Leveraging Machine Learning and Feature Extraction from Physiological Signals for Multimodal Emotion Detection and Mental Health Support

Ashvik Raina

Cambridge Center of International Research

Dr. Shadi Ghiasi

Abstract

Mental health conditions like anxiety and depression often involve emotional dysregulation, but most methods used to measure emotions depend on people's own subjective reports. The goal of this study is to use signal processing and machine learning to create a system that can objectively classify emotional states based on physiological signals, especially EEG and ECG, to support real-time mental health monitoring and intervention.

For this research, the DREAMER dataset and features from EEG and ECG channels for 23 subjects were used. Bandpass filtering and z-score normalization were used for preprocessing to remove noise and make the data more consistent. Band power, entropy, and HRV were extracted from 14 EEG and 2 ECG channels. To try and distinguish between baseline and stimulus conditions, multiple classification models were evaluated, including Gradient Boosting, Random Forest, AdaBoost Classifier, Logistic Regression, and K-Nearest Neighbors. Subject-exclusive train-test splits ensured generalizability. The Gradient Boosting model achieved the highest performance among all tested algorithms, with an F1 score of 76.74%, an accuracy of

77.78%, a precision of 80.49%, and a recall of 73.33%, suggesting that EEG and ECG signals offer a promising foundation for developing objective, physiology-based emotion recognition systems. The multimodal machine learning approach used in this study also lends future work to incorporate additional bio-signals, such as skin temperature from smartwatches, to enhance emotion detection accuracy in clinical and personal health applications.

Keywords: Physiological signals, emotions, machine learning, feature extraction, classification

Introduction

Emotions are central to human cognition, influencing decision-making, social interactions, and overall psychological well-being (Picard, 2000). They control behavior, shape memory, and attention, and help individuals manage stress better. Emotional dysregulation is also one of the main symptoms of two of the more common mental illnesses: anxiety and depression. According to recent epidemiological studies, over 970 million people globally suffer from a mental disorder, with anxiety and depressive disorders accounting for the majority (Kumar, et al., 2024). Traditionally, emotional assessments rely heavily on self-reported data. This approach is skewed by personal preferences and differences in how individuals see things. Objective physiological measures, on the other hand, like those derived from neurophysiological signals, show promise to make mental health evaluations more accurate and reliable. By figuring out how feelings are expressed through measurable biological responses, researchers make better tools for early detection, personalized treatment, and constant mental health monitoring.

Physiological signals serve as objective indicators of internal emotional and cognitive states by capturing the body's neurological and autonomic responses. For example, Electroencephalography (EEG) records electrical signals made by neuronal oscillations and is often used to measure brain function. These signals are typically decomposed into distinct

frequency bands, each associated with different cognitive and affective functions, such as attention, relaxation, and arousal (Klimesch, 1999).

Electrocardiography (ECG) keeps track of the heart's electrical activity and is often used to figure out how one's feeling and what's going on with their bodies. Features like heart rate variability, mean heart rate, and frequency-domain measures can show how stressed, emotionally aroused, and active the autonomic nervous system is. Combining EEG and ECG gives us different but complementary views on how emotions work, which makes it possible to recognize emotions in a more complete and objective way than with standard self-report methods.

The Valence-Arousal-Dominance (VAD) model used in the DREAMER dataset derives feelings in a continuous three-dimensional space. The first dimension, valence, is the degree of good or bad feelings that are associated with an emotional experience. The second dimension, arousal, shows how calm or excited somebody is feeling. The third dimension used by this model is dominance. Dominance represents the level of control or impact a person thinks he/she has over a situation. The VAD model derives emotional states in a more flexible and continuous way than standard categorical emotion models, which divide emotions into discrete values like fear, anger, or happiness. This three-dimensional method helps us understand emotions in more depth and works especially well for tracking them in real time for things like mental health evaluations. The VAD model is widely used in affective computing for representing complex emotional states.

This study explores how EEG and ECG signals can be used to build machine learning models for emotion recognition, aiming to improve the accuracy and objectivity of mental health assessments.

Approach

The research question guiding this work is: Can emotion-related physiological states be accurately classified using features extracted from EEG and ECG signals specifically by focusing on the two-way separation of baseline states and stimulus-induced states?

For this study, to understand the subject's mental and physiological state, Electroencephalography (EEG) and Electrocardiography (ECG) signals were combined across multiple trials and channels to create a multimodal feature set. This representation captures both central and peripheral physiological responses to emotional stimuli and serves as the input for subsequent machine learning classification models. Studies have shown that band power and entropy, two features of the EEG, greatly relate to cognitive load and emotional processing (Lin, et al., 2010). All data was preprocessed to get rid of noise and improve quality before feature extraction. To remove low-frequency drift and high-frequency artifacts and focus on the important neural activity, i.e., remove inter-subject variability, the EEG data was run through a fourth-order Butterworth bandpass filter, which had cutoff frequencies set between 0.5 Hz and 45 Hz. After the signals were filtered, they were normalized using Z-score normalization to ensure they could be compared across people and channels. After sampling the ECG signals at a higher frequency, the NeuroKit2 library was used to get clean heartbeats and heart rate variability measures (Makowski et al., 2021). Then, features such as time-domain, frequency-domain, entropy-based, and heart rate variability (HRV) were extracted from both types of signals. These features were used to train several machine learning models, such as K-Nearest Neighbors, Gradient Boosting, Random Forest, AdaBoost Classifier, and Logistic Regression.

Welch's method was used to figure out the frequency-domain features of the EEG and the power spectral density (PSD) for each channel. This method is frequently used in EEG analysis as it provides stable spectral values and is not affected by noise (Cohen, 2014). These

were used to find the absolute and relative band strengths for five common frequency bands: delta, theta, alpha, beta, and gamma. Signal entropy, peak power, and nonlinear descriptors like Hjorth parameters were some of the other traits used in this study. Heart rate variability was calculated using the ECG data for time and frequency. These included the mean of Normal-to-Normal intervals, the standard deviation of Normal-to-Normal intervals, the root mean square of successive differences, and the ratio of low-frequency to high-frequency power. Autonomic regulation and emotional arousal can be seen in these measures (Shaffer & Ginsberg, 2017).

Past studies have shown that physiological signals like EEG and ECG can be helpful in determining how people are feeling. The DEAP dataset introduced by Koelstra et al. (2022) is among the earliest large-scale efforts to collect EEG and peripheral physiological signals during various emotional states. Their work laid the foundation for machine learning to classify emotional states. Alarcao & Fonseca (2017) did a thorough review of EEG-based emotion recognition methods. They showed how important preprocessing, feature extraction, and model selection are for getting good results. These studies show that EEG data can be used to classify emotions, and they have led to more research that builds on these signal processing and machine learning methods.

Materials and Methods

Dataset Description

This study utilizes the DREAMER dataset. It is a set of neurophysiological data that was made to help people recognize emotions. Among the 23 people who took part were 13 men and 10 women, ages 18 to 25. 18 emotionally charged video clips were shown to the participants to get a range of emotional reactions. After watching each movie, they used the Valence-Arousal-Dominance (VAD) model to rate their own emotional state.

There are both EEG and ECG signals in the sample. EEG was recorded with 14 channels at a sampling rate of 128 Hz, which picked up electrical activity in the brain in a few different areas of the scalp. ECG was taken at 256 Hz at the same time to see how the heart was beating. This dataset includes a baseline recording and a stimulus recording for each trial, giving the ability to compare states of being at rest and being mentally stimulated. Using both types of data also makes it possible to investigate how emotions are reflected in different types of body data. The DREAMER dataset is perfect for building machine learning models that can classify emotions because it has both subjective rates of emotions and objective physiological signals (Katsigiannis & Ramzan, 2017).

Signal Preprocessing

EEG signals are prone to noise contamination from ocular artifacts, muscle activity, electrical interference, and other noise, so they need to be preprocessed before they can be used to extract features. A Butterworth bandpass filter was used to get rid of frequency components that weren't needed. For EEG, the cutoff frequencies were set between 0.5 and 45 Hz. The Butterworth filter transfer function is defined as:

$$H(f) = \frac{1}{\sqrt{1 + \left(\frac{f}{f_c}\right)^{2N}}}$$

where f_c is the cutoff frequency, and N is the filter order. After filtering, Z-score normalization was performed to standardize the signals across subjects and trials:

$$X' = \frac{X - \mu}{\sigma}$$

where X is the raw signal, μ is the mean, and σ is the standard deviation.

The NeuroKit2 library was used to do preprocessing on the ECG data. This included filtering, R-peak detection, and automatic artifact correction. Clean inter-beat intervals (NN intervals) were used for heart rate variability (HRV) research later. These steps made sure that the signals were not badly distorted and were best for extracting features and then classifying them. Here is an example of part of an EEG signal that was shown before and after preprocessing (Figure 1).

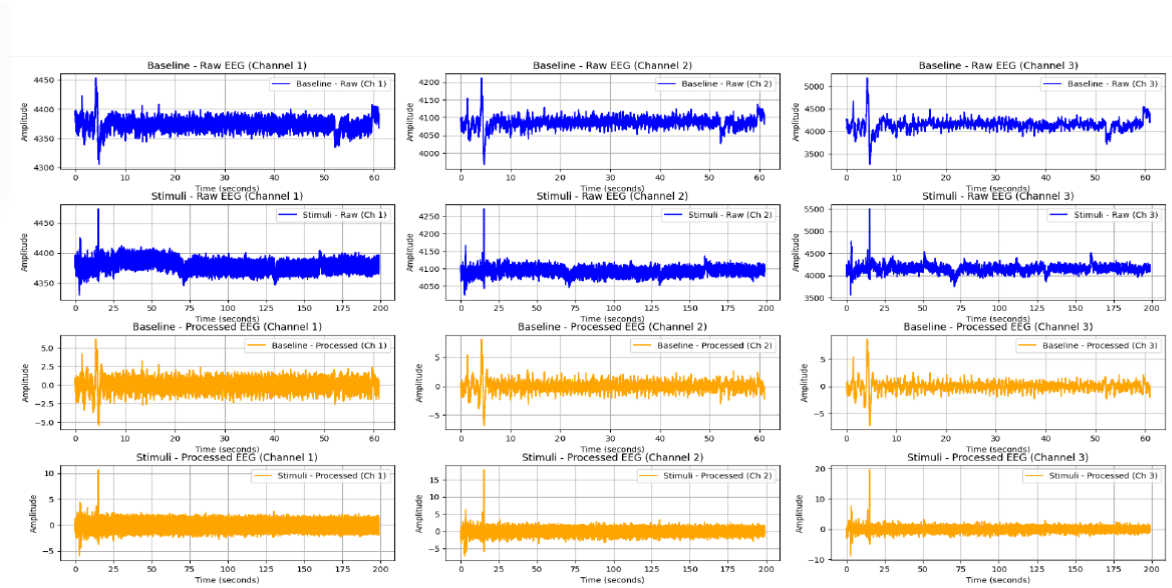

Figure 1. Comparison of raw EEG (top) and preprocessed EEG (bottom).

Feature Extraction

After preprocessing, features shown in Table 1 were extracted from EEG and ECG signals to quantify emotional states. These features were categorized into time-domain, frequency-domain, and nonlinear measures.

TABLE 1. EXTRACTED FEATURES PER EEG AND ECG CHANNEL

Feature Type	Description
Max/Min	Amplitude extrema in the signal
Total Power	Total spectral power
Relative Power	Max power / total power
Spectral Entropy	Signal complexity in frequency domain
Hjorth Complexity	Nonlinear mobility and complexity
Delta-Gamma Powers	Band-specific power (0.5–45 Hz)
Mean HR	Avg. heart rate from ECG
SDNN	Std. dev. of NN intervals
LF/HF Ratio	Frequency-domain HRV ratio

EEG Time-Domain Features

The extracted features include the maximum (max(x)) and minimum (min(x)) values of the signal, as well as energy, which represents the total signal power and is computed as:

$$E = \sum_{i=1}^{N} x_i^2$$

EEG Frequency-Domain Features

Frequency-domain features were derived using Welch's method for Power Spectral Density (PSD) estimation (P. Welch, 2003).

$$S_{xx}(f) = \frac{1}{M} \sum_{m=1}^{M} |X_m(f)|^2$$

where $X_m(f)$ is the Fourier Transform of the windowed signal, and M is the number of overlapping segments.

Extracted features include the peak frequency with its corresponding power, total power and relative power, and the band powers across Delta, Theta, Alpha, Beta, and Gamma frequency ranges.

$$P_{\text{band}} = \sum_{f=f_{\text{low}}}^{f_{\text{high}}} S_{xx}(f)$$

EEG Power Ratios

Power ratios were calculated to provide insights into neural oscillation relationships. These included the Theta/Beta ratio, which is linked to attentional control and emotional regulation; the Alpha/Theta ratio, which relates to cognitive processing; and the Alpha/Beta ratio, which is associated with relaxation and stress levels.

EEG Nonlinear Features

Nonlinear dynamics of EEG signals were captured using Spectral entropy and Hjorth complexity. Spectral entropy was computed from the PSD as:

$$H_s = -\sum_i p_i \log p_i$$

where p_i represents the normalized power at frequency bin *i,* and Hjorth complexity was computed as the ratio of mobility between the second derivative and the first derivative:

$$\text{Complexity} = \frac{\text{Mobility}(\frac{d^2x}{dt^2})}{\text{Mobility}(\frac{dx}{dt})}$$

These extracted features provide a comprehensive representation of EEG signals, enabling effective classification of emotional states.

ECG Features (per channel)

ECG features extracted using the NeuroKit2 library included standard heart rate variability (HRV) metrics such as Mean HR, which is the average inter-beat interval (Mean NN); SDNN, the standard deviation of NN intervals; and the LF/HF ratio, which represents the balance between low-frequency and high-frequency power.

Feature Extraction Process

Raw EEG and ECG signals were first preprocessed and then evaluated channel by channel to extract the features used in this study. For EEG, NumPy operations were used to get time-domain features like maximum, minimum, and energy directly from the signal amplitude arrays. To calculate energy, the signal values were squared and summed across the time axis. Frequency-domain features were extracted by applying Welch's method to estimate the Power Spectral Density (PSD) of each EEG channel using SciPy's welch() function. From the PSD, the power within standard EEG frequency ranges (delta, theta, alpha, beta, and gamma) was added up to get the band powers. The Shannon entropy formula was used to find the spectral entropy after normalizing the PSD to make a probability distribution. The Hjorth complexity was found by first and second-order differencing of the EEG signal and then finding the variance ratios between the derivatives to show the shape and irregularity of the signal.

For preprocessing and feature extraction from ECG data, the NeuroKit2 Python library was used. It found R-peaks and calculated inter-beat (NN) intervals algorithmically. These calculations were then used to determine standard heart rate variability (HRV) measurements, like Mean HR, SDNN (standard deviation of NN intervals), and LF/HF (low-frequency to high-frequency power ratio). These features were selected based on their known relevance to emotional arousal and autonomic nervous system function. All the extracted features were saved in a structured data frame, which had different columns for each channel and type of feature.

Machine Learning Model

This study used supervised machine learning models that were trained on features taken from both EEG and ECG signals. To tell the difference between baseline (resting) and stimuli (emotionally induced) situations, the task was set up as a binary classification problem. The

dataset was split up by subject to ensure that all the trials from each individual were only in the training set or the testing set. This method helps check how well the model can apply to people it hasn't seen before. 80% of the people who were tested were put in the training set and 20% were kept for testing.

To make the feature matrix, all the columns that weren't linked to the subject's metadata or labels were taken out, including Valence, Arousal, Dominance, Trial, and Subject. This left only quantitative EEG and ECG features which made up data set X. The goal vector y was the binary class label (”Baseline” or ”Stimuli”). Features were not scaled up or down in any other way. Random seeds were also chosen to make sure that the results could be repeated between runs. The study's random number was 42. Cross-validation was not used in this study because the data was split up by person using the 'traintestsplit′ function from scikit-learn.

The classifiers evaluated included Random Forest, Gradient Boosting, AdaBoost, Logistic Regression, and K-Nearest Neighbors. These models were chosen because they had previously performed well in bio-signal classification tasks and could be used in psychological studies (Jenke et al., 2014). For example, Random Forest and Gradient Boosting have constantly been ranked among the best classifiers for figuring out emotions from physiological data (Zheng & Lu, 2015). The scikit-learn library was used to build all the models with 100 estimators for ensemble methods and a set random seed to make sure the results were able to be replicated (Pedregosa et al., 2011). Z-score normalization was then used to make the features standardized, and all models used the same train-test splits.

Results

Model Performance Evaluation

Several performance measures, such as accuracy, precision, recall, and F1 score, were calculated to see how well the classification models worked. These measures give a fuller picture of how the model works than just its overall accuracy. Table II shows how well all five models did on the test set. With an accuracy of 77.78% and an F1 score of 76.74%, the Gradient Boosting model did the best. Random Forest did well on all of the same measures. Logistic regression had the highest recall rate of 78.89%, but it had a lower accuracy rate. The worst overall performance was by K-Nearest Neighbors (KNN).

TABLE II. CLASSIFICATION PERFORMANCE ON TEST SET

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.7500	0.7848	0.6889	0.7337
Gradient Boosting	0.7778	0.8049	0.7333	0.7674
AdaBoost	0.7056	0.7761	0.5778	0.6624
Logistic Regression	0.6500	0.6174	0.7889	0.6927
KNN	0.6111	0.6667	0.4444	0.5333

Confusion Matrix Analysis

Figure 2 shows confusion matrices that show how well each classifier can tell the difference between the “Baseline” and “Stimuli” classes. For each model, these models show how many true positives, true negatives, false positives, and false negatives there are.

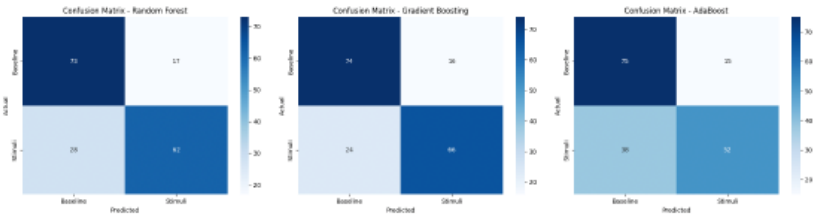

Figure 2. Confusion matrices for the top 3 models. Random Forest (left), Gradient Boosting (middle), and AdaBoost (right). Rows represent actual classes; columns represent predicted classes.

The Gradient Boosting model had the best mix of accuracy and recall, especially when it came to correctly identifying “Stimuli” instances. Gradient Boosting is a type of ensemble

learning that creates a set of decision trees one after the other. Each new tree is trained to correct the prediction errors made by the previous trees by minimizing a specified loss function. In contrast to Random Forest, which builds trees separately and averages their outcomes, Gradient Boosting works on making the model better over time by learning from the residuals. Random Forest was slightly better at finding "Baseline" trials, but it also had more false positives. AdaBoost did not do as well overall, but it still managed to correctly identify most of the trials.

Discussion

The study shows that machine learning models trained on EEG and ECG signal data can determine the difference between normal and emotionally charged states. Out of the five models that were tested, Gradient Boosting performed best overall across all measures, showing that it is good at dealing with structured, high-dimensional feature sets that are common in physiological data. Random Forest model also performed well, as it was able to handle noise and find complicated, nonlinear relationships in the data. Models like K-Nearest Neighbors (KNN) and AdaBoost did not work as well because they seemed sensitive to noise. KNN assumes local similarity in feature space, which may not hold consistently across subjects due to high intersubject variability in physiological responses. Model variance may have also been caused by the fact that the feature space was very multidimensional, and there were only a few examples per subject. A subject-exclusive train-test split is an important thing to factor in, as this validation method acts like the real world by ensuring that the model can work with people it hasn't seen before instead of just remembering patterns from the same subjects. This may have adversely impacted the performance compared to random splitting, but it did make the results more reliable and like real-world deployments. The idea to include both EEG and ECG features

worked well. The EEG measured activity in the central nervous system by looking at band power, entropy, and complexity while the ECG measured heart rate variability (HRV). This multimodal approach enabled the models to consider both cognitive and physiological aspects of emotional response, resulting in a richer feature space for classification. Despite promising results, limitations include the relatively small dataset size and lack of cross-validation.

In the future, researchers could investigate deep learning techniques for feature extraction and classification, subject calibration strategies, or temporal models that show how signals change over time. Incorporating other bio-signals, such as electrodermal activity (EDA) or breathing, along with computer vision and voice analysis, should help with detecting emotions more accurately real time.

Conclusion

This study concludes that emotional states can be effectively classified using physiological signals from EEG and ECG recordings. By applying signal preprocessing, extracting informative features, and training the right machine learning models, a strong classification performance between baseline and emotionally stimulated conditions was achieved. Random Forest and Gradient Boosting classifiers had accuracy scores exceeding 75% and were found well-suited for emotion recognition tasks. These findings support the use of physiological signals as reliable, objective indicators of emotional state, laying the groundwork for future applications in real-time emotion tracking.

References

Alarcao, S. M., & Fonseca, M. J. (2017). "Emotions recognition using eeg signals: A survey," *IEEE transactions on affective computing*, vol. 10, no. 3, (pp. 374–393).

Cohen, M. X. (2014). Analyzing neural time series data: theory and practice. *MIT press*.

Jenke, R., Peer, A., & Buss, M. (2014). "Feature extraction and selection for emotion recognition from eeg," *IEEE Transactions on Affective computing*, vol. 5, no. 3, (pp. 327–339).

Katsigiannis, S., & Ramzan, N. (2017). "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off the-shelf devices," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, (pp. 98–107).

Klimesch, W. (1999). "Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis," *Brain research reviews*, vol. 29, no. 2-3, (pp. 169 195).

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & Patras, I. (2011). "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, (pp. 18–31).

Kumar, P. M., Kumar, V. U., Meenakshi, S., Bahekar, T. N., Narapaka, P. K., Dhingra, S., & Murti. K. (2024). "Epidemiology and risk factors of mental disorders," in *A Review on Diverse Neurological Disorders*, (pp. 3–12), Elsevier.

Lin, 4Y.-P., Wang, C.-H., Jung, T.-P., Wu, T.-L., Jeng, S.-K., Duann, J.-R., & Chen, J.-H. (2010). "Eeg-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. (1798–1806).

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., SchÅNolzel, C., & Chen, S. A. (2021). "Neurokit2: A python toolbox for neurophysiological signal processing," *Behavior research methods*, (pp. 1–8).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. (2825–2830).

Picard, R. W. (2000). Affective Computing. MIT Press.

Shaffer, F., & Ginsberg, J. P. (2017). "An overview of heart rate variability metrics and norms," *Frontiers in public health*, vol. 5, (p. 258).

Welch, P. (2003). "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on audio and electroacoustics*, vol. 15, no. 2, pp. (70–73).

Zheng, W.-L., & Lu., B.-L. (2015). "Investigating critical frequency bands and channels for eeg based emotion recognition with deep neural networks," *IEEE Transactions on autonomous mental development,* vol. 7, no. 3, pp. (162–175).