Cureus Journal of Computer Science

Open Access Original Article

Development of Real-Time Electroencephalography-Based Emotion Recognition Models Using Machine Learning Techniques

Received 06/13/2025
Review began 06/29/2025
Review ended 09/16/2025
Published 09/26/2025

DOI:
https://doi.org/10.7759/s44389-025-06161-0

Tijesunimi D. Olashore [1] , Olutayo K. Boyinbode [2] , Oladunni A. Daramola [2] , Mary T. Kinga [2]

1. Computer Science, Federal University of Technology, Akure, NGA 2. Information Technology, Federal University of Technology, Akure, NGA

Corresponding authors: Tijesunimi D. Olashore, tijesunimiolashore@gmail.com, Olutayo K. Boyinbode, okboyinbode@futa.edu.ng, Oladunni A. Daramola, oadaramola@futa.edu.ng, Mary T. Kinga, mtkinga@futa.edu.ng

Abstract

Emotion recognition using electroencephalography (EEG) signals has gained significant attention in affective computing, human-computer interaction, and mental health applications. Owing to the complex and non-stationary nature of EEG signals, diverse feature extraction methods - spanning time-domain (statistical), frequency-domain, and deep learning approaches - have been explored to capture intrinsic signal characteristics. This study presents a critical evaluation of three machine learning models for EEG-based emotion classification, leveraging both time- and frequency-domain features. Specifically, the performance of eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and K-Nearest Neighbors (KNN) was investigated using the DEAP (Database for Emotion Analysis using Physiological Signals) dataset, a widely adopted benchmark for emotion recognition research. To enhance signal quality, EEG data were segmented into 1-second epochs, from which a comprehensive set of statistical and mathematical descriptors were extracted: mean, standard deviation, root mean square, minimum, maximum, peak value, activity, mobility, complexity, entropy, skewness, and kurtosis in the time-domain, along with entropy in the frequency-domain. A 10-fold cross-validation procedure was applied under a cross-subject evaluation scheme to assess generalization across individuals. Experimental results demonstrated that KNN achieved the highest classification accuracy (96.77%), surpassing XGBoost (88.55%) and LightGBM (84.08%), thereby challenging the conventional preference for more complex ensemble models in EEG-based emotion recognition. These findings highlight the critical role of feature engineering and classifier selection in optimizing system performance. The outcomes have practical implications in clinical diagnostics (e.g., early detection of depression and anxiety) and adaptive systems, including mood-aware interfaces, neurofeedback applications, and emotionally adaptive gaming environments.

Categories: Affective Computing, Health Informatics, Machine Learning (ML)
Keywords: emotion recognition, electroencephalography (eeg), segmentation, statistical features, k nearest neighbours (knn), light gradient boosting (lightgbm), extreme gradient boosting (xgboost), machine learning, frequency features, deep learning

Introduction

Human emotion plays a pivotal role in both affective computing and human-machine interaction. As Picard et al. [1] observed, affective computing seeks to bridge the gap between computational systems and the human emotional experience, recognizing that emotions significantly influence user behavior and the effectiveness of interactive systems. Emotional regulation is also closely linked to mental health, as difficulties in managing emotions are associated with conditions like attention-deficit disorders and depression [2]. In this context, electroencephalography (EEG)-based emotion detection serves a dual purpose: Diagnostically, it provides objective neural markers that can support the early detection and monitoring of mental health conditions such as anxiety, depression, and Attention-Deficit/Hyperactivity Disorder; behaviorally, it enables real-time emotion monitoring in human-machine interaction systems, allowing adaptive responses in applications like emotion-aware user interfaces, gaming, and neurofeedback-based interaction.

Traditional emotion recognition approaches, including facial expression analysis, gesture recognition, and speech processing, have demonstrated promise but face limitations. These methods are often hindered by individual variability in cultural background, age, language, and behavior. Furthermore, because they rely on observable external cues, such approaches are susceptible to contextual influences and may not accurately reflect an individual's true emotional state. In contrast, physiological signal-based emotion recognition offers a more objective and reliable avenue for assessing affective states [3].

Several studies have shown that EEG-based models consistently outperform traditional observable

How to cite this article
Olashore T D, Boyinbode O K, Daramola O A, et al. (September 26, 2025) Development of Real-Time Electroencephalography-Based Emotion Recognition Models Using Machine Learning Techniques. Cureus J Comput Sci 2 : es44389-025-06161-0. DOI https://doi.org/10.7759/s44389-025-06161-0

methods in emotion recognition tasks. For example, Han et al. [4] reported that EEG-based models achieved an average classification accuracy of up to 85% on the Database for Emotion Analysis using Physiological Signals (DEAP) dataset, compared to 65-70% for facial expression and audio-based modalities. Alotaibi and Fawad [5] further demonstrated superior F1-scores from EEG signals, particularly in noisy environments where facial or speech cues become unreliable. These findings highlight the robustness of EEG signals in capturing internal affective states with minimal susceptibility to external bias.

Among various physiological modalities, EEG has emerged as one of the most effective tools for emotion recognition. EEG signals are involuntarily generated, immune to voluntary manipulation, and directly reflect brain activity associated with emotional processing. This makes EEG particularly valuable in contexts where traditional methods fail; for instance, among individuals with physical disabilities who may not be able to express emotions through speech or gestures [3]. Recent advancements in deep learning have significantly improved the performance of EEG-based emotion recognition systems. However, model complexity, computational cost, and real-time applicability remain critical challenges. This section reviews several state-of-the-art approaches categorized by their architectural strategies and contributions.

One prominent approach is the development of end-to-end deep learning models tailored for EEG signals. For example, Han et al. [4] proposed E2ENNet, a lightweight end-to-end neural network that integrates depth-wise separable convolution for spatial feature extraction with a long short-term memory (LSTM) network to capture temporal dependencies. The model achieved high classification accuracies of 96.28% and 98.1% on the DEAP and DREAMER datasets, respectively (binary valence-arousal classification), and 41.73% on the seven-class MPED dataset. Despite its efficiency compared to traditional convolutional neural networks (CNNs), the model's training time (reported at 72 seconds) could still pose a challenge for real-time implementations.

Building upon the spatial-temporal analysis paradigm, Alotaibi and Fawad [5] introduced an AI-inspired spatio-temporal neural network known as the Bag-of-Hybrid-Deep-Features model. To address the high-dimensional feature space issue inherent in deep models, this architecture combined GoogLeNet deep features with handcrafted texture descriptors using a novel extraction method called OMTLBP_SMC. The authors applied classifiers such as support vector machine and K-nearest neighbors (KNN), achieving recognition accuracies of 93.83% (DEAP) and 96.95% (SEED). Although the performance was impressive, the model’s multi-step feature engineering and hybridization pipeline introduced significant complexity.

Other studies have focused on deep CNNs to enhance spatial and temporal learning from EEG signals. In Chen et al. [6], CNN models were trained to automatically learn emotional features across EEG dimensions. Compared with shallow classifiers, CNNs improved performance by 3.29% and 3.58% in arousal and valence classification, respectively. The model yielded the highest recognition accuracy when combining temporal and frequency features. However, it struggled to extract robust cross-subject discriminative features, emphasizing the need for more generalized EEG-based models.

In an effort to model multi-modal dependencies and improve temporal feature learning, Ma et al. [7] proposed a multimodal residual LSTM network. This architecture integrated both residual shortcut connections (spatial) and temporal LSTM paths, sharing temporal weights across modalities to capture interdependencies between EEG and other physiological signals. The model achieved classification accuracies of 92.87% (arousal) and 92.30% (valence). Nevertheless, its complex architecture made it less suitable for real-time deployment, raising concerns about scalability and computational demand.

In contrast to deep-only models, Zhong et al. [8] explored a hybrid deep architecture combining Hidden Markov Models with CNNs, utilizing multiscale sample entropy features from EEG data. This system targeted emotion recognition in human-computer interaction, attaining average accuracies of 79.77% (arousal), 83.09% (valence), and 81.83% (dominance). While the hybrid design improved accuracy, it introduced challenges related to processing speed and recognition latency, which are critical in real-time settings.

Several studies have focused on enhancing signal preprocessing and feature extraction through advanced decomposition techniques. Alhalaseh and Alasasfeh [9] proposed a system leveraging Variational Mode Decomposition and Empirical Mode Decomposition - techniques commonly used in biomedical signal processing. Feature extraction was performed using entropy measures and Higuchi’s fractal dimension, followed by classification using Naïve Bayes, CNN, KNN, and decision trees. The CNN-based model yielded the highest accuracy at 95.20%, but again, its computational complexity limited real-time feasibility. The key contribution of this study lies in its robust noise-reduction and decomposition framework, which enhanced signal quality for emotion recognition.

A more application-oriented approach was adopted by Hassouneh et al. [10], who developed a real-time emotion recognition system combining EEG signals and facial expressions. Aimed at classifying emotional states in children with autism and individuals with physical disabilities, the system used optical flow algorithms to process facial landmarks and applied CNN and LSTM classifiers for both facial and EEG data. While CNN achieved 99.81% accuracy on facial landmarks, LSTM reached 87.25% using EEG inputs. Despite its real-time focus, the EEG component fell short in precision, revealing the need for more accurate and efficient EEG-based emotion detection frameworks.

In response to these challenges, the present study seeks to develop statistically grounded yet computationally lightweight feature extraction techniques for EEG-based emotion recognition. Emphasis is placed on extracting time-domain statistical descriptors (e.g., mean, root mean square [RMS], peak-to-peak, entropy) that are easier to compute and interpret. Combined with efficient machine learning classifiers such as KNN, Light Gradient Boosting Machine (LightGBM), and Extreme Gradient Boosting (XGBoost), the aim is to construct an emotion recognition pipeline that ensures high classification performance, low-latency inference, and scalability for real-time applications. This study contributes to the field by aligning simplicity, accuracy, and efficiency to support applications such as mental health monitoring, adaptive user interfaces, and emotion-aware gaming systems.

Materials And Methods

The workflow was broken down into five phases, as shown in Figure *1*: data collection, data preprocessing, feature extraction, model training, and model evaluation. Each stage was created to ensure the precision and resilience of the emotion recognition system while preserving processing efficiency for real-time applications. This study developed a real-time EEG-based emotion recognition system that uses Russell’s two-dimensional valence-arousal scale to classify emotions. Time-domain and frequency-domain feature extraction methods were employed in conjunction with machine learning algorithms to improve the model's accuracy.

The dataset used in this study consists of EEG recordings obtained from 32 participants, each recorded using 32 scalp electrodes based on the international 10-20 system. These 32 EEG channels served as the physiological inputs for analyzing emotional responses. The data underwent a two-step processing pipeline to enhance signal quality and reliability. In the first stage, the raw EEG signals were subjected to bandpass filtering, downsampling, and artifact removal. In the second stage, the cleaned EEG signals were segmented into one-second non-overlapping epochs to ensure temporal consistency and facilitate efficient feature extraction.

Each one-second EEG segment was then processed to extract both time-domain and frequency-domain features, which were used as input for machine learning-based emotion classification. Emotions were not labeled directly as discrete categories but were inferred from participants’ self-reported ratings along the valence (pleasant-unpleasant) and arousal (active-passive) dimensions. These ratings were used to categorize emotional states into four quadrants: high valence-high arousal (e.g., excited), high valence-low arousal (e.g., relaxed), low valence-high arousal (e.g., anxious), and low valence-low arousal (e.g., sad). These discrete categories served as the target classes for training and evaluating the emotion recognition models.

Three machine learning algorithms - KNN and two gradient boosting techniques, LightGBM and XGBoost - were used to classify the gathered EEG features. Several common evaluation criteria were used in a thorough evaluation to gauge the trained model's performance. These included F1-score, Matthew's correlation coefficient (MCC), accuracy, precision, recall, and area under the curve (AUC).

The proposed system architecture for real-time EEG-based emotion recognition was designed to efficiently acquire, process, analyze, and classify EEG signals to detect emotional states. The architecture was structured into five main stages: data acquisition, data preprocessing, feature extraction, classification, and emotion recognition output, as shown in Figure *1* [11]. Each of these components played a critical role in the accurate and efficient recognition of emotional states based on EEG data.

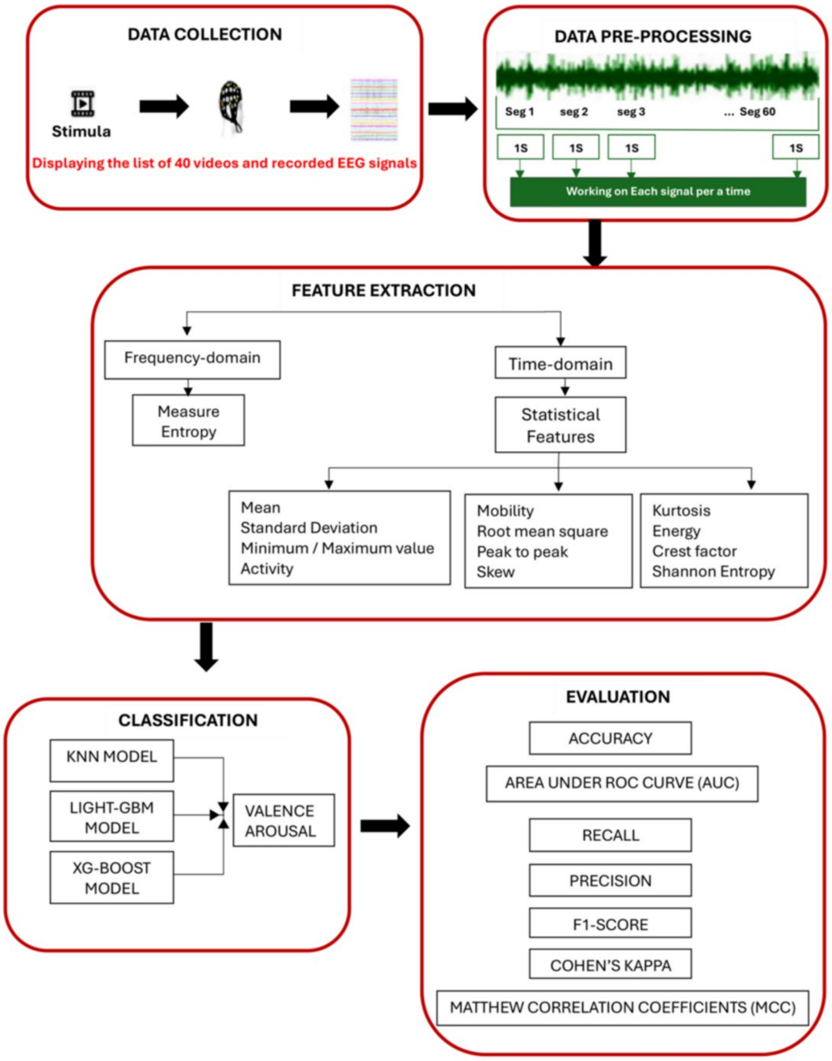

FIGURE 1: System Architecture

Data acquisition

The study employed the DEAP dataset [12], a widely used benchmark for emotion recognition from physiological signals, to analyze participants' emotional responses to audiovisual stimuli. The dataset comprises recordings from 32 participants, each of whom watched 40 one-minute-long music video clips designed to elicit a wide range of emotional responses. Data collection included both EEG and peripheral physiological signals such as electromyography, electrooculography, galvanic skin response, respiration, and blood volume pulse.

DEAP was selected over other publicly available emotion datasets, such as SEED and DREAMER, due to its multimodal nature, fine-grained emotion annotations, and wide emotional coverage using the valence-arousal scale. Unlike SEED, which focuses primarily on discrete emotions using visual stimuli alone, or DREAMER, which uses fewer participants and lacks peripheral signals, DEAP offers a more diverse and rich dataset with continuous ratings of valence, arousal, dominance, and liking. These characteristics make DEAP particularly suitable for building robust, generalizable emotion recognition systems that rely on both central and peripheral physiological responses.

For this research, the focus was placed on EEG signals due to their direct connection to brain activity and their suitability for affective state analysis. EEG signals were recorded using 32 electrodes positioned according to the International 10-20 system, with a sampling rate of 512 Hz. The channels analyzed included: Frontal regions: FP1, FP2, AF3, AF4, F3, F4, F7, F8, Fz; Central regions: FC1, FC2, FC5, FC6, C3, C4, Cz; Parietal regions: CP1, CP2, CP5, CP6, P3, P4, P7, P8, Pz; Temporal regions: T7, T8; Occipital and Parietal: O1, O2, Oz and Occipital regions: PO3, PO4, O1, O2, Oz.

These channels were categorized into five major regions: frontal, prefrontal, central, temporal, and occipital. The electrode placements used for signal acquisition are illustrated in Figure 2 [13], following the standard 10-20 international system. In addition to physiological recordings, the DEAP dataset includes participants' subjective emotional responses, collected through the Self-Assessment Manikin (SAM). SAM is a non-verbal pictorial assessment tool that captures individual ratings across three emotional dimensions: valence, arousal, and dominance.

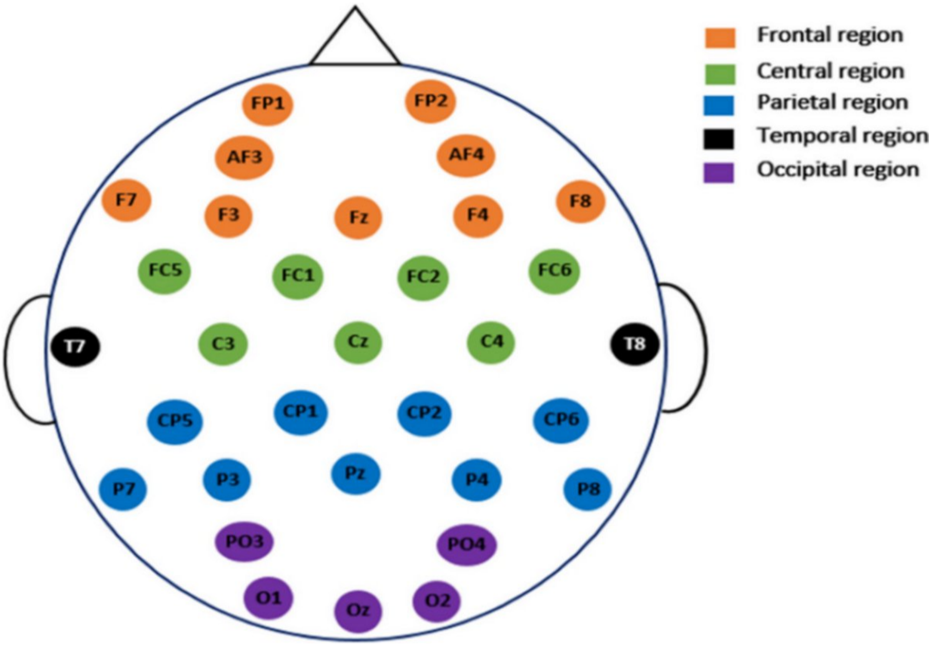

FIGURE 2: Layout of 32-Channel Electroencephalography Electrodes

Figure 3 shows a screenshot of the dataset used in this study. The dataset consists of 8 columns and 29,440 rows including the partcipants, video number, channel names, channel data, valence, arousal, dominance and liking rating.

	participant	video	channel no	channel data	valence	arousal	dominance	liking
0	s01.dat	1	Fp1	[0.948231680995192, 1.65333532651348, 3.013725...	7.71	7.60	6.90	7.83
1	s01.dat	1	AF3	[0.12470658952969527, 1.390082695001436, 1.835...	7.71	7.60	6.90	7.83
2	s01.dat	1	F3	[-2.2165109922713175, 2.292016820720129, 2.746...	7.71	7.60	6.90	7.83
3	s01.dat	1	F7	[1.005733531703222, 1.2979270362383235, 2.3676...	7.71	7.60	6.90	7.83
4	s01.dat	1	FC5	[5.095918660570228, 5.000869323443265, 4.17616...	7.71	7.60	6.90	7.83
...	...	...	...	...	...	...	...	...
29435	s24.dat	40	CP2	[30.26159680089255, 17.628873440414907, 3.5036...	5.00	2.97	4.99	4.04
29436	s24.dat	40	P4	[25.3160353460341, 16.030208513904697, 9.45232...	5.00	2.97	4.99	4.04
29437	s24.dat	40	P8	[15.616913361410415, 5.301099997194811, -7.034...	5.00	2.97	4.99	4.04
29438	s24.dat	40	PO4	[3.03038485339734, -15.74567625219008, -30.833...	5.00	2.97	4.99	4.04
29439	s24.dat	40	O2	[8.15624498353247, -7.627520674982776, -22.496...	5.00	2.97	4.99	4.04

29440 rows × 8 columns

FIGURE 3: Data Showing the Electroencephalography Channel Names, Channel Data, and Target Labels

Preprocessing

Raw EEG signals were subjected to a series of preprocessing steps aimed at improving signal quality and reducing computational complexity, thereby ensuring reliable feature extraction. First, the signals were downsampled from 512 Hz to 128 Hz, significantly reducing data volume while preserving the essential temporal dynamics relevant for emotion recognition. This was followed by bandpass filtering to eliminate baseline drift and high-frequency noise, thereby isolating the frequency components most relevant to affective processing.

Additionally, artifact removal techniques were applied to mitigate the impact of common EEG contaminants such as eye blinks, muscle movements, and electrical interference. In line with the DEAP dataset's preprocessing protocol [12], Independent Component Analysis was employed to identify and remove ocular artifacts, such as eye blinks and eye movements. The ICA components were manually inspected and those corresponding to non-neural activity were rejected, thereby improving signal clarity and ensuring the reliability of extracted features. These preprocessing steps collectively enhanced the clarity of the EEG signals, minimized noise, and produced clean input data for subsequent segmentation and feature extraction.

In addition to filtering and artifact removal, data segmentation was performed to facilitate the analysis of continuous EEG signals. As EEG is inherently a time-series signal capturing dynamic brain activity, segmentation into smaller time windows - referred to as epochs - is essential for enabling meaningful interpretation and efficient feature extraction.

To capture short-term fluctuations associated with transient emotional states, each 60-second EEG recording was divided into non-overlapping 1-second epochs, resulting in 60 segments per trial. This segmentation strategy offers a practical balance between temporal resolution and computational efficiency. Although sliding windows can increase the number of training samples and capture finer-grained transitions, they introduce temporal redundancy and greater computational overhead. Non-overlapping windows reduce this redundancy and preserve the independence of each epoch, which is advantageous when training machine learning models that assume uncorrelated input samples. Moreover, given that the DEAP dataset uses one-minute audiovisual stimuli designed to evoke relatively consistent emotional states, a 1-second non-overlapping window is sufficient to capture meaningful neural variations without sacrificing temporal context.

To enable the application of supervised learning algorithms, the continuous self-assessment ratings of valence and arousal provided in the DEAP dataset were discretized into binary class labels. A threshold-based encoding scheme was applied, where ratings greater than or equal to 5 were categorized as high valence or high arousal, while ratings below 5 were categorized as low valence or low arousal. This binarization effectively transformed the continuous affect modeling task into two distinct two-class classification problems, one for each emotional dimension. Although the DEAP dataset also includes dominance ratings, this study focuses exclusively on valence and arousal, in line with Russell's circumplex model of affect, which defines core emotional states using these two dimensions. Dominance was excluded due to its relatively lower relevance in EEG-based emotion recognition and the tendency of participant ratings to cluster around the mid-point, which may lead to class imbalance and reduced discriminative power.

Feature extraction

To enhance the model's ability to distinguish between different emotional states, both time-domain and frequency-domain features were extracted from each segmented EEG epoch. This dual-domain approach aimed to capture a wide range of signal characteristics, thereby improving the model's predictive performance. The feature extraction process involved transforming raw EEG time-series data into a structured set of parameters comprising both statistical descriptors (e.g., mean, standard deviation, skewness, and kurtosis) and signal-based descriptors (e.g., root mean square, peak-to-peak amplitude, Hjorth parameters, and spectral band power), which collectively encapsulate the temporal variability and spectral dynamics of emotional processing.

Time Domain Feature Extraction

In order to offer important insights into the temporal dynamics of brain activity, this study used a number of feature extraction criteria in time-domain. The activity, mobility, RMS, standard deviation, mean, peak-to-peak amplitude, skewness, kurtosis, and crest factor were among the statistical and mathematical characteristics that were included in these properties. Each of these parameters captured distinct properties of the EEG signal, offering a comprehensive representation of its temporal structure.

The mean and standard deviation provide measured of central tendency and dispersion respectively, while activity and mobility reflected the overall signal power and its rate of variation. Minimum and maximum values indicated the lowest and highest signal amplitudes, while energy quantified the total power of the signal. RMS served as an indicator of signal energy, whereas peak-to-peak amplitude quantified the range of fluctuations. Activity, defined as the signal's variance, measured power variability, whereas mobility tracked the variation of the signal over time to reflect frequency content. Complexity further analyzed frequency dynamics by quantifying how closely the signal's shape resembled that of its first derivative. Skewness and kurtosis characterized the asymmetry and peakedness of the signal distribution respectively, and the crest factor evaluated the ratio of peak amplitude to the RMS value, highlighting transient changes in the signal. Finally, Shannon entropy was included to measure the randomness or unpredictability present in the EEG signal.

In the time domain, a comprehensive set of features was derived to describe various aspects of the EEG signals. The mean represented the central tendency of the signals within each epoch, while the standard deviation measured the dispersion of amplitude values. Minimum and maximum values indicated the lowest and highest signal amplitudes, respectively. Energy quantified the total power of the signal, offering insight into the intensity and strength of neural activity. RMS assessed signal power by computing the square root of the mean of squared values. Activity, defined as the signal's variance, measured power variability, whereas mobility tracked the variation of the signal over time to reflect frequency content. Complexity further analyzed frequency dynamics by quantifying how closely the signal's shape resembled that of its first derivative. Peak-to-peak amplitude captured the highest range of signal fluctuation, and skewness assessed the asymmetry of the amplitude distribution. Kurtosis evaluated the tailedness of the signal's probability distribution, and crest factor was calculated as the ratio of the peak value to the RMS, indicating peak sharpness. Finally, Shannon entropy was included to measure the randomness or unpredictability present in the EEG signal.

1. Mean: The mean was computed to represent the central tendency of the signals within each epoch. Mean of the signal was calculated using the equation below.

$$\text{Mean} = \frac{1}{N}\sum_{i=1}^{N} x[i]$$

where N was the total number of samples and $x[i]$ was the signal's value at the i-th sample.

2. Standard deviation: Standard deviation measured the dispersion of the signal's amplitude values. Standard deviation was calculated using the equation below.

$$\text{Standard deviation} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x[i] - \text{Mean})^2}$$

where N was the total number of samples, Mean was the signal's mean, and $x[i]$ was the signal's value at the i-th sample.

3. Minimum and Maximum values: The EEG signal's minimum and maximum values, respectively, indicated its lowest and highest amplitudes for a specific epoch. The equations below were used to determine the signal's minimum and maximum values, respectively.

Minimum value = min($x[i]$, $x[ii]$, $x[iii]$, …, $x[N]$)

Maximum value = max($x[i]$, $x[ii]$, $x[iii]$, …, $x[N]$)

where N was the total number of samples.

4. Energy: The energy measured the total power of the signal within a given epoch, to understand the intensity and strength of neural activity. Energy of the signal was calculated using the equation below.

$$\text{Energy} = \sum_{i=1}^{N} (x[i])^2$$

where N represented the total number of samples and $x[i]$ represented the signal's value at the i-th sample.

5. RMS: The RMS value calculated the signal's power by taking the square root of the mean of squared data. The equation below was used to calculate the signal's RMS value.

$$\text{Root mean square} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x[i])^2}$$

where N was the total number of samples and $x[i]$ was the signal's value at the i-th sample.

6. Activity: The activity computed the variance to quantify signal power. The equation below was used to calculate the signal's activity

$$\text{Activity (Variance)} = \frac{1}{N}\sum_{i=1}^{N}(x[i] - \text{Mean})^2$$

where N was the total number of samples, Mean was the signal's mean, and $x[i]$ was the signal's value at the i-th sample.

7. Mobility: Mobility quantified the signal's frequency content by tracking its variation over time. It was defined as the ratio between the original signal's standard deviation and the standard deviation of the signal's first derivative. The equation below was used to determine the signal's mobility.

$$\text{Mobility} = \sqrt{\frac{Var\left(\frac{dx}{dt}\right)}{Var(x(t))}}$$

where *Var* is the signal's variance and $\dfrac{dx}{dt}$ is the signal's first derivative with respect to time.

8. Complexity: Complexity quantified the degree to which the EEG signal's form resembled that of its first derivative. It is the first derivative's mobility divided by the original signal's mobility. The equation below was used to calculate the signal's complexity.

$$\text{Complexity} = \frac{M_{\frac{dx}{dt}}}{M_x}$$

where M_x denoted the mobility of the original signal and $M_{\frac{dx}{dt}}$ denoted the mobility of the signal's first derivative.

9. Peak-to-Peak Value: The peak-to-peak value measured the highest range of signal variation. It calculated the difference between each epoch's maximum and minimum signal amplitudes. The equation below was used to determine the signal's peak-to-peak value.

$$\text{Peak-to-peak} = \max(x[i]) - \min(x[i])$$

where $\max(x[i])$ denoted the signal's maximum amplitude value at the i-th sample and $\min(x[i])$ denoted its minimum amplitude value at the i-th sample.

10. Skewness: Skewness quantified how asymmetrical the EEG signal values' probability distribution was. It helped to analyze whether the EEG amplitude distribution was more concentrated on one side of the mean. Skewness of the signal was calculated using the equation below.

$$\text{Skewness} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\sigma^3}$$

where N stood for the total number of samples, μ for the signal mean, σ for the standard deviation, and $x[i]$ for the signal's value at the i-th sample.

11. Kurtosis: Kurtosis measured the tailedness of the EEG signal's amplitude distribution. It quantifies whether the EEG values have more or fewer extreme deviations compared to a normal distribution. Kurtosis of the signal was calculated using the equation below.

$$\text{Kurtosis} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4}$$

where N stood for the total number of samples, μ for the signal mean, σ for the standard deviation, and $x[i]$ for the signal's value at the i-th sample.

12. Crest Factor: The crest factor calculated the ratio between the RMS value and the peak value. It was used to highlight the sharpness of signal peaks. Crest Factor of the signal was calculated using the equation below.

$$\text{Crest Factor} = \frac{\max|x(t)|}{\text{RMS}\,|x(t)|}$$

where RMS stood for the signal's root mean square and max|$x(t)$| for the signal's maximum absolute value.

13. Shannon Entropy: The Shannon entropy measured the randomness or unpredictability in the signal's data. The EEG signal values were first converted into a probability distribution by binning the amplitude values. The probability p_i was the percentage of total values that fell into the i-th bin, and each bin indicated a range of signal amplitudes. The equation below was used to calculate the signal's entropy.

$$\text{Shannon Entropy} = -\sum_{i=1}^{N} p_i log2_{pi}$$

where p_i was the likelihood that the i-th amplitude value in the signal would occur, N was the number of distinct amplitude levels (or bins) in the signal segment, and $log2p_i$ was the logarithm (base 2) of the probability.

Figures 4 and 5 displays a snapshot of the time domain properties of each 1-second epoch of the EEG signal data.

	channel	valence_label	arousal_label	dominance_label	mean 0	std_dev 0	min_val 0	max_val 0	activity 0	mobility 0	...	activity 59	mobility 59
0	FP1	1	1	1	-0.146521	3.993753	-12.469560	9.900878	15.950060	0.698273	...	9.470493	0.920991
1	AF3	1	1	1	-0.073709	4.270408	-13.270521	9.878666	18.236382	0.664475	...	11.362700	0.837165
2	F3	1	1	1	-0.095725	4.534111	-11.572778	9.214996	20.558162	0.676792	...	11.682807	0.793693
3	F7	1	1	1	-0.330268	4.056349	-9.898021	10.248160	16.453966	0.711457	...	12.548717	0.786450
4	FC5	1	1	1	-0.223199	2.940199	-7.231208	8.011139	8.644770	0.709156	...	6.871972	0.903855
...	...	...	...	...	...	...	...	...	...	...	...	...	...
40955	CP2	1	0	1	0.685193	14.735845	-33.648205	48.215899	217.145140	0.547181	...	166.089070	0.696379
40956	P4	1	0	1	0.239408	6.784246	-17.946195	17.877497	46.025997	0.744204	...	28.534151	1.064057
40957	P8	1	0	1	0.005869	7.237277	-20.838630	18.055395	52.378174	0.837767	...	26.962381	1.049130
40958	PO4	1	0	1	0.975887	27.011600	-69.248418	93.152754	729.626520	0.504537	...	619.249522	0.656369
40959	O2	1	0	1	0.089327	5.933386	-13.474728	14.741871	35.205074	0.714649	...	43.465359	0.744996

40960 rows × 844 columns

FIGURE 4: Time-Based Feature Extraction

activity 59	mobility 59	complexity 59	shannon_entropy 59	root_mean_square 59	peak_to_peak 59	skew 59	kurtosis 59	energy 59	crest_factor 59
9.470493	0.920991	1.313763	2.616749	3.077417	13.577976	0.241071	2.514778	1212.223318	4.412134
11.362700	0.837165	1.312286	2.266763	3.371005	16.498194	0.288905	2.872834	1454.550153	4.894147
11.682807	0.793693	1.396533	2.256643	3.418031	16.676408	0.173676	2.646491	1495.415571	4.878952
12.548717	0.786450	1.446276	2.249472	3.542451	16.925197	-0.089695	2.408587	1606.266543	4.777821
6.871972	0.903855	1.456322	2.484452	2.621965	13.731931	0.146826	2.795397	879.961870	5.237266
...	...	...	...	...	...	...	...	...	...
166.089070	0.696379	1.817662	0.779484	12.894288	71.737214	-0.055921	3.985619	21281.619911	5.563488
28.534151	1.064057	1.338055	1.631304	5.346712	26.674009	0.313707	2.960379	3659.178285	4.988862
26.962381	1.049130	1.365127	1.760450	5.199483	24.779377	0.104831	2.373735	3460.432286	4.765738
619.249522	0.656369	1.894523	0.456846	24.893111	142.665929	-0.063077	4.201861	79317.372226	5.731141
43.465359	0.744996	1.699828	1.419096	6.592965	32.971563	0.547923	3.170976	5563.800388	5.001022

FIGURE 5: Time-Domain Feature Extraction II

Frequency Domain Analysis of EEG Signals

In this research, the EEG signals' frequency-domain characteristic, spectral entropy, was extracted. It was calculated in order to measure the EEG power spectrum's unpredictability. After estimating the power spectral density (PSD) using Welch's approach, the normalized PSD's entropy was computed. Each EEG recording was divided into 1-second epochs, and spectral entropy was calculated for each epoch using the Welch method for power spectral density estimation. The equation below displays the mathematical formula for calculating spectral entropy.

$$\text{Spectral entropy} = -\sum_{i=1}^{N} P(f_i) log_2 p_{fi}$$

where $P(f_i)$ represented the normalized PSD at frequency f_i and N was the number of frequency bins. While lower entropy levels revealed dominating frequency components, usually seen during calm or unconscious states, higher entropy values indicated wider spectrum dispersion, which is linked to alertness and cognitive involvement. To train machine learning models to categorize emotional states, the retrieved spectral entropy characteristics were utilized as input features. Figures 6 displays a snapshot of the frequency domain properties of each 1-second epoch of the EEG signal data.

abel	emtropy 0	emtropy 1	emtropy 2	emtropy 3	emtropy 4	emtropy 5	...	emtropy 50	emtropy 51	emtropy 52	emtropy 53	emtropy 54	emtropy 55	emtropy 56	emtropy 57	emtropy 58	emtropy 59
1	4.253388	4.184880	4.102895	4.400221	4.239568	4.016697	...	3.845704	3.925847	4.241608	4.218194	4.583336	4.120355	4.126251	4.025469	4.305704	4.416862
1	4.144137	4.265561	4.243319	4.199149	4.031556	4.169508	...	3.901461	3.677438	4.320534	4.022160	4.450974	4.126974	4.117979	4.402833	4.242477	4.333276
1	4.232622	4.205201	4.221551	4.219390	3.950648	4.480395	...	3.946106	3.799766	4.275722	3.898344	4.541193	4.237118	4.187895	4.570081	4.464024	4.301240
1	3.984093	4.114149	3.742872	4.264199	4.072495	4.299512	...	4.151636	4.336159	3.999319	4.423865	4.679060	4.049580	4.536614	4.276591	4.572580	4.191792
1	4.032882	3.916700	4.357818	3.834407	3.671401	4.469807	...	4.629931	4.151270	4.661906	4.702124	4.781600	4.379865	4.985091	4.325961	4.512291	4.508690
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1	3.776813	3.160940	2.357512	4.703229	3.797712	4.007635	...	4.594736	4.753255	4.531559	4.712242	3.968378	4.227432	4.456448	4.848147	4.696950	4.202124
1	3.824490	2.566792	2.455398	3.959559	4.182803	4.807766	...	4.469614	4.790995	4.808144	4.884467	4.327647	4.857619	3.754829	3.913131	4.877455	4.850797
1	4.464546	3.082185	2.433856	4.748195	4.860288	4.392385	...	4.879660	4.515216	4.692838	4.719798	4.275055	4.775704	4.122075	4.026127	4.335378	4.784591
1	3.635052	3.261251	2.287884	4.721355	3.605886	3.154811	...	4.692974	4.653973	4.791174	4.575068	3.690390	3.955771	4.666577	4.406934	4.691203	4.214819
1	4.115128	3.399567	2.417111	4.542942	4.456361	4.448363	...	4.344070	4.631446	4.469778	4.614197	3.717564	3.678037	4.423332	4.803598	4.458034	3.946060

FIGURE 6: Frequency-Based Feature Extraction

To ensure that the model was trained on the most informative and non-redundant features, a feature selection strategy was also employed. Dimensionality reduction techniques - such as Principal Component Analysis - and feature importance rankings derived from ensemble tree-based models (e.g., XGBoost) were explored to identify and retain only the most discriminative features. This step reduced computational overhead and minimized overfitting while preserving the features most essential to accurately classifying emotional states. The retrieved characteristics were then used as inputs to supervised machine learning models tasked with categorizing emotional states based on the valence-arousal framework.

Classification

Following feature extraction, machine learning classifiers were trained on the derived feature set to recognize emotional states from EEG signals. Three supervised learning algorithms were employed for classification: KNN, LightGBM, and XGBoost. Each model was selected for its strengths in handling structured numerical data and was optimized through cross-validation.

K-Nearest Neighbours

KNN algorithm was implemented to classify emotional states based on the proximity of a sample's features to labeled examples in the feature space. It is a non-parametric, instance-based learning technique that assigns a class label to an input based on the majority label of its k nearest neighbors. To ensure consistency in the feature space, all input features were first scaled using Min-Max normalization, which transforms values to a (0, 1) range, thereby preventing features with larger scales from dominating the distance computation.

The classification is based on the Euclidean distance between feature vectors. For a given test sample x, its distance to each training sample x_i in an n-dimensional space is calculated as:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{ij})^2}$$

where:

$x = [x_1, x_2, ..., x_n]$ is the feature vector of the test instance.

$x_i = [x_{i1}, x_{i2}, ..., x_{in}]$ is a feature vector from the training set.

$d(x, x_i)$ is the Euclidean distance between x and x_i.

While Euclidean distance was adopted as the default metric, other distance measures such as Manhattan was also considered to assess the robustness of the classifier's performance.

The KNN classifier assigns to x the majority class y among its k closest training instances:

$$\widehat{y} = mode(\{y_i | x_i \epsilon N_k(x)\})$$

where $N_k(x)$ denotes the set of KNN to x, and y_i is the class label of neighbor x_i.

In this study, grid search was used to optimize the value of k, testing values from 1 to 5. The value $k = 3$ was selected for its optimal trade-off between bias and variance, ensuring balanced classification performance. The model was trained separately for valence and arousal dimensions using binary labels (i.e., high/low) obtained from the DEAP dataset. To ensure robustness and generalization, 10-fold cross-validation was applied during model training. This helped mitigate overfitting and provided a reliable estimate of the model's performance on unseen data.

Light Gradient Boosting Machine

LightGBM is a highly efficient implementation of gradient boosting decision trees, designed for speed and memory optimization on large-scale, high-dimensional datasets. Unlike traditional level-wise tree construction, LightGBM grows trees leaf-wise, selecting the leaf with the highest gain. This results in faster convergence and deeper, more accurate trees, albeit with a higher risk of overfitting - which is controlled using built-in regularization mechanisms.

LightGBM employs advanced techniques such as Gradient-based One-Side Sampling and Exclusive Feature Bundling to reduce computational complexity without compromising predictive performance. These innovations make it particularly well-suited for real-time EEG-based emotion recognition tasks, where rapid inference and low latency are critical.

In this study, LightGBM automatically selected informative features from the extracted EEG feature vectors and built an ensemble of weak learners (decision trees), each trained to minimize the residual error of the preceding ensemble. The objective function minimized by the model consists of two components: a loss function that quantifies prediction error and a regularization term that penalizes model complexity:

$$L = \sum_{i=1}^{n} l(y_i, y_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k)$$

where y_i is the true label, $y_i^{(t)}$ is the prediction after t iterations, and f_k is the k-th decision tree. The regularization term $\Omega(f_k)$ is defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

where T is the number of leaves, w_j are the leaf weights, and γ and λ are regularization parameters.

To ensure fair and unbiased performance comparison with other classifiers, systematic hyperparameter tuning was conducted. Parameters such as the learning rate, number of estimators, maximum tree depth, feature fraction, and minimum data in leaf were optimized using grid search combined with five-fold cross-validation for initial exploration. Final model evaluation was performed using 10-fold cross-validation to confirm robustness and generalization performance. LightGBM was trained independently for both arousal and valence classification tasks.

Its scalability, embedded regularization, and capacity to handle imbalanced and high-dimensional EEG data confirmed LightGBM as a highly competitive model for real-time EEG-based emotion recognition.

Extreme Gradient Boost

XGBoost is a powerful and scalable ensemble learning algorithm based on gradient boosting frameworks. It constructs an ensemble of decision trees where each new tree is trained to correct the residual errors of the previous ones. Known for its efficiency, regularization capability, and scalability, XGBoost has become a popular choice for high-dimensional and structured data, making it particularly well-suited for EEG-based emotion recognition tasks.

In this study, XGBoost was employed to classify emotional states using features extracted from segmented EEG epochs. The model was configured with a maximum tree depth of 7 to capture complex feature interactions without excessive overfitting. The log loss (logarithmic loss) function was adopted as the objective function for binary classification tasks, such as distinguishing between high and low arousal or valence. The objective function can be represented as:

$$L = -\frac{1}{n}\sum_{i=1}^{n}[y_i log(\widehat{y_i}) + (1-y_i)log(1-\widehat{y_i})]$$

where:

$y_i \epsilon \left\{0, 1\right\}$ is the true binary label for the i-th instance.

$\widehat{y_i}$ is the predicted probability output by the model.

A logistic sigmoid function was used as the activation function to map the raw output of each tree into a probability for binary classification. Additionally, a carefully tuned learning rate was applied to ensure a balance between training speed and convergence stability, promoting gradual updates that minimize overfitting.

XGBoost's performance was further enhanced through L1 (Lasso) and L2 (Ridge) regularization, which helped control model complexity by penalizing large leaf weights. These regularization terms, integrated into the objective function, contributed to robust generalization on unseen data. Moreover, XGBoost employs techniques such as column subsampling, parallel tree construction, and cache-aware data structures to reduce training time and computational overhead.

In the context of EEG emotion classification, XGBoost's ability to handle noisy and high-dimensional inputs, alongside its proficiency in modeling non-linear relationships, enabled it to effectively capture subtle patterns in the temporal and spectral EEG features. Hyperparameters were optimized through grid search, and 10-fold cross-validation was used to validate model generalizability and mitigate overfitting.

Evaluation metrics

To comprehensively assess the performance of the classification models used for EEG-based emotion recognition, multiple evaluation metrics were adopted. These metrics account for different aspects of model behavior, particularly under conditions of class imbalance or probabilistic output. Below are the definitions and mathematical formulations of each:

i. Accuracy: Accuracy represents the proportion of correctly classified instances (both positive and negative) out of the total instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where: TP: True Positives; TN: True Negatives; FP: False Positives; FN: False Negatives.

ii. Area Under the Receiver Operating Characteristic Curve (AUC-ROC): AUC evaluates the ability of the model to distinguish between the positive and negative classes. It is computed as the area under the receiver operating characteristic (ROC) curve. There is no simple closed-form formula for AUC, but it can be estimated as:

$$\text{AUC} = \int_0^1 TPR\big(FPR^{-1}(x)\big)\,dx$$

Where: TPR: True Positive Rate; FPR: False Positive Rate.

iii. Recall (Sensitivity or True Positive Rate): Recall measures the model's capacity to identify true positives by showing the percentage of real positive cases that it properly detects.

$$\text{Recall} = \frac{TP}{TP + FN}$$

iv. Precision: Precision measures the percentage of expected positive cases that turn out to be positive, demonstrating the model's dependability in classifying data as positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

v. F1 Score: The F1 Score is the harmonic mean of precision and recall, balancing the two in a single metric.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

vi. Cohen's Kappa: Cohen's Kappa measures the degree of agreement between the actual and predicted classifications while taking the likelihood that the agreement will occur by chance into account.

$$k = \frac{P_o - P_e}{1 - P_e}$$

where P_o is the observed agreement and P_e is the expected agreement by chance.

vii. MCC: MCC measures true positives, true negatives, false positives, and false negatives to offer a more balanced assessment than accuracy alone.

$$\text{MCC} = \frac{TP.TN - FP.FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives respectively. MCC values range from -1 (total disagreement between prediction and observation) to +1 (perfect prediction), with 0 indicating random performance.

viii. Confusion Matrix: A confusion matrix is a tabular visualization of the actual versus predicted classifications made by a model. It allows detailed analysis of model performance for each class and helps identify specific areas of misclassification. Each row of the matrix represents the instances in an actual class, while each column represents the instances in a predicted class.

Results

Experimental setup

All experiments were conducted on an HP Pavilion x360 equipped with an Intel Core i5-10210U CPU, 16 GB of RAM, and a 1 TB SSD, running Windows 11. The implementation was carried out in Python 3.13 using several scientific libraries, including Scikit-learn, XGBoost, and LightGBM for classification, along with NumPy, Pandas, and MNE for signal processing and feature extraction.

The study utilized the DEAP dataset, comprising multichannel EEG recordings collected from participants during emotion-eliciting tasks. The EEG signals were sampled at 512 Hz; later down-sampled to 128 Hz and recorded from 32 electrodes based on the international 10-20 system. To improve signal quality, a bandpass filter was applied in the range of 0.5-45 Hz to remove baseline drift and high-frequency noise.

EEG data were segmented into fixed-length windows of 1 second. Each segment was normalized using min-max normalization. Feature extraction focused on statistical and mathematical descriptors from the time and frequency domains, including mean, standard deviation, RMS, peak-to-peak amplitude, skewness, kurtosis, activity, mobility, and crest factor.

For classification, three machine learning algorithms were employed:

1. KNN with k = 3.

2. LightGBM with default boosting parameters.

3. XGBoost with a learning rate of 0.1 and 100 estimators.

Model evaluation was performed using 10-fold cross-validation to ensure generalizability and a thorough and objective assessment of the predictive model. This method reduced the possibility of overfitting, offered a more thorough and comprehensive performance evaluation, and maximized data utilization. Performance metrics included accuracy, precision, recall, F1-score, MCC, and AUC.

Valence classification performance using KNN

For the valence dimension, as illustrated in Figure *7*, the KNN model demonstrated outstanding classification performance, achieving an average accuracy of 96.77%. This high accuracy reflects the model's robustness in distinguishing between low- and high-valence emotional states derived from EEG signals. The corresponding AUC score of 96.68% further underscores the model's strong discriminative capability.

Additionally, the precision and recall scores - both at 96.77% - indicate that the model maintained a well-balanced ability to identify true positives without compromising specificity. The F1-score, also at 96.77%, confirms the consistency of the model across both precision and recall dimensions. Importantly, Cohen's Kappa and the MCC - each reaching 93.42% - suggest strong agreement beyond chance and robust predictive validity.

To ensure the integrity of the evaluation process and prevent data leakage, a 10-fold subject-independent cross-validation strategy was employed. In this configuration, EEG samples from each subject were strictly confined to either the training or testing set within each fold. This subject-wise split ensured that the classifier's performance reflected its generalization ability to unseen individuals, which is critical for real-world deployment.

The high values of Cohen's Kappa and MCC carry significant practical implications. In clinical settings, where accurate emotional assessment can support mental health diagnosis or therapeutic interventions, strong agreement metrics ensure dependable outcomes even in ambiguous cases. Similarly, in adaptive user environments such as affective gaming, personalized learning, or human-computer interaction systems, this level of predictive reliability supports stable and context-aware system responses, enhancing user experience and system trustworthiness.

Figure *7* provides a visual summary of these evaluation metrics, illustrating both central performance tendencies and variance across subsets of the valence-labeled dataset.

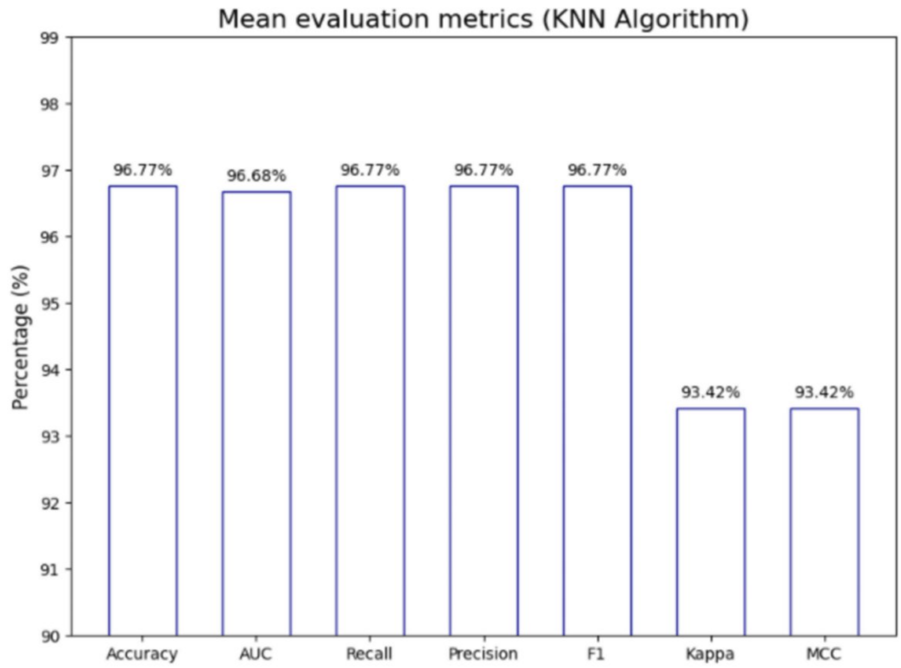

FIGURE 7: Graphical Illustration of KNN Evaluation Metrics (Valence)

KNN, K-Nearest Neighbor

Arousal classification performance using KNN

For the arousal dimension, as shown in Figure 8, the KNN model achieved an average accuracy of 96.68%, indicating its strong capability in classifying low- and high-arousal emotional states derived from EEG signals. The model also recorded an impressive AUC score of 96.49%, highlighting its excellent discriminative power between arousal classes.

Both the precision and recall scores reached 96.68%, confirming the model's reliability in correctly identifying high-arousal instances without introducing class imbalance bias. The corresponding F1-score of 96.67% reinforces this balance, demonstrating the classifier's stable performance across sensitivity and precision metrics. Additionally, the agreement metrics - Cohen's Kappa at 93.12% and MCC at 93.13% - indicate substantial inter-class agreement and predictive validity across multiple folds of evaluation.

To safeguard against overfitting and ensure fair evaluation, a 10-fold subject-independent cross-validation scheme was applied. This setup ensured that data from each subject appeared in only one split-either training or testing-thus reflecting the model's capacity to generalize to previously unseen individuals, a critical requirement for practical emotion-aware systems.

The high reliability scores carry meaningful implications for real-world applications. In clinical neurofeedback systems, where monitoring arousal levels is essential for stress or anxiety management, such robust agreement ensures consistent performance over diverse patient data. Similarly, in adaptive human-computer interfaces or brain-computer interaction systems, dependable arousal classification can enable dynamic adjustments to user engagement levels, thereby improving system responsiveness and user satisfaction.

Figure 8 provides a visual summary of these evaluation metrics, illustrating both central performance tendencies and variance across subsets of the arousal-labeled dataset.

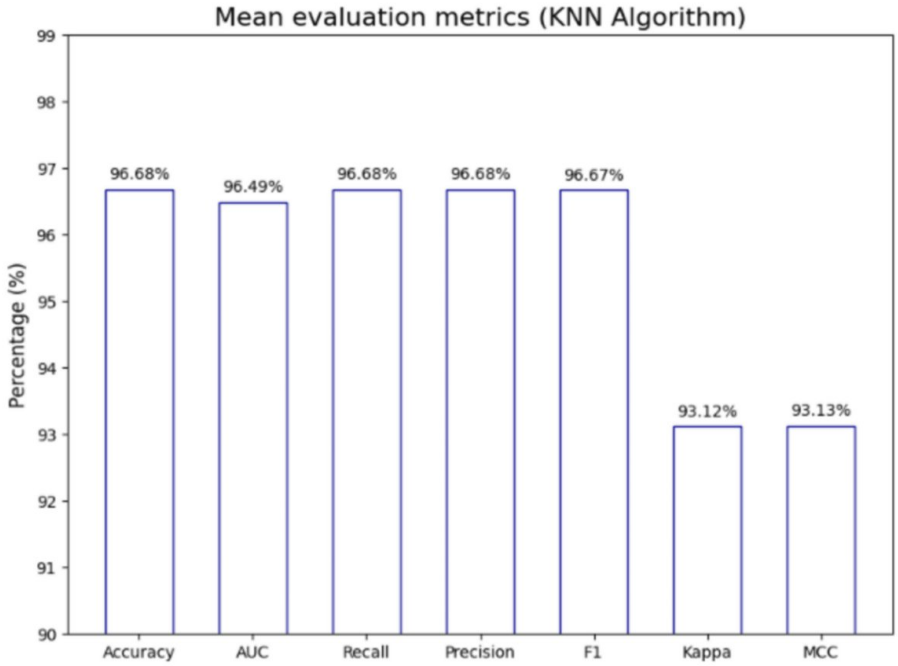

FIGURE 8: Graphical Illustration of KNN Evaluation Metrics (Arousal)

KNN, K-Nearest Neighbor

Valence classification performance using LightGBM

For the valence dimension, as shown in Figure *9*, the LightGBM model achieved an average accuracy of 84.08%, demonstrating a strong ability to distinguish between low- and high-valence emotional states from EEG signals. The model also recorded an AUC score of 82.71%, indicating solid discriminative capacity between the two emotional classes.

The recall and precision scores - 84.08% and 84.71%, respectively - highlight the model's competence in identifying true positive cases while maintaining a relatively low false positive rate. The corresponding F1-score of 83.79% confirms that the model maintained a balanced trade-off between sensitivity and precision across validation folds.

Additionally, agreement metrics such as Cohen's Kappa (66.87%) and MCC (67.87%) provide insight into the model's moderate to strong reliability beyond chance-level agreement. Although these values are lower than those obtained using KNN, they still reflect meaningful predictive power-particularly given the complexity and variability inherent in EEG-based valence detection.

As with other models in this study, LightGBM was evaluated using a 10-fold subject-independent cross-validation strategy. This approach ensured that data from individual participants were not shared between training and testing splits, thus preserving the model's generalization capacity to unseen subjects.

From a practical standpoint, these results suggest that LightGBM may be well-suited for non-critical or semi-real-time applications such as affective feedback in intelligent tutoring systems or adaptive multimedia environments, where moderate emotional accuracy can still provide meaningful user personalization without requiring clinical-grade precision.

Figure *9* provides a visual summary of the model's average evaluation metrics, capturing performance variability across different subsets of the valence-labeled EEG dataset.

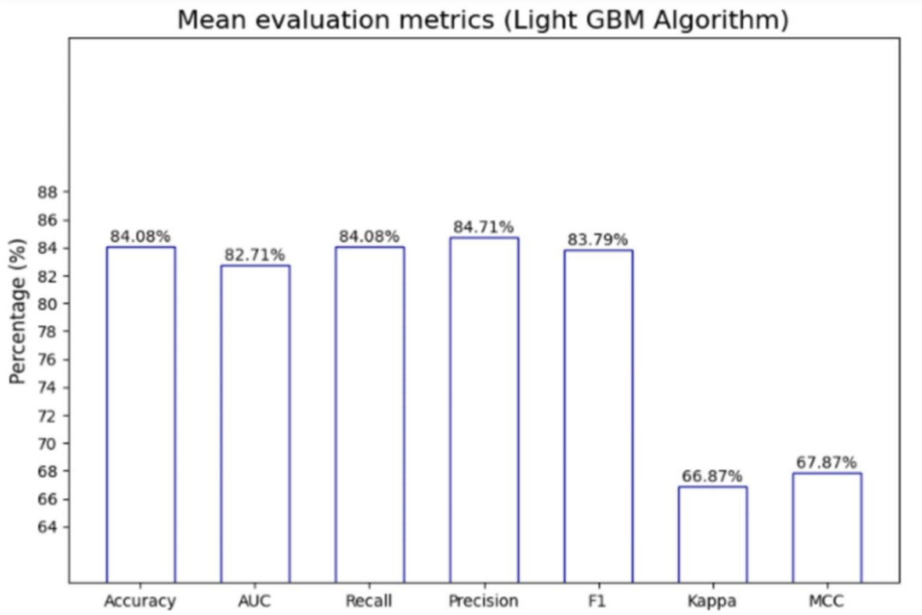

FIGURE 9: Graphical Illustration of LightGBM Evaluation Metrics (Valence)

Arousal classification performance using LightGBM

For the arousal dimension, as illustrated in Figure *10*, the LightGBM model achieved an average accuracy of 83.34%, demonstrating strong overall performance in classifying low- and high-arousal emotional states from EEG signals. The model also attained an AUC score of 81.10%, indicating a commendable capacity to differentiate between arousal levels.

The recall and precision values - 83.34% and 84.02%, respectively - highlight the model's reliability in correctly detecting positive arousal instances while maintaining a low rate of false positives. The resulting F1-score of 82.89% confirms a stable and well-balanced trade-off between sensitivity and specificity. Furthermore, agreement metrics such as Cohen's Kappa (64.38%) and MCC (65.67%) reflect a moderate level of inter-class agreement and predictive reliability across validation folds.

To ensure the robustness of evaluation and to prevent data leakage, a 10-fold subject-independent cross-validation procedure was employed. This strategy guaranteed that EEG recordings from the same subject were never present in both training and testing sets, thereby simulating real-world generalization to new, unseen users.

Although the agreement metrics are lower compared to KNN, they remain acceptable for contexts where moderate emotional state classification is sufficient. For instance, LightGBM's performance is appropriate for adaptive systems such as interactive learning platforms, biofeedback applications, or personalized media, where responsiveness to arousal fluctuations can enhance engagement without necessitating clinical precision.

Figure *10* presents a graphical summary of the model's average evaluation metrics, capturing variability across different data subsets labeled for arousal.

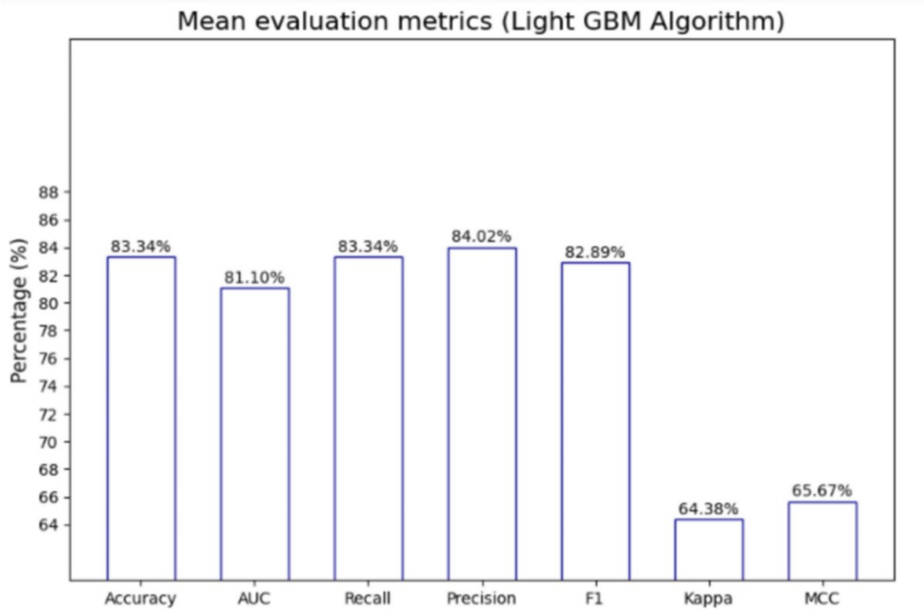

FIGURE 10: Graphical Illustration of LightGBM Evaluation Metrics (Arousal)

Valence classification performance using XGBoost

For the valence dimension, as presented in Figure *11*, the XGBoost model achieved an average accuracy of 88.55%, reflecting strong performance in distinguishing between low- and high-valence emotional states based on EEG signals. The model also recorded a high AUC of 87.80%, indicating excellent discriminatory capability between emotional classes.

The recall and precision values - 88.55% and 88.70%, respectively - demonstrate the model's consistency and reliability in detecting positive valence states while minimizing misclassification. The F1-score of 88.47% further confirms the model's balanced performance across sensitivity and specificity dimensions.

Moreover, agreement metrics such as Cohen's Kappa (76.44%) and MCC (76.72%) reflect substantial agreement and predictive robustness beyond chance levels, indicating that the model maintained reliable classification performance across folds.

To ensure fairness in evaluation and guard against data leakage, a 10-fold subject-independent cross-validation procedure was implemented. This ensured that EEG data from the same individual were not included in both training and test sets, thereby validating the model's ability to generalize to new, unseen subjects.

Given its superior agreement metrics and high overall accuracy, XGBoost presents a reliable and generalizable solution for valence classification. This makes it particularly suitable for deployment in emotion-aware systems such as intelligent user interfaces, real-time affective computing applications, or mental health monitoring tools where consistent classification of emotional valence is critical.

Figure *11* presents a graphical summary of these evaluation metrics, illustrating performance consistency across different subsets of the valence-labeled EEG dataset.

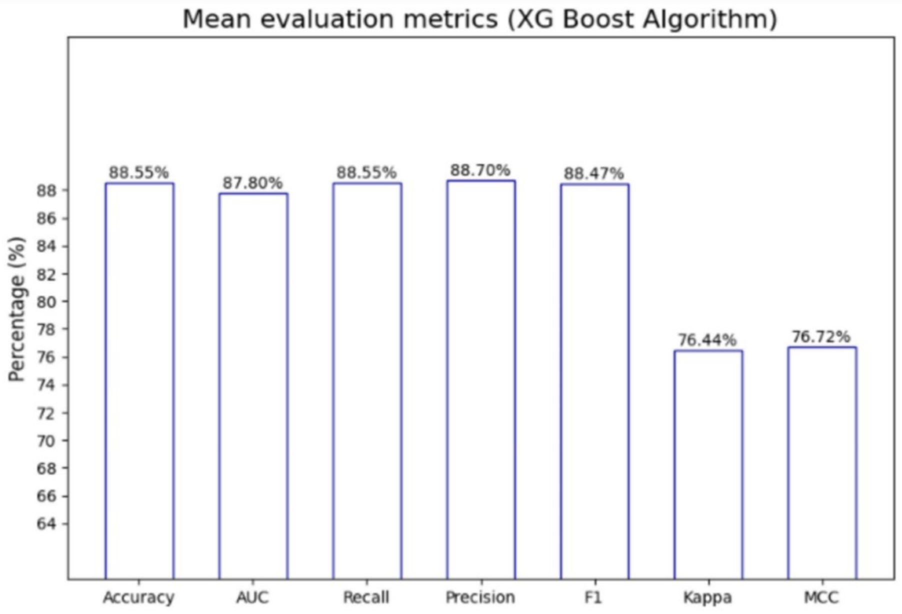

FIGURE 11: Graphical Illustration of XGBoost Model Evaluation Metrics (Valence)

Arousal classification performance using XGBoost

For the arousal dimension, as illustrated in Figure *12*, the XGBoost model achieved an average accuracy of 87.54%, demonstrating strong effectiveness in distinguishing between low- and high-arousal emotional states from EEG data. The corresponding AUC score of 86.21% further confirms the model's robust discriminative capability across the binary arousal classes.

With recall and precision values of 87.54% and 87.72%, respectively, the model consistently identified positive arousal instances while maintaining a low false-positive rate. The resulting F1-score of 87.39% reflects a well-balanced trade-off between sensitivity and specificity.

In terms of reliability, the model exhibited substantial agreement and predictive validity, as evidenced by a Cohen's Kappa of 73.77% and an MCC of 74.20%. These metrics suggest that the XGBoost classifier maintained reliable and generalizable performance across multiple folds.

To ensure rigorous evaluation, a 10-fold subject-independent cross-validation strategy was employed. This approach prevented data leakage by ensuring that EEG recordings from the same subject were not used simultaneously for training and testing, thus supporting the model's generalization capability to unseen individuals.

Given its strong performance across all evaluation metrics, XGBoost demonstrates considerable potential for real-world applications such as affective brain-computer interfaces, stress monitoring systems, and other adaptive user-state modeling environments where accurate detection of arousal levels is essential.

Figure *12* provides a graphical summary of these averaged performance metrics, depicting the model's effectiveness across the arousal-labeled EEG dataset.

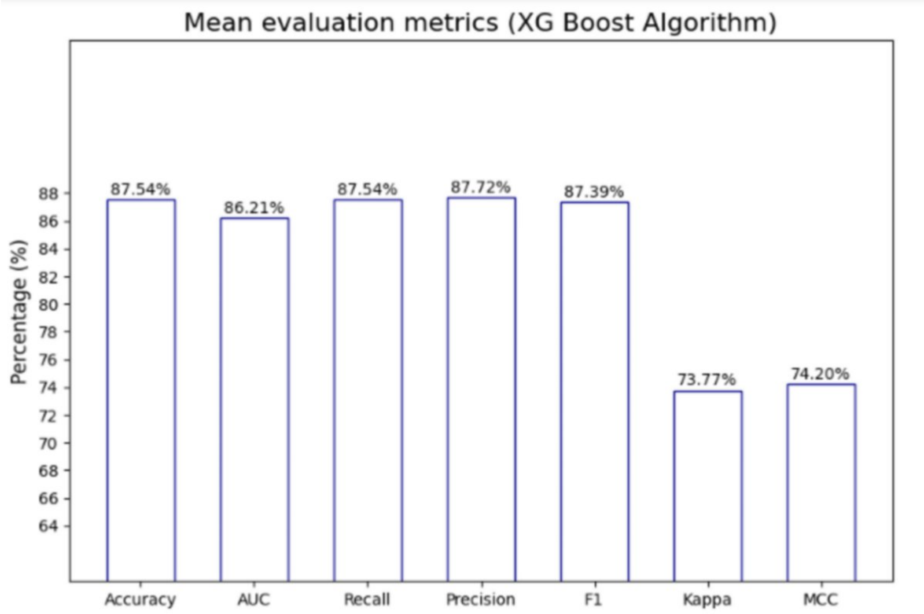

FIGURE 12: Graphical Illustration of XGBoost Model Evaluation Metrics (Arousal)

Discussion

Experimental analysis

The experimental results demonstrate that the selected statistical features, when combined with appropriate machine learning classifiers, can effectively capture emotional patterns in EEG signals while maintaining computational efficiency. Contrary to the typical expectation that ensemble methods outperform simpler algorithms, this study found that KNN consistently delivered the highest classification accuracy and reliability across both valence and arousal dimensions.

Among the three classifiers evaluated are KNN, LightGBM, and XGBoost. KNN emerged as the top-performing model, achieving an average accuracy of 96.77% and 96.67% for valence and arousal, respectively. It also produced high scores across other evaluation metrics, including AUC, precision, recall, F1-score, Kappa, and MCC. These results demonstrate that, under suitable conditions, well-engineered traditional models like KNN can outperform more complex ensemble methods, particularly when paired with effective feature extraction and normalization strategies.

While XGBoost and LightGBM also showed strong performance with accuracies ranging from 83% to 88% - they did not surpass KNN in overall evaluation. XGBoost performed better than LightGBM, particularly in terms of MCC and AUC, suggesting stronger predictive agreement and discriminative power. However, these gains came at the cost of increased model complexity and longer training time, which may limit their practicality in low-latency or resource-constrained environments.

Interestingly, LightGBM, despite being designed for fast training and low memory consumption, underperformed relative to both KNN and XGBoost. This may be due to its sensitivity to hyperparameter configurations and its reliance on sufficient data diversity to optimize its leaf-wise growth strategy effectively.

The success of simple time- and frequency-domain statistical features, including mean, RMS, standard deviation, entropy, crest factor, and kurtosis, highlights the potential of lightweight feature engineering for real-time emotion classification. These descriptors captured meaningful distinctions in EEG patterns associated with emotional states, supporting the premise that complex deep feature extraction pipelines are not always necessary for achieving high classification performance.

Overall, this experimental analysis confirms the viability of developing lightweight, interpretable, and highly accurate EEG-based emotion recognition systems using statistical features and classical machine learning models. The findings advocate for the practical deployment of such systems in real-time applications such as mental health monitoring, emotion-aware gaming, and adaptive human-computer interaction, particularly where low latency and computational efficiency are critical.

Comparative evaluation of classifiers (valence dimension)

As summarized in Table 1, the KNN model delivered the highest overall classification performance among the three evaluated models. It achieved a remarkable accuracy of 96.77%, along with equally high recall, precision, and F1-score values (each at 96.77%), suggesting a highly balanced classifier. Agreement metrics such as Cohen's Kappa (93.42%) and MCC (93.42%) further reinforce the model's predictive reliability and strong agreement with ground truth labels. These outcomes highlight KNN's suitability for EEG-based emotion recognition tasks, especially in binary classification settings such as valence recognition.

In contrast, the LightGBM model exhibited the lowest performance across nearly all metrics. It achieved an accuracy of 84.08% and an AUC score of 82.71%, indicating a more limited ability to distinguish between low- and high-valence emotional states. Although precision (84.71%) slightly exceeded recall (84.08%), the difference was marginal. However, the relatively lower Cohen's Kappa (66.87%) and MCC (67.87%) scores suggest a moderate level of agreement between predicted and actual values. This may indicate LightGBM's sensitivity to noise or overfitting on high-dimensional EEG data.

The XGBoost model offered a middle ground, with an accuracy of 88.55% and an AUC score of 87.80%, outperforming LightGBM but falling short of KNN's metrics. Both precision (88.70%) and recall (88.55%) were nearly equal, indicating balanced performance. Its Kappa (76.44%) and MCC (76.72%) scores demonstrated fair-to-strong agreement, making it a reliable and robust classifier. Furthermore, XGBoost's efficiency and scalability make it a strong candidate for real-time deployment, especially in resource-constrained environments such as edge devices and mobile health applications.

In summary, while KNN emerged as the most accurate and consistent model for valence classification, XGBoost presented a compelling trade-off between performance and scalability. LightGBM, though relatively less effective in this configuration, could still offer advantages when optimized further or integrated into ensemble strategies.

S/N	Metric	KNN	LightGBM	XGBoost
1	Accuracy	96.77%	84.08%	88.55%
2	AUC	96.68%	82.71%	87.80%
3	Recall	96.77%	84.08%	88.55%
4	Precision	96.77%	84.71%	87.80%
5	F1-Score	96.77%	83.79%	88.47%
6	Kappa	93.42%	66.87%	76.44%
7	MCC	93.42%	67.87%	76.72%
8	Training Time	< 1 seconds	9.6 seconds	38.56 seconds

TABLE 1: Models Comparison Valence Dimension

AUC, Area Under the Curve; KNN, K-Nearest Neighbor; LightGBM, Light Gradient Boosting Machine; MCC, Matthew's Correlation Coefficient; XGBoost, Extreme Gradient Boosting

Comparative evaluation of classifiers (arousal dimension)

As presented in Table 2, the KNN model demonstrated the highest overall accuracy (96.68%) among the evaluated classifiers for the arousal dimension. The recall, precision, and F1-score, all at 96.68%, reflect the model's excellent consistency and balance in identifying both low- and high-arousal emotional states. Additionally, the Cohen's Kappa (93.12%) and MCC (93.12%) values underscore the model's strong agreement with the true labels, reaffirming its robustness and reliability in binary emotion classification tasks.

In contrast, the LightGBM yielded the lowest classification accuracy (83.34%) among the three models. The AUC score (81.10%), indicative of its discriminative capacity, was also the lowest. While the precision (84.02%) slightly surpassed recall (83.34%), the overall difference was minimal, suggesting a slight bias toward minimizing false positives. However, the Cohen's Kappa (64.38%) and MCC (65.67%) values indicate only moderate agreement, suggesting potential limitations in LightGBM's performance consistency for arousal state detection across diverse EEG input distributions.

The XGBoost model again served as an intermediate performer. With an accuracy of 87.54% and AUC of 86.21%, it performed significantly better than LightGBM and approached KNN's level of performance. Precision (87.72%) was slightly higher than recall (87.54%), indicating minimal bias and a well-calibrated prediction distribution. The Cohen's Kappa (73.77%) and MCC (74.20%) further support its fair-to-strong agreement, validating the model's robustness. Given its scalability and optimized execution speed, XGBoost remains a compelling candidate for deployment in adaptive systems and mobile platforms, particularly in applications demanding real-time arousal detection from EEG signals.

In summary, while KNN provided superior performance across all metrics for arousal classification, XGBoost offered a practical balance of accuracy, efficiency, and generalizability. LightGBM, although the weakest performer in this context, may benefit from hyperparameter tuning or hybrid ensemble integration.

S/N	Metric	KNN	LightGBM	XGBoost
1	Accuracy	96.68%	83.34%	87.54%
2	AUC	96.49%	81.10%	86.21%
3	Recall	96.68%	83.34%	87.54%
4	Precision	96.68%	84.02%	87.72%
5	F1-Score	96.67%	82.89%	87.39%
6	Kappa	93.12%	64.38%	73.77%
7	MCC	93.13%	65.67%	74.20%
8	Training Time	< 1 seconds	8.3 seconds	36.27 seconds

TABLE 2: Models Comparison Arousal Dimension

AUC, Area Under the Curve; KNN, K-Nearest Neighbor; LightGBM, Light Gradient Boosting Machine; MCC, Matthew's Correlation Coefficient; XGBoost, Extreme Gradient Boosting

Model comparison with existing works 1

Table 3 provides a comparative analysis of classification accuracy between the proposed models and the approach reported by Han et al. [4]. This comparison reveals that the KNN model in this study achieved higher accuracy than the method used by Han et al. [4], demonstrating the effectiveness of the selected features and machine learning pipeline. While the comparison is limited to accuracy, it offers a useful benchmark, suggesting that well-designed traditional models can compete with or outperform more complex approaches under similar conditions.

Feature Extraction Methods	Classification Methods	Valence Accuracy (%) [4]	Feature Extraction Methods	Classification Methods	Proposed Work Valence Accuracy (%)
Depth-wise separable convolution and LSTM	E2ENNet (LSTM)	63.46	Time (Statistical) and Frequency (Entropy) Features	LightGBM (10 cross-validation)	84.08
Depth-wise separable convolution and LSTM	E2ENNet (CNN)	94.89	Time (Statistical) and Frequency (Entropy) Features	XGBoost (10 cross-validation)	88.55
Depth-wise separable convolution and LSTM	E2ENNet (CNN + LSTM)	96.25	Time (Statistical) and Frequency (Entropy) Features	KNN (10 cross-validation)	96.77

TABLE 3: Comparison of the Proposed Work with Existing Work 1

CNN, Convolutional Neural Network; KNN, K-Nearest Neighbor; LightGBM, Light Gradient Boosting Machine; LSTM, Long Short-Term Memory; XGBoost, Extreme Gradient Boosting

Model comparison with existing works 2

Table 4 presents a comparative analysis of classification accuracy between the proposed models and the approach reported by Alotaibi and Fawad [5]. This comparison highlights notable performance differences across classification methods, with the current study's KNN model outperforming the method in Alotaibi and Fawad [5] in terms of accuracy. These findings underscore the potential of traditional machine learning approaches when paired with informative feature extraction techniques to deliver superior results in EEG-based emotion recognition tasks.

Feature Extraction Methods	Classification Methods	Valence Accuracy (%) [5]	Feature Extraction Methods	Classification Methods	Proposed Work Valence Accuracy (%)
Alex Net	KNN Weighted	93.20	Time (Statistical) and Frequency (Entropy) Features	LightGBM (10 cross-validation)	84.08
ResNet-50	KNN Weighted	92.70	Time (Statistical) and Frequency (Entropy) Features	XGBoost (10 cross-validation)	88.55
ResNet-101	KNN Weighted	91.50	Time (Statistical) and Frequency (Entropy) Features	KNN (10 cross-validation)	96.77
InceptionV3	KNN Weighted	92.40			
Google Net	KNN Weighted	93.83			

TABLE 4: Comparison of the Proposed Work with Existing Work 2

KNN, K-Nearest Neighbor; LightGBM, Light Gradient Boosting Machine; XGBoost, Extreme Gradient Boosting

The findings of this study demonstrate that a relatively simple classifier such as KNN can achieve predictive performance that rivals or even surpasses more complex deep learning models under certain conditions. Specifically, the KNN model attained a valence classification accuracy of 96.77%, slightly outperforming the best results reported by Chen et al. [6] and Ma et al. [7], who employed deep learning-based feature extraction techniques. This outcome highlights the effectiveness of the proposed time-domain and frequency-domain feature extraction strategy, suggesting that high classification accuracy in EEG-based emotion recognition does not strictly depend on deep learning architectures provided that the features are well-engineered and sufficiently discriminative.

Moreover, the model's consistent performance across different validation schemes underscores its robustness and generalizability within the current dataset. However, it is important to recognize that deep learning models may offer superior generalization in cross-subject or cross-dataset applications, where they can automatically learn complex, non-linear representations that better capture inter-subject variability and broader context.

Additionally, while KNN performed remarkably well in the present binary classification setup (low vs. high valence/arousal), future investigations should assess whether this performance advantage persists in more challenging classification frameworks, such as multi-class emotion recognition or continuous dimensional modeling (e.g., using a broader range of valence and arousal values). These finer-grained emotional states may require more expressive models to capture subtle variations in EEG patterns, potentially shifting the advantage toward deep learning approaches.

Conclusions

This study introduces a lightweight and interpretable EEG-based emotion recognition framework that integrates statistical features from both the time and frequency domains with classical machine learning models. Leveraging the DEAP dataset, the proposed system demonstrates high classification accuracy while maintaining low computational complexity, underscoring its potential for real-time applications. Among the evaluated classifiers, the KNN model achieved the highest performance, attaining a valence classification accuracy of 96.77%, thereby outperforming XGBoost (88.55%) and LightGBM (84.08%). These results highlight the effectiveness of traditional machine learning techniques, when combined with robust feature engineering, in achieving performance levels that rival or even exceed those of more computationally intensive deep learning approaches.

Beyond performance, the simplicity and efficiency of the proposed framework make it highly adaptable for real-time emotion-aware systems, including deployment on resource-constrained platforms such as edge devices and mobile systems. Nevertheless, it is worth noting that deep learning models may offer superior robustness in cross-dataset or cross-subject scenarios, where generalization is a critical challenge. Future work will therefore focus on expanding validation across unseen subjects to improve reliability, exploring multi-modal integration (e.g., combining EEG with facial expressions or physiological signals), and investigating user-specific calibration strategies to enhance personalization. Furthermore, extending the framework to accommodate multi-class and dimensional emotion representations will broaden its applicability in diverse affective computing contexts, particularly in healthcare, adaptive human-computer interaction, and mental health monitoring.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Tijesunimi D. Olashore, Olutayo K. Boyinbode, Oladunni A. Daramola, Mary T. Kinga

Acquisition, analysis, or interpretation of data: Tijesunimi D. Olashore

Drafting of the manuscript: Tijesunimi D. Olashore

Critical review of the manuscript for important intellectual content: Olutayo K. Boyinbode, Oladunni A. Daramola, Mary T. Kinga

Supervision: Olutayo K. Boyinbode, Oladunni A. Daramola

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Picard RW, Vyzas E, Healey J: Toward machine emotional intelligence: analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2001, 23:1175-1191. 10.1109/34.954607
2. American Psychiatric Association: What is ADHD?. https://www.psychiatry.org/patients-families/adhd/what-is-adhd.
3. Wang F, Wu S, Zhang W, Xu Z, Zhang Y, Wu C, Coleman S: Emotion recognition with convolutional neural network and EEG-based EFDMs. Neuropsychologia. 2020, 146:107506. 10.1016/j.neuropsychologia.2020.107506
4. Han Z, Chang H, Zhou X, Wang J, Wang L, Shao Y: E2ENNet: An end-to-end neural network for emotional brain-computer interface. Frontiers in Computational Neuroscience. 2022, 16:942979. 10.3389/fncom.2022.942979
5. Alotaibi FM, Fawad: An AI-inspired spatio-temporal neural network for EEG-based emotional status. Sensors. 2023, 23:498. 10.3390/s23010498
6. Chen JX, Zhang PW, Mao ZJ, Huang YF, Jiang DM, Zhang YN: Accurate EEG-based emotion recognition on combined features using deep convolutional neural networks. IEEE Access. 2019, 7:44317-44328. 10.1109/access.2019.2908285
7. Ma J, Tang H, Zheng WL, Lu BL: Emotion recognition using multimodal residual LSTM network. MM '19: Proceedings of the 27th ACM International Conference on Multimedia. 2019, 176-183. 10.1145/3343031.3350871
8. Zhong Q, Zhu Y, Cai D, Xiao L, Zhang H: Electroencephalogram access for emotion recognition based on a deep hybrid network. Frontiers in Human Neuroscience. 2020, 14:589001. 10.3389/fnhum.2020.589001
9. Alhalaseh R, Alasasfeh S: Machine-learning-based emotion recognition system using EEG signals. Computers. 2020, 9:95. 10.3390/computers9040095
10. Hassouneh A, Mutawa AM, Murugappan M: Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods. Informatics in Medicine Unlocked. 2020, 20:100372. 10.1016/j.imu.2020.100372
11. Attia A, Amin K, Ibrahim M: Machine learning algorithms for enhancing emotion recognition from EEG signals. IJCI International Journal of Computers and Information. 2023, 10:54-65. 10.21608/ijci.2023.199502.1099
12. Koelstra S, Muhl C, Soleymani M, et al.: DEAP: A database for emotion analysis using physiological signals. IEEE Transactions on Affective Computing. 2012, 3:18-31. 10.1109/t-affc.2011.15

13. Latifoğlu F, Penekli S, Orhanbulucu F, Chowdhury MEH: A novel approach for Parkinson’s disease detection using Vold-Kalman order filtering and machine learning algorithms. Neural Computing and Applications. 2024, 36:9297-9311. 10.1007/s00521-024-09569-2