset	Feature selection	Classifier	Accuracy
1	[17]	DEAP	-	DT	Arousal = 64.7% Valence = 65.3%
2	Present work	Phymer	Mutual information	RF	Arousal = 97% Valence = 97%

6. CONCLUSION

This study investigates the use of physiological signals, including Galvanic Skin Response (GSR), Blood Volume Pulse (BVP), and Skin Temperature (SKT), for emotion recognition, demonstrating their effectiveness in accurately identifying emotions. Emotion classification offers advantages across various fields, such as healthcare, mental health, customer service, human-computer interaction, education, e-learning, automation, and market research. In the healthcare and mental health sectors, applications range from autism and Parkinson's disease management to stress management and mental health monitoring. Recent advancements in machine learning and deep learning techniques have significantly enhanced these applications. This study makes a novel contribution to affective computing by developing a unimodal and multimodal emotion classification model that evaluates all three physiological parameters: GSR, BVP, and SKT. This is also the first study to use GSR, BVP, and SKT for both early and late fusion in emotion classification. Various feature selection techniques were explored, and a new technique,

the Chatterjee Correlation, was employed in this study. Furthermore, the accuracy of emotion classification using the ML classifiers improved compared to existing studies. Thus, this study makes a valuable contribution to the use of next-generation AI in healthcare systems. The PHYMER dataset is limited in scope as it contains physiological signal data exclusively from students at the same university and with a similar cultural background. Future research could test the model on individuals from diverse cultural backgrounds, across various age groups, and with different health conditions. This is particularly relevant, as the combination of physiological signals is often required for effective emotion recognition, yet many existing datasets focus on specific types of physiological signals. Moving forward, it will be essential to assess the model's performance across a broader demographic by incorporating a variety of physiological signals. The performance of the multimodal system may be affected by variations in ambient conditions, stress, and motion artifacts. These factors can lead to the extraction of signals that are substantially compromised by noise. The model was evaluated using a dataset collected within a controlled environment, where both temperature and motion were maintained consistently. Consequently, it is essential to move beyond laboratory settings to real-world applications in future research. Future efforts should focus on incorporating additional physiological signals, including respiration rate, electromyography (EMG), electrooculography (EOG), electrocardiography (ECG), and optomyography (OMG). By leveraging deep learning methodologies, we can enhance multiclass emotion classification to more accurately capture the complex and nuanced nature of human emotions. It is also important to test the model on different datasets.

ACKNOWLEDGEMENT

The authors are thankful to the Goa College of Engineering, affiliated with Goa University, for supporting the work carried out in this study.

DECLARATION OF THE USE OF AI TOOLS

The authors used Grammarly for grammatical error and typo correction. All AI-assisted content was reviewed and edited by the authors, who take full responsibility for the integrity and accuracy of this work. No AI tool is listed as an author, and AI tools are not cited as primary sources.

REFERENCES

[1] G. Pei, H. Li, Y. Lu, Y. Wang, S. Hua, T. Li, "Affective Computing: Recent Advances, Challenges, and Future Trends", Intelligent Computing, Vol. 3, 2024. doi: 10.34133/icomputing.0076.

[2] A. U. Dessai, H. G. Virani, "Emotion Detection Using ECG Signals and a Lightweight CNN Model", Computer Systems Science and Engineering, Vol. 48, No. 5, 2024, pp. 1193-1211. doi: 10.32604/csse.2024.052710.

[3] A. Saavedra, R. Chocarro, M. Cortinas, N. Rubio, "Exploring unconscious user responses to affective computing in interactive prototypes: a consumer neuroscience study", Behaviour & Information Technology, Vol. 44, No. 20, 2025, pp. 1-20. doi: 10.1080/0144929X.2025.2504514.

[4] M. Endah Hiswati, E. Utami, K. Kusrini, A. Setyanto, "Insights into the Emotion Classification of Artificial Intelligence: Evolution, Application, and Obstacles of Emotion Classification", Engineering Proceedings, Vol. 103, No. 1, 2025, p. 24. doi: 10.3390/engproc2025103024.

[5] Y. S. Can, B. Mahesh, E. André, "Approaches, Applications, and Challenges in Physiological Emotion Recognition—A Tutorial Overview", Proceedings of the IEEE, Vol. 111, No. 10, 2023, pp. 1287-1313. doi: 10.1109/JPROC.2023.3286445.

[6] N. Gudikandula, R. Janapati, R. Sengupta, S. Chintala, "Brain computer interface based emotion recognition with error analysis and challenges: an interdisciplinary review", Discover Applied Sciences, Vol. 7, No. 8, 2025, p. 857. doi: 10.1007/s42452-025-06692-0.

[7] E. M. Polo, F. Iacomi, A. V. Rey, D. Ferraris, A. Paglialonga, R. Barbieri, "Advancing emotion recognition with Virtual Reality: A multimodal approach using physiological signals and machine learning", Computers in Biology and Medicine, Vol. 193, 2025, p. 110310. doi: 10.1016/j.compbiomed.2025.110310.

[8] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014-2023) and research recommendations", Information Fusion, Vol. 102, 2024, p. 102019. doi: 10.1016/j.inffus.2023.102019.

[9] S. N. M. Sayed Ismail, N. A. Ab. Aziz, S. Z. Ibrahim, M. S. Mohamad, "A systematic review of emotion recognition using cardio-based signals", ICT Express, Vol. 10, No. 1, 2024, pp. 156-183. doi: 10.1016/j.icte.2023.09.001.

[10] L.-C. Yu *et al*. "Building Chinese Affective Resources in Valence-Arousal Dimensions", Proceedings of the Conference of the North American Chapter of the

Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, June 2016, pp. 540-545. doi: 10.18653/v1/N16-1066.

[11] A. Dessai, H. Virani, "Multimodal and Multidomain Feature Fusion for Emotion Classification Based on Electrocardiogram and Galvanic Skin Response Signals", Sci, Vol. 6, No. 1, p. 10, 2024. doi: 10.3390/sci6010010.

[12] A. Dessai, H. Virani, "Lightweight 1D CNN model for emotion classification using GSR signal", International Journal of Innovative Research and Scientific Studies, Vol. 8, No. 3, 2025, pp. 5100-5116. doi: 10.53894/ijirss.v8i3.7709.

[13] M. Saleem, S. Ullah, D.-H. Kim, "Dual-Branch Residual CNN-LSTM Model for Emotion Classification using PPG and GSR", Proceedings of the IEEE International Conference on Big Data and Smart Computing, Kota Kinabalu, Malaysia, 9-12 February 2025, pp. 252-255. doi: 10.1109/BigComp64353.2025.00055.

[14] A. Dessai, H. Virani, "Emotion Classification Based on CWT of ECG and GSR Signals Using Various CNN Models", Electronics, Vol. 12, No. 13, 2023, p. 2795. doi: 10.3390/electronics12132795.

[15] N. A. Ab. Aziz *et al*. "Asian Affective and Emotional State (A2ES) Dataset of ECG and PPG for Affective Computing Research", Algorithms, Vol. 16, No. 3, 2023, p. 130. doi: 10.3390/a16030130.

[16] S. N. M. Sayed Ismail, N. A. Ab. Aziz, S. Z. Ibrahim, "A comparison of emotion recognition system using electrocardiogram (ECG) and photoplethysmogram (PPG)", Journal of King Saud University - Computer and Information Sciences, Vol. 34, No. 6, 2022, pp. 3539-3558. doi: 10.1016/j.jksuci.2022.04.012.

[17] J. Gohumpu, M. Xue, Y. Bao, "Emotion recognition with multi-modal peripheral physiological signals", Frontiers in Computer Science, Vol. 5, 2023. doi: 10.3389/fcomp.2023.1264713.

[18] H. Alharbi, "Explainable feature selection and deep learning based emotion recognition in virtual reality using eye tracker and physiological data", Frontiers in Medicine, Vol. 11, 2024. doi: 10.3389/fmed.2024.1438720.

[19] Z. Zhu *et al*. "An emotion recognition method based on frequency-domain features of PPG", Frontiers in Physiology, Vol. 16, 2025. doi: 10.3389/fphys.2025.1486763.

[20] M. F. Bamonte, M. Risk, V. Herrero, "Determining the Optimal Window Duration to Enhance Emotion Recognition Based on Galvanic Skin Response and Photoplethysmography Signals", Electronics, Vol. 13, No. 16, 2024, p. 3333. doi: 10.3390/electronics13163333.

[21] E. Mattern, R. R. Jackson, R. Doshmanziari, M. Dewitte, D. Varagnolo, S. Knorn, "Emotion Recognition from Physiological Signals Collected with a Wrist Device and Emotional Recall", Bioengineering, Vol. 10, No. 11, 2023, p. 1308. doi: 10.3390/bioengineering10111308.

[22] A. Kumar, A. Kumar, "Human emotion recognition using Machine learning techniques based on the physiological signal", Biomedical Signal Processing and Control, Vol. 100, 2025, p. 107039. doi: 10.1016/j.bspc.2024.107039.

[23] D.-H. Kang, D.-H. Kim, "1D Convolutional Autoencoder-Based PPG and GSR Signals for Real-Time Emotion Classification", IEEE Access, Vol. 10, 2022, pp. 91332-91345. doi: 10.1109/ACCESS.2022.3201342.

[24] M. Saleem, D.-H. Kim, "Multimodal Biosignals for Affective State Recognition Using Differential Multimodal Transformer", IEEE Access, Vol. 13, 2025, pp. 202395-202407. doi: 10.1109/ACCESS.2025.3637310.

[25] S. Pant, H.-J. Yang, E. Lim, S.-H. Kim, S.-B. Yoo, "PhyMER: Physiological Dataset for Multimodal Emotion Recognition With Personality as a Context", IEEE Access, Vol. 11, 2023, pp. 107638-107656. doi: 10.1109/ACCESS.2023.3320053.

[26] P. Bota, R. Silva, C. Carreiras, A. Fred, H. P. da Silva, "BioSPPy: A Python toolbox for physiological signal processing", SoftwareX, Vol. 26, 2024, p. 101712. doi: 10.1016/j.softx.2024.101712.

[27] J. Shukla, M. Barreda-Angeles, J. Oliver, G. C. Nandi, D. Puig, "Feature Extraction and Selection for Emotion Recognition from Electrodermal Activity", IEEE Transactions on Affective Computing, Vol. 12, No. 4, 2021, pp. 857-869. doi: 10.1109/TAFFC.2019.2901673.

[28] L. Huang, X. Zhou, L. Shi, L. Gong, "Time Series Feature Selection Method Based on Mutual Information", Applied Sciences, Vol. 14, No. 5, 2024, p. 1960. doi: 10.3390/app14051960.

[29] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines", npj Digital Medicine, Vol. 3, No. 1, 2020, p. 136. doi: 10.1038/s41746-020-00341-z.

[30] S. Abdumalikov, J. Kim, Y. Yoon, "Performance Analysis and Improvement of Machine Learning with Various Feature Selection Methods for EEG-Based Emotion Classification", Applied Sciences, Vol. 14, No. 22, 2024, p. 10511. doi: 10.3390/app142210511.

[31] J. A. Domínguez-Jiménez, K. C. Campo-Landines, J. C. Martínez-Santos, E. J. Delahoz, S. H. Contreras-Ortiz, "A machine learning model for emotion recognition from physiological signals", Biomedical Signal Processing and Control, Vol. 55, 2020, p. 101646. doi: 10.1016/j.bspc.2019.101646.

[32] D. Ayata, Y. Yaslan, M. E. Kamasak, "Emotion Based Music Recommendation System Using Wearable Physiological Sensors", IEEE Transactions on Consumer Electronics, Vol. 64, No. 2, 2018, pp. 196-203. doi: 10.1109/TCE.2018.2844736.

[33] D. Ayata, Y. Yaslan, M. E. Kamasak, "Emotion Recognition from Multimodal Physiological Signals for Emotion Aware Healthcare Systems", Journal of Medical and Biological Engineering, Vol. 40, No. 2, 2020, pp. 149-157. doi: 10.1007/s40846-019-00505-7.