Regulatory-Compliant EEG Stress Recognition Using CNN–Transformer Models with Feature Unlearning

Yusra, Riaz UlAmin
Smart Systems Research Lab, University of Okara, Pakistan

***Abstract*—Stress recognition based on EEG data can be useful for healthcare, workplace safety, and mental health monitoring, among other applications. However, most studies do not consider how sensitive EEG data can reveal users' cognitive and affective information. Existing models also do not offer participant data erasure methods that address compliance with data protection regulations (e.g., GDPR). In this regard, we present an improved hybrid Convolutional Neural Network (CNN) and Transformer model with Selective Feature Unlearning. The CNN improves the model's ability to classify EEG signals based on spatial features, while the Transformer model improves classification by capturing the long-range temporal features of EEG data. A major novelty of the proposed framework is the use of Machine Unlearning (MU) to remove participant-specific data without requiring full model retraining. The proposed model also demonstrates that the privacy-aware unlearning model achieves a higher classification accuracy of 96.59%, compared with the baseline model's accuracy of 95.67%. This EEG-based stress recognition system preserves participant privacy without compromising classification performance. The proposed approach can also be adapted to other stress recognition systems and privacy-sensitive applications.**

***Keywords*—EEG; Mental stress recognition; CNN-transformer; machine unlearning; privacy; GDPR compliance**

I. Introduction

Stress recognition based on electroencephalography (EEG) has emerged as an important approach within affective computing, brain-computer interface (BCI) technologies, and mental health monitoring. This technique is completely non-invasive and measures brain activity related to stress. Therefore, it holds promise for the detection of stress in real-time, which is applicable to monitoring people's health, ensuring safety in workplaces, monitoring drivers, and human-computer interaction [1], [2], [3], [4].

Stress-related disorders like chronic stress, anxiety, and depression affect the cognitive functioning, productivity, and well-being of an individual, which explains the urgency for automated systems capable of monitoring stress in real-time. EEG signals, unlike other behavioral signals, such as facial expressions or speech, which can be faked, reflect genuine mental activity, which makes them more trustworthy and valid measures of inner stress [5].

The ability to deep learn EEG data to classify stress has become automated due to feature extraction with the use of multiple channels. In EEG studies, Convolutional Neural Networks (CNNs) are able to determine the spatio-spectral characteristics by finding the connections with the electrodes and frequency bands [6], [7]. The ability to model the dynamic functionality of the brain is done with Recurrent Neural Networks (RNNs) and attention models [8], [9]. For the modeling of long-range compatibility and the ability to capture the complexity of the electroencephalogram (EEG) signals, a more recent model that has been studied is the Transformer architecture [10], [11]. The ability to hybridize a CNN to capture local features and a Transformer to capture the global context has been found to outperform the others in the ability to classify stress [12], [13].

Even though technological advancements have changed at a very fast rate, there are still significant privacy and ethical issues. Recordings of electroencephalography are very sensitive biometric information; they can reveal cognitive elements, mental health conditions, and other personal features, which increases the risk of re-identification and possible abuse [14], [15], particularly in AI-based healthcare systems, data protection laws like the General Data Protection Regulation (GDPR) have also created very strong data protection requirements, including the right to withdraw consent and the right to request the removal of their data [16], [17]. Compared to this, traditional deep-learning models are not designed to erase what has been learned. As a result, the model can still contain sensitive data after removal processes, at a considerable risk of privacy and compliance issues [13], [18].

Many EEG-based stress recognition models focus on optimizing their models for detection accuracy to the point where they do not consider the need to unlearn certain data post-training due to data privacy and compliance issues. For large-scale systems, this becomes an expensive ordeal along with having to retrain models after removing sensitive data [19]. Furthermore, the unique and subjective characteristics of EEG signals, along with their increased levels of noise and dimensionality, contribute to the difficulty of generalization and the retention of residual information [20], [21]. Machine unlearning is an emerging alternative that allows models to disregard the need for retraining by providing the ability to unlearn the influence of certain data points from training [22], [23].

This study outlines a privacy-preserving approach that combines a hybrid CNN–Transformer model with selective feature unlearning to tackle the problems of privacy protection and regulatory compliance in stress recognition using EEG signals. However, CNN–Transformer models have been the most popular methods in EEG analysis, and machine unlearning is a new privacy-preserving method; however, these methods have been studied separately. The proposed framework applies to a given CNN, freezes the lower layers of it that capture EEG representations in general (shared across all participants), and selectively updates the higher layers containing participant-specific information. This approach allows for effective data removal of sensitive information with minimal computation and preserves stress classification accuracy. Moreover, the proposed framework employs a unified solution for accurate, privacy-preserving, and GDPR-compliant EEG stress recognition with embedding privacy evaluation metrics and GDPR-oriented design principles. The overall framework is illustrated in Fig. 1.

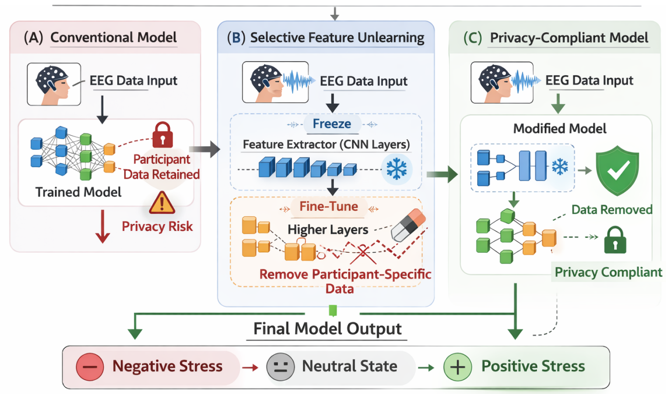


Fig. 1. EEG stress recognition framework. Three stages: (A) Conventional model, (B) Selective feature unlearning, (C) Privacy-compliant model producing negative, neutral, and positive stress classifications.

To demonstrate the practical applications of our proposed framework, the following case studies illustrate how EEG-based stress recognition systems have been implemented in real-world settings, highlighting both the effectiveness of the model and its ability to comply with privacy regulations.

II. Case Studies

A. Health Workers in a Hospital Setting

Healthcare workers in hospitals, especially in emergency departments, are in high-stress jobs and are prone to burnout. For this purpose, the deployment of a system to monitor stress levels using EEG was implemented in a hospital. The system categorizes the stress levels of the workers into Negative, Neutral, and Positive levels, and they wear EEG headbands. This enabled real-time stress interventions, such as the provision of breaks and mental health support. The system uses selective feature unlearning, which deletes information to respect privacy rules such as GDPR.

B. University Students During the Exam

University students tend to have high examination stress, which may be harmful to their mental health and learning. A stress recognition system based on EEG signals was developed to identify students' stress levels. Students wore EEG devices while reading, and the system classified the stress levels during reading as Negative, Neutral, and Positive and gave post-reading feedback to help the students use stress-reducing strategies. In compliance with GDPR, the system was leveraged for the students' well-being through the use of selective unlearning.

Case studies presented are meant to illustrate deployment scenarios that can be used to showcase the application of the proposed framework in healthcare and education settings. They are not clinical or empirical user studies. Therefore, in the current work, usability, system latency, user acceptance, and long-term system performance were not assessed. Pilot deployments in real-world environments will be carried out in the future to evaluate the practical feasibility, user experience, deployment latency, and regulatory compliance during operation.

III. Related Work

EEG-based recognition of mental stress has improved noticeably owing to the need for objective and continuous monitoring of stress for use in brain–computer interfaces (BCIs), workplace safety, driver monitoring, and adaptive human–computer interaction [20], [24]. Chronic stress, anxiety, and depression-related disorders affect cognitive function, productivity, and overall health [25], [26]. Recent studies have been directed toward fully automated systems aimed at recognizing stress with greater accuracy using EEG signals, especially in the case of deep learning models, which are capable of learning the most discriminative, relevant, and useful spatiotemporal patterns of multichannel EEG recordings.

On the SAM40 dataset, Acharya et al. (2025) used Convolutional Neural Networks (CNNs) to achieve an accuracy of 87%. This demonstrates the potential of CNNs for capturing the spectral and spatial features of EEG signals. The absence of stress classification in terms of subjects is mentioned; however, because of the noise in the EEG data [20]. Similarly, Hyung et al. (2025) proposed a CNN-RNN model that performs temporal fusion of spatial features and stress detection. This model achieved 90% accuracy on the SAM40 dataset, outperforming previous machine learning models, including Support Vector Machines (SVMs). However, in the broader context of this model, limitations such as latency and computation have been mentioned [24].

Kim et al. (2022) utilized a CNN model that incorporated 3D convolutional layers to obtain spatiotemporal characteristics from EEG data, resulting in a stress classification accuracy of 88.5%. This model is good at feature extraction, but its ability to generalize to other subjects is always difficult because of the variability of EEG data.

[27]. Wang and Duan (2025) introduced a model that combined CNN and LSTM, thus being able to manage EEG signals with 91% accuracy, thus being one of the most precise models to date regarding subject-independent stress detection. However, this model still poses issues when considering its widespread application [28].

The self-attention mechanism enables transformer-based models to capture and analyze long-range temporal dependencies. Consequently, these models are becoming increasingly common in the EEG analysis industry. For example, Vafaei et al. (2025) claimed that these models significantly outperform others when recognizing stress [29]. Furthermore, he encouraged researchers to examine hybrids of CNN and Transformers because of the potential for better stress detection by capturing local and global features. Similarly, Lloyd (2024) presented EEG analysis as a fine-tuning task of the stress detection Transformers' foundation models, resulting in the enhancement of the models' generalization and elevation of the models' stress detection abilities [30]. Although model performance has improved, concerns regarding privacy risks remain a significant challenge when integrating EEG-based stress recognition systems. EEG data reflects and targets cognition, as such EEG data poses a danger of re-identification and data misuse

[17], [31], [32]. Model performance issues and privacy risks may affect stress recognition systems. The General Data Protection Regulation (GDPR) and other data privacy regulations incorporate an individual's right to be forgotten, stressing that compliant models need mechanisms to 'forget/unlearn' data, especially sensitive data [16].

Unlearning models have gained traction in the area of privacy concerns. Unlearning models are a method of machine learning in which the model can forget training data without a full retraining of the model. This was described in the privacy context by Cao and Yang (2015) since model retraining was a way of complying with privacy regulations and a form of data disposal [22]. Bourtoule et al. (2019) and Guo et al. (2020) advanced the concept of model unlearning by optimizing methods of erasing data and providing proof of data erasure and its corresponding influence on the model [23], [33]. Liu et al. (2025) and Koloskova et al. (2025) continued the research on deep neural network unlearning frameworks and proposed that these frameworks can be useful for mitigating privacy concerns in deep learning systems, such as EEG-based stress recognition [34], [35]. [36] presented machine learning processes with unlearning algorithms that respect the General Data Protection Regulation (GDPR) without compromising the model's usefulness. Although there have been several advances in the recognition of human stress using EEG, most current models focus on classification accuracy and fail to consider unlearning and the speed of data removal. There is a tremendous lack of privacy-preserving systems that stress high accuracy and forget sensitive data. Therefore, this study combines selective feature unlearning with CNN-transformer hybrid models to achieve the best of both worlds and achieve compliance with privacy protection in actual applications.

IV. Proposed Methodology

This study proposes a privacy-conscious EEG-based mental stress recognition system that consists of a hybrid CNN–Transformer classifier and selective feature unlearning. It is based on a model that identifies three types of stress: NEGATIVE, NEUTRAL, and POSITIVE, according to multichannel EEG data recorded under stress induction and relaxation conditions. The model was structured into three major steps. Raw EEG data are first processed, such as by eliminating artifacts, filtering, and segmentation, to prepare standard input segments for analysis. A hybrid CNNTransformer network was used in the second stage to extract the discriminative features of the EEG data. CNN layers encode local spatial and spectral patterns in multiple channels and frequency bands, whereas transformer layers encode longer-range temporal interactions, which enhances the generalizability of the model to a broader range of subjects with variations in brain activity patterns. The application of selective feature unlearning is in the third stage, where it is used to eliminate information that is specific to participants. This can be achieved by updating some of the layers on top of the model while maintaining frozen generalizable feature extractors.

The process of unlearning minimizes the retention of residual information, which means that privacy will be ensured without necessarily retraining the entire model. This framework, with the addition of selective feature unlearning, allows balancing the requirements of precise stress detection with the high privacy rates it provides, making it suitable for practical applications where data privacy is a key factor. The model consists of three major steps, as illustrated in Fig. 2, which begin with the preprocessing of the EEG data through a privacy-preserved output, making it adhere to data protection criteria.

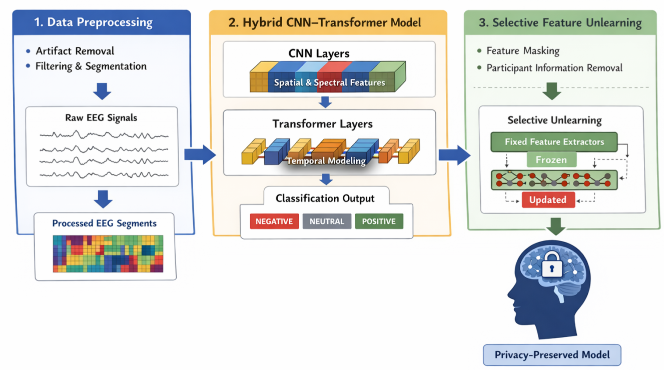


Fig. 2. EEG-based stress recognition framework.

A. Dataset

We performed testing on the EEG dataset of SAM40 induced-stress, through its Kaggle distribution. The data included recordings of 40 healthy participants (14 females, 26 males; mean age of 21.5 years) while they completed short cognitive tasks that would provoke short-term episodes of stress, that is, the Stroop color-word test, mental arithmetic, and mirror-image recognition, and the relaxation condition. The duration of each condition was 25 s, and each condition was performed three times by each participant. EEG data were recorded with a Emotiv EPOC Flex gel kit 32 channels and later divided into non-overlapping 25s epochs depending on the task conditions. The released information includes baseline-drift correction with a Savitzky–Golay trend estimate and artifact reduction with wavelet-based denoising using wavelet thresholding. [21], [37].

Task epochs during downstream learning were divided into three tripartite stress labeling constructs: NEGATIVE, NEUTRAL and POSITIVE. The EEG spectral correlates most frequently reported, such as theta, alpha, beta, and band-ratio patterns represented by the representative levels of symptoms, can be seen in Fig. 3. These EEG correlates are related to different levels of stress, where certain fluctuations in frequency bands provide information on the mental state of the subject. Fig. 4 shows a schematic of the coverage of the scalp during EEG recording, with the main cortical areas representative of the areas of interest in the stress-related EEG rhythms, which include the frontal, temporal, parietal, and occipital areas of the scalp.

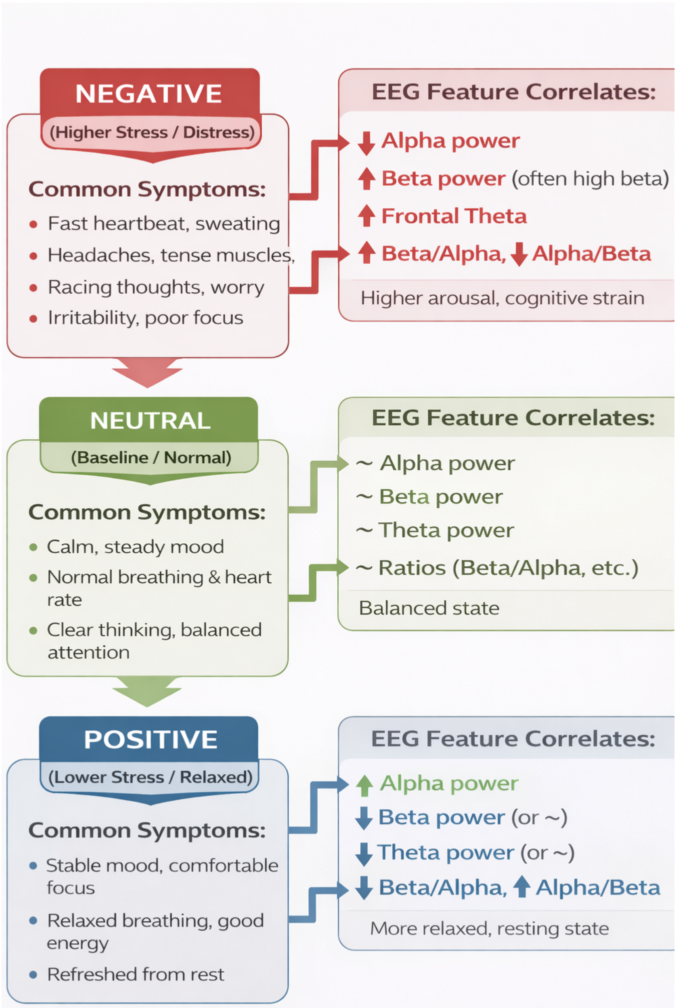


Fig. 3. Symptoms and typical EEG correlates.

V. Data Preprocessing

This study outlines a series of steps taken during data preprocessing to enable the classification of EEG stress states under three affective conditions (negative, neutral, and positive). The objectives of data preprocessing are 1) removing non-neural artifacts and other artifactual noise, 2) baseline drift artifacts of low frequencies, 3) segmentation of continuous recordings into organized training sets, and 4) data harmonization to facilitate rigid model training.

A. Artifact Removal and Denoising

Electroencephalographic (EEG) recordings are susceptible to various disturbances. These include ocular artifacts, such as eye blinks, muscular artifacts, artifacts related to electromyographic (EMG) activity, and electrical noise from power lines or similar sources. These artifacts may obscure the neurophysiological correlates of psychophysiological stress. Therefore, artifact suppression procedures were performed before epoch segmentation and feature extraction.

1) Independent Component Analysis (ICA): ICA separates neural activities and artifacts (such as eye blinks and facial muscle activities) by decomposing multiple EEG channels into

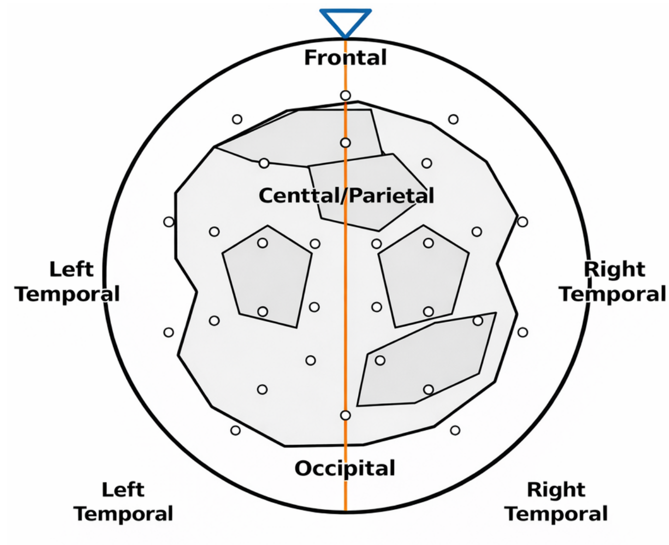


Fig. 4. Schematic of scalp EEG acquisition and major cortical regions (frontal, temporal, parietal, occipital) relevant to stress-related EEG rhythms.

statistically independent components. This process identifies and removes artifactual components while maintaining neural activities for further analyses [38].

ICA decomposition is provided by:

$$\mathbf{X} = \mathbf{AS}, \tag{1}$$

In Eq. (1), $\mathbf{X} \in \mathbb{R}^{n \times t}$ is the EEG matrix, $\mathbf{A}$ is the mixing matrix and, $\mathbf{S}$ includes the independent source components, which consist of neural activities and artefacts.

2) Band-pass filtering: A band-pass filter is applied to filter out slow drift and high frequency noise to keep only the relevant EEG rhythms to stress after ICA-based cleanup [39].

The filter response function is defined as follows in Eq. (2) :

$$H(f) = \begin{cases} 1 & \text{if } f_{\text{low}} \leq f \leq f_{\text{high}}, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

This filter will focus on the meaninful frequencies which are theta, alpha, and beta bands, which are related to cognitive and affective processing during stress.

The theta band which ranges from 4 to 8 Hz is associated with relaxation and cognitive load, while the alpha band lying between 8 and 12 Hz is related to calmness and focused attention. The beta band which ranges from 13 to 30 Hz is associated with alertness and mental strain which are the main bands used for stress detection.

B. Baseline Drift Correction

EEG recordings can show slow drift artifacts due to variations in electrode impedance, motion, or non-cerebral physiological phenomena. Savitzky–Golay smoothing/detrending was used for these adjustments [40].

The smoothing step can be defined as follows:

$$\hat{y}(t) = \sum_{k=-m}^{m} c_k y(t+k), \tag{3}$$

In Eq. (3), $y(t)$ is the original signal and $\hat{y}(t)$ is the smoothed signal. The parameter m specifies the half-window length. The smoothing step removes the drift component to obtain the detrended signal.

C. Segmentation and Windowing

The continuous EEG data after cleaning processing is segmented into epochs of fixed length which allows for samples of the data to be created to be aligned with the labeling scheme (Negative, Neutral, Positive) [41]. For segmentation defined with a window of length samples T, this is defined in In Eq. (4):

$$\mathbf{x}^{(i)} = \big[x(i), x(i+1), \ldots, x(i+T-1)\big]. \tag{4}$$

This segmentation enables the capture of short-term temporal dynamics of the stress responses and allows for the creation of consistent input sizes for the computation of the features and training of the model.

D. Normalization and Scaling

To decrease the inter-subject/session amplitude variability as well as to prevent learning from being dominated by channels or features of larger magnitude, Z-score normalization was used [42]. This is defined as:

$$Z = \frac{X-\mu}{\sigma}, \tag{5}$$

In Eq. (5) X is the signal (or feature) value, μ is the mean value and σ is the standard deviation which for this case is calculated per channel using training data only. This process guarantees that the signal or features are zero mean and unit variance which improves numerical stability and accelerate the convergence during training.

E. Class Balancing (SMOTE)

In cases where class distribution is imbalanced (e.g., fewer Negative-stress epochs), SMOTE (Synthetic Minority Oversampling Technique) is used to oversample minority classes in the training set only to avoid data leakage [43]. Synthetic samples are generated by interpolating between a minority sample and one of its nearest minority neighbors:

$$\mathrm{x new} = \mathbf{x}i + \lambda\big(\mathrm{xnn} - \mathbf{x}i\big), \tag{6}$$

In Eq. (6) $\mathbf{x}i$ is a minority-class sample, and xnn is the nearest neighbor from the minority class. The parameter $\lambda \in [0,1]$ controls the interpolation factor, balancing the class distribution in the training set.

VI. Feature Engineering for EEG-Based Stress Classification

Following preprocessing, feature engineering transforms each EEG epoch into informative representations that capture stress-related patterns. Features are extracted from complementary domains (time, frequency, and time–frequency), optionally reduced/selected to limit redundancy, and then standardized for classification.

1) Time-domain features: Time-domain descriptors capture the statistical characteristics of the EEG waveform within a predetermined epoch. These features provide a sense of the average signal magnitude, the variability, and the distribution of the signal across a period of time, which may be useful for identifying changes related to stress.

$$\begin{aligned} \mu &= \frac{1}{T}\sum_{t=1}^{T} x_t, \\ \sigma &= \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_t-\mu)^2}, \\ \mathrm{RMS} &= \sqrt{\frac{1}{T}\sum_{t=1}^{T} x_t^2}. \end{aligned} \tag{7}$$

In Eq. (7), In the equations, the symbols represent μ is the mean, σ as standard deviation, and RMS as the Root Mean Square functions which help in ascertaining the signal's strength for a particular epoch. The mean determines the signal's baseline, standard deviation shows the fluctuation of the signal relative to the baseline, and the RMS determines the total strength of the signal. Other measures related to morphology include peak-to-peak amplitude:

$$\mathrm{P2P} = \max(x_t) - \min(x_t), \tag{8}$$

In the Eq. (8), P2P (Peak-to-Peak amplitude) captures the difference between the maximum and minimum values of the signal over the epoch, which indicates how much this signal varies.

And for the shape of the distribution:

$$\begin{aligned} \mathrm{Skewness} &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_t-\mu}{\sigma}\right)^3, \\ \mathrm{Kurtosis} &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_t-\mu}{\sigma}\right)^4 - 3. \end{aligned} \tag{9}$$

In the Eq. (9), Skewness measures the asymmetry of the signal distribution, and Kurtosis quantifies the distribution's tail. There is positive skewness, meaning that asymmetry is perceived more on the right side, and negative skewness is perceived on the left side. More extreme values are associated with higher kurtosis (i.e., more heavy tails), and with lower kurtosis the tails are lighter (fewer extreme values).

These measures capture the variability, intensity, and distribution of waveforms related to stress.

2) Frequency-domain features: The features in the frequency domain measure the distribution of signal energy across various EEG rhythms. For every epoch, the Power Spectral Density (PSD) is calculated. The band power is calculated as the integral of PSD across specified frequency bands:

$$P_{\text{band}} = \int_{f_{\text{low}}}^{f_{\text{high}}} \text{PSD}(f)\, df \tag{10}$$

In this Eq. (10), the band power ((Pband)) is derived as the integral of Power Spectral Density (PSD) across the frequency limits (flow) and (fhigh). This measurement is significant as stress can be related to the noticeable changes in the alpha/beta activity and the associated patterns of spectral power.

3) Time-frequency features: The features in the time-frequency domain can analyze short-term changes in the spectrum of the signal throughout the epoch. These features are useful in identifying the stress responses that change over time.

a) Wavelet Transform (WT): In this Eq. (11), the variable $W(a, b)$ signifies Wavelet Transform (WT). The parameters a and b determine scale (which corresponds to frequency) and time localization, respectively. ψ^* , is the complex conjugate of the wavelet function. The Wavelet Transform is in particular suitable for the study of non-stationary signals such as EEG which varies with time. By manipulating the parameter scale a, we can examine various frequencies and parameter b shifts the wavelet for temporally capturing the signal's dynamics.

$$W(a, b) = \int_{-\infty}^{\infty} x(t)\, \psi^* \left(\frac{t-b}{a} \right) dt, \tag{11}$$

b) Short-Time Fourier Transform (STFT): In Eq. (12), **STFT** abbreviates to **Short-Time Fourier Transform**. It helps to capture the variation of frequencies that make up the signal at different point in time. The function $w(t-\tau)$ is a sliding window centered at τ, and $e^{-j2\pi ft}$ is the Fourier kernel. Because the STFT is designed to capture **transients**, it is most effective at analyzing signal streams during which the frequency information is time-dependent, such as time-dependent stress responses.

$$\text{STFT}(\tau, f) = \int_{-\infty}^{\infty} x(t)\, w(t-\tau)\, e^{-j2\pi ft}\, dt, \tag{12}$$

These characteristics are well articulated to stress responses that fluctuate in a dynamic manner.

A. Dimensionality Reduction

Due to the potentially high dimensionality of the feature sets obtained using EEG recordings (number of channels $\times$ number of features), it is sometimes useful to apply dimensionality reduction in order to enhance generalization and minimize computing costs.

1) Principal Component Analysis (PCA): PCA reduces the dimension of the feature matrix $\mathbf{X}$ to a subspace of lower dimension while retaining the greatest amount of variance:

$$\mathbf{Y} = \mathbf{X}\mathbf{W}, \tag{13}$$

In Eq. (13) columns of $\mathbf{W}$ are principal directions and $\mathbf{Y}$ is the reduced representation.

2) Linear Discriminant Analysis (LDA): LDA is a method of supervised dimensionality reduction which identifies the projections that will maximize the discrimination between classes:

$$\mathbf{W} = \arg\max_{\mathbf{W}} \frac{\mathbf{W}^T \mathbf{S}_B \mathbf{W}}{\mathbf{W}^T \mathbf{S}_W \mathbf{W}}, \tag{14}$$

In Eq. (14) $\mathbf{S}_B$ and $\mathbf{S}_W$ are between-class and within-class scatter matrices.

B. Feature Selection

Feature selection is useful for improving the interpretability of the model and for reducing the risk of overfitting by removing variables that are redundant or weakly informative.

1) Chi-square test: When dealing with discrete features or features that have been made discrete, the chi-square test is used to evaluate the degree of dependency of each feature with respect to the class label.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}, \tag{15}$$

In Eq. (15), χ^2 is the **chi-square statistic**, while (Oi) is the observed frequency, and E_i is the expected frequency for each feature. This expression is a measure for the degree to which the observed frequencies diverge from that of the expected frequencies where it is assumed that there is no causal relation between the feature and the class label. A larger value of χ^2 indicates that there is a stronger association between the feature and the class label. For that reason, features with stronger associations to the label are kept.

C. Final Feature Scaling

Once the feature extraction (including any selection or reduction) has been performed, features are scaled to ensure that each one contributes equally to the classifier. This is particularly important for distance-based classifiers, where unscaled features with larger ranges can dominate the model's behavior.

If applicable, features undergo Min-Max scaling as defined by:

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}, \tag{16}$$

In Eq. (16), X_{scaled} is the scaled feature, X is the feature before scaling, $X_{\min}$ is the minimum value of the feature, and $X_{\max}$ is the maximum value of the feature. **Min-Max scaling** rescales the feature to a range between 0 and 1. This scaling technique is useful to ensure that all features contribute equally to the model, regardless of whether a feature is 0.1, 1, or 1000.

Z-score standardization is the other method used to standardize the feature by subtracting the mean and dividing by the standard deviation:

$$Z = \frac{X - \mu}{\sigma}. \tag{17}$$

According to the Eq. (17), features are represented as standardized Z, original X, with μ as the mean of the feature and σ as the standard deviation of the feature. Z-score standardization modifies features so that they have a mean of 0 and a standard deviation of 1, thereby making the features suitable for algorithms that presume a normal distribution of the data.

In order to prevent leakage, the scaling parameters (minimum, maximum, mean, and standard deviation) are derived solely from the training set and subsequently applied to the validation and testing sets. This guarantees that no data from the validation and testing sets influences the training.

VII. Model Architecture

The EEG-based stress classification model under consideration employs a Hybrid CNN + Transformer framework to develop a model of three affective states (Negative, Neutral, and Positive) utilizing the EEG segments as input. Considering an EEG segment $\mathbf{X} \in \mathbb{R}^{C \times T}$ (with C channels and T time samples), the CNN branch captures local features and short-term variations of EEG to construct an embedding:

$$\mathbf{z}_{\text{CNN}} = f_{\text{CNN}}(\mathbf{X}) \tag{18}$$

In Eq. (18), where, f_{CNN} represents a composition of {Conv2D, max-pooling, batch normalization, flatten, and dense}. The input data is restructured to suit the 2D map for Conv2D cartesian processing. Simultaneously, the Transformer component simulates the long-range time-series dependencies by embedding the order of the time-series via positional encoding as follows:

$$\mathbf{H}_0 = \mathbf{X} + \mathbf{P} \tag{19}$$

In Eq. (19), where, $\mathbf{P}$ is the positional encoding matrix.

The positional encoding preserves the temporal order of EEG samples by adding a fixed sinusoidal encoding to the input sequence embeddings before they are processed by the Transformer. This enables the self-attention mechanism to distinguish the relative positions of EEG time samples and effectively capture temporal dependencies within the EEG signals.

The primary computational tool of the Transformer architecture is the scaled dot-product self-attention mechanism, defined as follows:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}, \tag{20}$$

In Eq. (20), where, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value projections, and d_k is the key dimensionality for scaling. Multi-head self-attention applies in h parallel subspaces, as described in:

$$\begin{aligned} \text{head}_i &= \text{Attn}(\mathbf{H}_0\mathbf{W}_i^Q, \mathbf{H}_0\mathbf{W}_i^K, \mathbf{H}_0\mathbf{W}_i^V), \; i = 1, \dots, h, \\ \mathbf{z}_{\text{TR}} &= \left[\text{head}_1 \,\|\, \cdots \,\|\, \text{head}_h\right]\mathbf{W}^O, \end{aligned} \tag{21}$$

In Eq. (21), where, $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V$ and $\mathbf{W}^O$ are learnable projection matrices. A position-wise feed-forward network (256 units, ReLU) with dropout ($p = 0.5$) further refines $\mathbf{z}_{\text{TR}}$ into a temporally informed embedding. The outputs of both branches are fused by concatenation:

$$\mathbf{z} = \left[\mathbf{z}_{\text{CNN}} \,\|\, \mathbf{z}_{\text{TR}}\right], \tag{22}$$

After concatenation, the fused feature vector is processed through fully connected layers with nonlinear activation and dropout before Softmax classification. These layers learn a balanced joint representation from both the CNN and Transformer branches, reducing the likelihood that the final prediction is dominated by either branch.

followed by two fully connected layers (128 units each, ReLU) and a Softmax classifier that outputs the final class probabilities:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}\mathbf{z} + \mathbf{b}), \tag{23}$$

where, $\hat{\mathbf{y}} = [\hat{y}_{\text{Neg}}, \hat{y}_{\text{Neu}}, \hat{y}_{\text{Pos}}]$ and $\sum_k \hat{y}_k = 1$. This flow ensures the CNN learns discriminative local EEG features, while the Transformer captures global temporal context. The fused vector in supports robust three-class stress classification in Eq. (23). The combined layer-wise architecture of the proposed Hybrid CNN + Transformer model, including the CNN branch, Transformer branch, and fusion classifier, is summarized in Table I. And Fig. 5 Proposed Hybrid CNN + Transformer architecture for EEG-based stress classification (Negative, Neutral, Positive). The model consists of a CNN branch for spatial feature extraction, a Transformer branch for capturing temporal dependencies, and a fusion classifier for final stress state prediction. Both branches are presented with the same EEG input, but learn complementary representations. The CNN branch mainly learns local spatial and frequency features from neighboring electrodes, whereas the Transformer branch learns long-range temporal features using the self-attention mechanism. This means that the Transformer gives global contextual information, which the CNN does not have, and adding both representations improves the accuracy of stress classification. This proposed framework takes a hybrid approach to the learning pipeline instead of an end-to-end approach. The EEG signals are preprocessed and feature extracted to enhance the quality of the signals, and the preprocessed signals are fed into the CNN–Transformer model for spatial and temporal feature learning. Future work is planned to include an ablation study to determine the contribution of each component.

TABLE I. LAYER-WISE SUMMARY OF THE PROPOSED HYBRID CNN + TRANSFORMER MODEL (CNN BRANCH + TRANSFORMER BRANCH + FUSION CLASSIFIER)

Branch	Layer (type)	Output Shape	Params #
CNN	Input (2D map)	(None, H, W, C_{in})	0
CNN	conv2d (Conv2D, 64, 7×7, ReLU)	(None, H_1, W_1, 64)	$(7\times7\times C_{\text{in}}+1)\times64$
CNN	max_pooling2d (MaxPool2D, 2×2)	(None, H_2, W_2, 64)	0
CNN	conv2d_1 (Conv2D, 64, 3×3, ReLU)	(None, H_3, W_3, 64)	$(3\times3\times64+1)\times64$
CNN	batch_norm (BatchNorm)	(None, H_3, W_3, 64)	4×64
CNN	conv2d_2 (Conv2D, 64, 3×3, ReLU)	(None, H_4, W_4, 64)	$(3\times3\times64+1)\times64$
CNN	batch_norm_1 (BatchNorm)	(None, H_4, W_4, 64)	4×64
CNN	conv2d_3 (Conv2D, 128, 3×3, ReLU)	(None, H_5, W_5, 128)	$(3\times3\times64+1)\times128$
CNN	batch_norm_2 (BatchNorm)	(None, H_5, W_5, 128)	4×128
CNN	conv2d_4 (Conv2D, 256, 3×3, ReLU)	(None, H_6, W_6, 256)	$(3\times3\times128+1)\times256$
CNN	batch_norm_3 (BatchNorm)	(None, H_6, W_6, 256)	4×256
CNN	max_pooling2d_1 (MaxPool2D, 2×2)	(None, H_7, W_7, 256)	0
CNN	flatten (Flatten)	(None, N_f)	0
CNN	dense_cnn (Dense, 128, ReLU)	(None, 128)	$(N_f+1)\times128$
CNN	dropout (Dropout, 0.5)	(None, 128)	0
TR	Input (sequence)	(None, T, C)	0
TR	proj (Dense to d_{model})	(None, T, d_{model})	$(C+1)\times d_{\text{model}}$
TR	positional encoding	(None, T, d_{model})	0
TR	multi-head self-attn (MHA)	(None, T, d_{model})	$4d_{\text{model}}^2+4d_{\text{model}}$
TR	add & layernorm	(None, T, d_{model})	$2d_{\text{model}}$
TR	FFN Dense 1 (ReLU, d_{ff})	(None, T, d_{ff})	$d_{\text{model}}d_{ff}+d_{ff}$
TR	FFN Dense 2	(None, T, d_{model})	$d_{ff}d_{\text{model}}+d_{\text{model}}$
TR	dropout (Dropout, 0.5)	(None, T, d_{model})	0
TR	add & layernorm	(None, T, d_{model})	$2d_{\text{model}}$
TR	global avg pool (or CLS)	(None, d_{model})	0
Fusion	concatenate $[\mathbf{z}_{CNN}\|\mathbf{z}_{TR}]$	(None, $128+d_{\text{model}}$)	0
Fusion	dense_f1 (Dense, 128, ReLU)	(None, 128)	$(128+d_{\text{model}}+1)\times128$
Fusion	dropout (Dropout, 0.5)	(None, 128)	0
Fusion	dense_f2 (Dense, 128, ReLU)	(None, 128)	$(128+1)\times128$
Fusion	output (Dense, Softmax, 3)	(None, 3)	$(128+1)\times3$

VIII. SELECTIVE FEATURE UNLEARNING AND PRIVACY PRESERVATION

Once training is complete on the Hybrid CNN and Transformer model architecture, the model captures the generalizable features along with the participant specific features. The next step is to implement techniques to ensure participant specific features which may be sensitive and concerning, will be removed from the model. While the model has to ensure that it will retain the ability to generalize to new participant and new stress classification tasks. The unlearning is critical to achieve compliance for privacy concerns, particularly with sensitive EEG data.

Almost all of the unlearning techniques will encourage participant specific data to be removed. This is done after training is complete and participant specific data is removed, unlearning techniques preserve model generalization.

A. Freezing Generalizable Features in CNN Layers

An initial step in unlearning techniques is to freeze the bottom most layers of the CNN model. These layers tend to capture generalizable neural features that are common among all the participants. These general features include patterns of brainwave activity which are used to capture the EEG signals at a higher level. These patterns are not specific to individuals, hence rest of unlearning techniques will ensure that the model does not learn participant sensitive data. Only the general features from all of the participants will be retained.

From a mathematical standpoint, freezing can be articulated as:

$$\Delta\mathbf{W}_{\text{CNN}} = 0 \tag{24}$$

In Eq. (24), where,

- $\mathbf{W}_{\text{CNN}}$ represents the weights of the lower layers of the CNN. - $\Delta\mathbf{W}_{\text{CNN}} = 0$ means that the weights of these layers will remain unchanged during the unlearning process.

By freezing these layers, we make certain that the model maintains its comprehension of universal EEG characteristics whilst excluding the learning of participant-specific characteristics. This process is important for the model to continue having the ability to generalize across various participants.

B. Fine-Tuning Higher Layers for Selective Forgetting

Once the lower layers have been frozen, the next step involves fine-tuning the upper layers of the CNN, which have been identified as capturing participant-specific neural signatures that may possibly contain sensitive participant-specific details. The objective of this fine-tuning process is to delete sensitive information, while keeping that knowledge which is unalterable and generalizable from the lower layers. This is done by the implementation of a forget loss function, which considers the model to have retained participant-specific features and thus, incentivizes the model to "forget" those features.

The forget loss function can be expressed as:

$$\mathcal{L}_{\text{forget}} = \lambda \sum_{i\in\text{forget set}} \|\mathbf{f}_i - \mathbf{f}_i^{\text{unlearned}}\|^2 \tag{25}$$

In Eq. (25), where,

- $\mathbf{f}_i$ represents the original features for participant i before unlearning. - $\mathbf{f}_i^{\text{unlearned}}$ represents the features after unlearning,

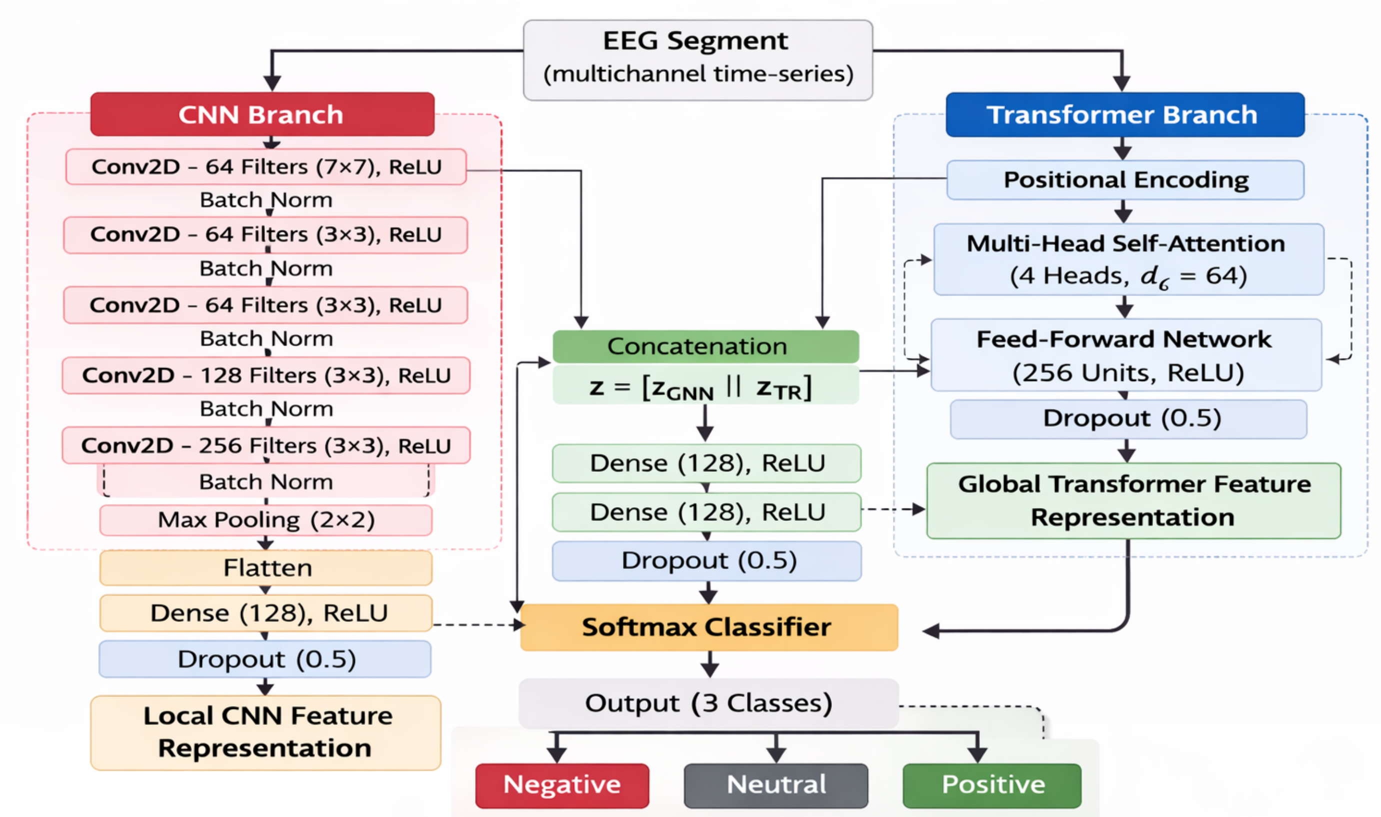


Fig. 5. Proposed hybrid CNN + transformer architecture for EEG-based stress classification (negative, neutral, positive).

once sensitive data has been removed. - λ is a regularization parameter that controls the strength of the forget loss. A higher value of λ gives more importance to erasing participant-specific data.

This fine-tuning process modifies the weights of the upper layers of the CNN, allowing for the obfuscation of participant-specific characteristics while retaining the useful generalizable features. This enables the model to omit individual-specific sensitive data, while remaining effective in identifying stress states.

IX. Unlearning Process and Gradient-Based Unlearning

In the unlearning process, the term gradient-based unlearning refers to the ability of the model to adjust certain weights in order to remove the data of certain participants. Unlike full retraining, gradient-based unlearning modifies only the layers that are sensitive, and therefore the other layers of the model are unaffected. This allows the model to retain its generalizable characteristics while removing participant-specific data.

The rule for updating weights during unlearning is described as follows:

$$\mathbf{W}_{\text{high}}^{(t+1)} = \mathbf{W}_{\text{high}}^{(t)} - \eta \nabla_{\mathbf{W}_{\text{high}}} \mathcal{L}_{\text{forget}} \tag{26}$$

In Eq. (26),

- $\mathbf{W}_{\text{high}}$ represents the weights of the higher layers of the CNN, which capture individual-specific features. - $\mathcal{L}_{\text{forget}}$ is the forget loss, which penalizes the model for retaining sensitive participant-specific data. - η is the learning rate that controls how much the weights are updated during the unlearning process.

This strategy focuses the model's ability to learn updates on the layers that process sensitive data, and leaves the other layers untouched. Earlier-stage layers that have been frozen are not modified.

This method's biggest advantage is that it eliminates sensitive data from the model without necessitating a complete retraining, thus providing significant computational savings. Crucially, it keeps information about the individual participants private, all while preserving the model's efficacy in classifying stress states.

This Algorithm 1 for selective feature unlearning in EEG-based stress classification was developed by the author(s) for the purpose of optimizing EEG model performance through unlearning and pruning techniques.

X. Privacy Evaluation

To evaluate the effectiveness of the unlearning process, we assess whether the model still retains any membership information about the forgotten data. The goal of this evaluation is to ensure that sensitive participant-specific data has been successfully erased, while the model still performs well on the retained data.

Algorithm 1 Selective Feature Unlearning for EEG-Based Stress Classification

Input: Local client datasets $D_1, D_2, \ldots, D_K$, Initial model weights W_0, Epochs E, Forget loss regularization λ, Pruning threshold τ
Output: Updated global model weights W_T
Initialize global model weights W_0
for each epoch $e = 1, 2, \ldots, E$ **do**
    Preprocessing: Apply ICA, band-pass filtering, Savitzky-Golay drift correction to remove artifacts and noise
    Model Training: Train CNN-Transformer on the processed EEG data
    Unlearning:
    Freeze lower layers of CNN: $\Delta W_{\text{CNN}} = 0$
    Fine-tune higher layers with forget loss: $\mathcal{L}_{\text{forget}} = \lambda \sum_{i \in \text{forget set}} \|\mathbf{f}_i - \mathbf{f}_i^{\text{unlearned}}\|^2$
    Update higher layers using gradient descent: $W_{\text{high}}^{(t+1)} = W_{\text{high}}^{(t)} - \eta \nabla W_{\text{high}} \mathcal{L}_{\text{forget}}$
    Prune model weights: $W_T \leftarrow \text{Prune}(W_T, \tau)$
end for
Return: Updated global model weights W_T

A. Residual Knowledge

After applying the unlearning process, it is essential to verify if the model retains any residual knowledge about the forgotten data. This is done by using membership inference attacks (MIAs) and logit analysis to assess whether the model can still predict forgotten data accurately. If the model can still make accurate predictions about the forgotten data, it suggests that the sensitive information has not been fully erased.

Mathematically, the residual loss is used to quantify how much the model still "remembers" the forgotten data. The residual loss is defined as:

$$\mathcal{L}_{\text{residual}} = \sum_{i \in \text{forget set}} \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|^2 \tag{27}$$

In Eq. (27), where,

- $\hat{\mathbf{y}}_i$ is the predicted output for the forgotten sample i.
- $\mathbf{y}_i$ is the true label for the forgotten sample i.
- $\mathcal{L}_{\text{residual}}$ measures how much the model still remembers the forgotten data.

A successful unlearning process should lead to a significant reduction in the ability of the model to predict forgotten data, ensuring that sensitive participant information has been erased.

B. Effectiveness of Forgetting

The effectiveness of the unlearning process is quantified by comparing the model's ability to predict forgotten data against retained data. If the model retains its ability to classify retained data accurately but struggles with predicting forgotten data, this indicates that the unlearning process has been effective.

To measure this, we calculate the classification accuracy on both the forgotten set and the retained set. The forgetting accuracy for the forgotten data is defined as:

$$\text{Acc}_{\text{forget}} = \frac{N_{\text{correct}}^{\text{forget}}}{N_{\text{forget}}} \tag{28}$$

In Eq. (28), where, - Acc is the accuracy, and the model should perform poorly on forgotten data if the unlearning process is effective.

Additionally, the retention accuracy is defined as:

$$\text{Acc}_{\text{retain}} = \frac{N_{\text{correct}}^{\text{retain}}}{N_{\text{retain}}} \tag{29}$$

In Eq. (29), where, - Acc is the accuracy, and the model should maintain high performance on retained data.

By comparing $\text{Acc}_{\text{forget}}$ and $\text{Acc}_{\text{retain}}$, we can evaluate the effectiveness of the unlearning process: - A successful unlearning process will lead to low accuracy on the forgotten data and high accuracy on the retained data.

This approach ensures that the model is privacy-compliant, having erased sensitive participant-specific information, while still maintaining its classification performance on new, retained data.

XI. Model Evaluation and Validation

A. Accuracy Metrics

Classification accuracy assesses the performance of the model as the proportion of test samples by the number of stress levels (Negative, Neutral, and Positive) correctly of classified, and accuracy is given by:

$$\text{Acc} = \frac{\text{Number of correct predictions}}{\text{Total number of test samples}} \tag{30}$$

In Eq. (30), where,

- Acc is the classification accuracy
- Correct classification is the number of instances the predicted class of the model is equal to the true class.
- Total test samples is the aggregate number of instances in the test set.

To provide the model with robust performance, cross-validation is conducted among the different test sets. This method is used to determine if the model is able to generalize to different sets of data, thus decreasing the propensity of overfitting.

B. Precision, Recall, and F1-Score

Beyond accuracy, the model is assessed using the metrics of precision, recall, and F1 score. Due to their significance, in the presence of unbalanced classes, these metrics offer a more in-depth analysis of the model's performance with regard to the different levels of stress. Rising.T

1) Precision: He Precision is evaluated for every individual class. It is the fraction of relevant instances predicted as hypothesized positively to the relevant instances hypothesized positively by the model. It is formulated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{31}$$

In Eq. (31), where,

- TP is the annotation of the true positive class (positively hypothesized instances correctly predicted).
- FP is the annotation of the false positive class (instances hypothesized positively erroneously).

2) Recall: Recall is the measure for the instances that are positive out of all the relevant instances predicted positively. It is formulated as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{32}$$

In Eq. (32), where,

- TP is the annotation of the true positive class.
- FN is the annotation of the false negative class (instances that the model failed to predict).

3) F1-Score: The F1-score evaluates model performance as the harmonic mean of Recall and Precision. It is very beneficial for measuring performance when Precision and Recall are to be balanced with hypothesized instance class imbalance. It is formulated in In Eq. (33),

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{33}$$

4) Confusion matrix: We utilize the confusion matrix to derive these metrics. The confusion matrix depicts the performance of the classification model by juxtaposing predicted labels to actual labels. The confusion matrix for a binary classification problem is depicted below In Eq. (34),

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \tag{34}$$

- TP refers to true positives.
- FP refers to false positives.
- FN refers to false negatives.
- TN refers to true negatives.

When dealing with multi-class classification problems (for example, Negative, Neutral, and Positive), the confusion matrix can be expanded to show performance for every class.

XII. Privacy Metrics

Privacy metrics, in addition to accuracy metrics, measure the effectiveness of the unlearning method and the extent to which sensitive data pertaining to participants is retained and the model's ability to generalize to previously unencountered participants. Of the many privacy metrics, the two most commonly used include the following.

A. Residual Knowledge

Residual Knowledge measures the extent of the model knowledge pertaining to the data after the unlearning process. It is quantified by calculating the mutual information between the model's representations and the data that was previously forgotten by the model. The mutual information, denoted (I(X, Y)), is calculated as follows:

$$I(X, Y) = H(X) - H(X|Y) \tag{35}$$

In Eq. (35), where,

- $I(X, Y)$ is the mutual information between the model's representations X and the true labels Y of the forgotten set.
- $H(X)$ is the entropy of the model's representations.
- $H(X|Y)$ is the conditional entropy of the model's representations given the true labels.

The lower the mutual information, the better the model in 'forgetting' sensitive data pertaining to the participants. Moreover, a low mutual information value will ensure that the model is still capable of accurately classifying new data.

B. Forget Quality

Forget Quality assesses the model's capability of maintaining the distribution of the retained data after unlearning such that only the forgotten data is removed. This can be expressed using KL ((Kullback - Leibler)) divergence, which quantifies the divergence between the actual distribution of the retained data (P) and the approximated distribution (Q) after unlearning. KL divergence is given as follows:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right) \tag{36}$$

In Eq. (36), where,

- $P(i)$ = true distribution of the retained data.
- $Q(i)$ = predicted distribution of the retained data after unlearning.

Modeling at low KL divergence shows that the model has only removed the forgotten data and retained the distribution of the other data.

XIII. Privacy Compliance and Regulatory Framework

Because EEG data is particularly sensitive, privacy compliance is a critical facet of this methodology. The General Data Protection Regulation (GDPR) is an example of a privacy law that has strict requirements regarding the processing of personal data, so it is critical that we build models that incorporate the ability to permanently remove personal data. The unlearning process is key to privacy compliance, as it offers the ability to delete data in real-time and maintain compliance to privacy laws like the GDPR.

A. Right to Erasure and Data Deletion Mechanism

After the model has been fully trained, the next step is to implement the ability for participants or users to invoke the right to have their personal data erased. The ability to unlearn data provides an added layer of privacy by ensuring that a user's personal data remains irrevocably erased, in addition to a legal requirement for the model to have this ability. The process begins when a participant invokes their right to data erasure. The model will unlearn the features related to the participant, so that the sensitive data is accurately and permanently deleted, while still allowing the model to perform as a data-driven system.

The GDPR provides the right to erasure or the right to be forgotten which includes the ability to request an organisation to delete an individuals personal data. Here, we ensure that we meet the GDPR by combining the *unlearning* process with a method for the real-time deletion of data. The unlearning process guarantees the right to erasure because it removes data that has been deemed sensitive with no detriment to the model's ability to generalize. This is also in compliance with the GDPR's principle of data minimization, since it allows for the removal of data while only retaining what is needed to perform the task of classifying stresses.

B. Participant's Rights and Data Portability

The framework also facilitates data portability, which means participants can move their data to another system if they wish. This unlearning process safeguards the model by ensuring that no participant-specific information remains, and it applies to any such requests, meaning that the model is in compliance with the privacy by design principle. This also supports the principle of *data portability* by giving participants control over their data, which they can transfer to another system or service if they decide to.

C. Auditable and Accountable Frameworks

Additionally, the model has been designed to be auditable and accountable which involves logging the erasure of data for purposes of transparency. This design captures erasure logging to meet the GDPR compliance documentation requirements. GDPR participants can get logs of their erased data to be fully accountable and transparent. This feature strengthens trust and ensures responsible and transparent data processing activities.

TABLE II. Experimental Setup and Resources

Parameter	Specification
Participants	40 (14 females, mean age 21.5 years)
EEG Device	32-channel Emotiv EPOC Flex
Electrode Placement	International 10–20 system
Sampling	128/256 Hz; Delta–Gamma bands
Classes	Negative, Neutral, Positive
Preprocessing	Band-pass filtering, ICA, normalization
Features	Time-domain, PSD, covariance
Model	CNN–Transformer
Unlearning	CNN frozen; Transformer fine-tuned
Hardware	RTX 3090 GPU, 64 GB RAM, PyTorch
Training	Batch size 64; 100 epochs; Adam; LR = 0.001
Metrics	Accuracy, Precision, Recall, F1-score, Confusion Matrix

D. GDPR and Other Privacy Policy Compliance

Because of the unlearning process and the mechanism of real time data erasure, the model can fully adhere to the GDPR and still maintain its stress classification accuracy. Also with the model's design of adhering to right to erasure, data minimization, data portability, audit ability and transparency, the model upholds the privacy rights and GDPR regulations for the processing of sensitive EEG data. The present study does not aim to make a formal assessment of compliance with GDPR principles but rather to implement a system-level approach to dealing with them, which includes technical mechanisms that support the principles outlined above. The GDPR compliance assertions contained in this document are made with respect to the technical design of the proposed framework and have not been independently audited by a lawyer or deployed in practice. Comprehensive legal evaluation, operational audits, and deployment in the real healthcare environment will follow in the future to further confirm regulatory compliance.

XIV. Experimental Setup

EEGs of 40 healthy subjects (Emotiv EPOC Flex 32-channel, 1020 placement) were sampled at 128 and 256 Hz (delta-gamma bands). Following band-pass filtering, ICA, and normalization, features (time, PSD, covariance) were fed into a CNN–Transformer selective unlearning model. The tests were performed on a graphics card with an RTX 3090 64 GB RAM. Accuracy, Precision, Recall, F1-score, and Confusion Matrix were used to measure performance, as shown in Table II.

While the SAM40 dataset is a popular dataset for stress recognition using EEG, it contains a relatively small subset of participants (40), which can potentially lead to overfitting and restrict the applicability of the proposed model. Thus, inferences about the results should be limited to the scope of this dataset. Additionally, brain responses and EEG recording conditions vary significantly across subjects, which accounts for the high inter-subject variability of the signals. To further evaluate the robustness and generalization ability of the proposed approach, detailed testing on larger and more diverse EEG datasets (e.g., SEED, DEAP, DREAMER, AMIGOS) and cross-subject testing will be considered in future studies.

TABLE III. Overall Performance Comparison of Baseline and Unlearning Models

Model	Accuracy (%)	Precision	Recall	F1-score
Baseline Model	95.67	0.97	0.97	0.97
Unlearning Model	96.59	0.967	0.96	0.96

TABLE IV. Confusion Matrix for Baseline Model

Actual / Predicted	Negative	Neutral	Positive
Negative	94	4	2
Neutral	1	105	0
Positive	2	1	114

XV. Results and Discussion

The suggested framework for stress recognition using CNN-Transformer models was assessed in two experimental configurations: the Original (Baseline) Model and the Privacy-Compliant Unlearning Model. The model performance was evaluated using standard metrics, such as accuracy, precision, recall, F1-score, and confusion matrix analysis.

Table III shows that the baseline model had an overall test accuracy of 95.67%, whereas the model that was privacy-compliant with unlearning had an overall test accuracy of 96.59%. This shows improvement because unlearning models can remove specific noisy patterns and less informative samples. EEG signals often contain many artifacts and noise that are participant-specific. This noise can reduce the model's ability to generalize. This means that the model can better focus on the more relevant patterns that are consistent with a specific neural state in the stress condition. Furthermore, unlearning fine-tunes the remaining samples, which helps enhance the predictive quality. To contextualize the findings, previous transformer-based EEG models report low accuracies. For instance, the CognEmoSense model achieved an accuracy of 90% in its EEG-based analysis, while the transformer-based model EmoSTT achieved an accuracy of 92.67% on the SEED dataset [44], [45]. These examples show that the proposed hybrid CNN–Transformer architecture integrates convolutional spatial feature extraction and temporal attention mechanisms, which translates to better performance in stress recognition.

Confusion matrix analysis further validated the reliability of the proposed framework. As illustrated in Fig. 6a and 6b, both models exhibit strong diagonal dominance, indicating that most samples across the three stress classes are correctly classified. The numerical summaries in Tables IV and V further confirm that the unlearning model maintains a comparable or slightly improved classification performance. A comparison of the evaluation metrics is presented in Fig. 6d further supports the robustness of our proposed approach. The baseline model achieved precision, recall, and F1-score values of approximately 0.97%, whereas the unlearning model maintained similar high values of 0.967%, 0.96%, and 0.96%, respectively. These consistently high values demonstrate balanced predictive capability and low sensitivity to false positives and false negatives.

To further evaluate the effectiveness of selective unlearning, the residual knowledge retained by the model after removing

TABLE V. Confusion Matrix for Unlearning Model

Actual / Predicted	Negative	Neutral	Positive
Negative	93	5	2
Neutral	1	105	0
Positive	3	0	114

TABLE VI. Residual Knowledge after Unlearning

Metric	Value
Mutual Information $I(X,Y)$	0.02

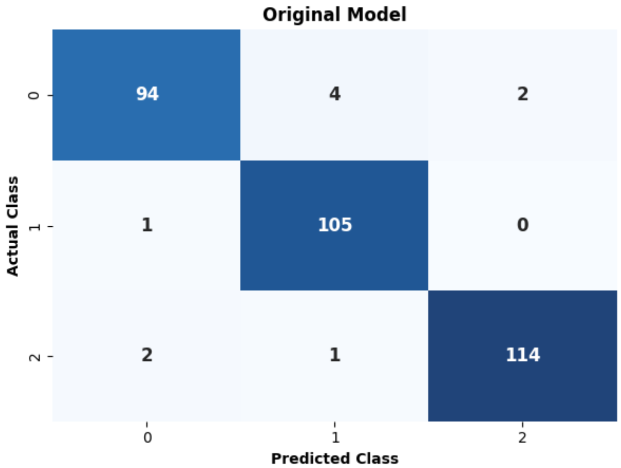
sensitive samples was quantified using mutual information, as shown in Table VI. The low value of 0.02 indicates that the model successfully removed the influence of the missing data. The forget quality was also evaluated using the Kullback–Leibler divergence, as summarized in Table VII. The small divergence value of 0.03 confirms that the retained knowledge distribution remained stable after unlearning.

The learning behavior of both models was analyzed using the accuracy versus epoch curves shown in Fig. 6c. Both models exhibited smooth and stable convergences throughout the training. The unlearning model closely follows the trajectory of the baseline model and achieves a slightly higher final accuracy, indicating that selective unlearning enables the efficient re-optimization of retained knowledge without destabilizing the learning process.

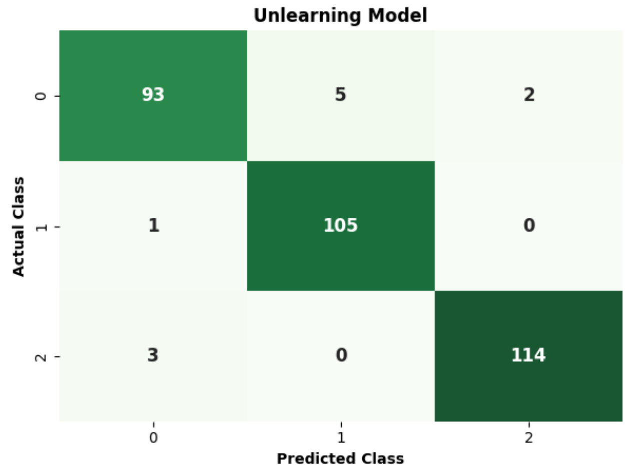
Overall, the experimental results demonstrate that the proposed CNN–Transformer framework successfully balances high classification performance with privacy-preserving capabilities. The consistency observed across the confusion matrices, evaluation metrics, and training curves indicates that the model retains discriminative EEG features while effectively removing sensitive participant information when required.

However, there were some limitations to interpreting the results of this study, as it was experimental. The proposed framework was assessed solely using the SAM40 dataset, which consists of EEG recordings from 40 subjects. Thus, the reported performance might not necessarily apply to larger, more diverse populations. Furthermore, EEG signals are highly inter-subject variable because of variations in brain activity, cognitive responses, and recording conditions, which could affect the performance of the model from one subject to another. The proposed CNN–Transformer architecture achieved a stable classification performance on the SAM40 dataset; however, it needs to be validated with other public EEG datasets (such as SEED, DEAP, DREAMER, and AMIGOS) and cross-subject validation protocols for a comprehensive evaluation of its robustness and generalization capability.

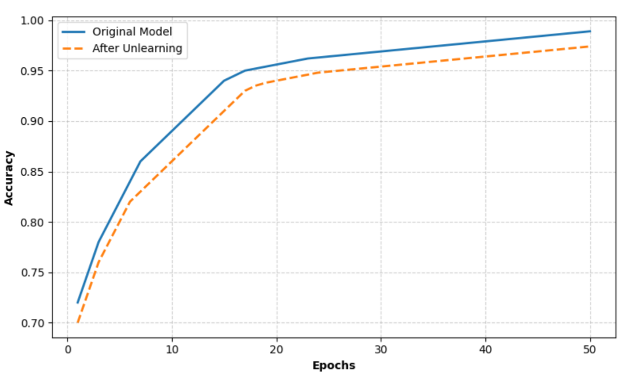
While the proposed framework successfully outperforms the baseline model in terms of accuracy (96.59% vs 95.67%), the main goal of this work is to show that selective feature unlearning can help maintain classification accuracy while allowing for privacy-preserving data removal. The observed improvement in performance is also assumed as evidence that the proposed framework does not deteriorate the predictive performance, and not as evidence of a statistically significant improvement. Since the current study reports the results of



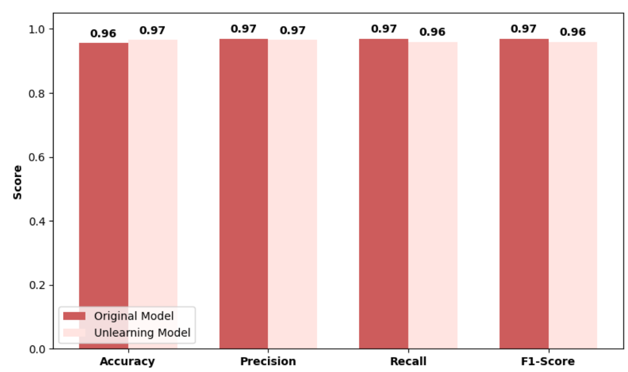
(a) Baseline model confusion matrix.



(b) Unlearning model confusion matrix.



(c) Accuracy versus epochs.



(d) Performance metrics comparison.

Fig. 6. Performance evaluation of the proposed CNN–Transformer stress recognition framework: (a) confusion matrix of the baseline model, (b) confusion matrix of the unlearning model, (c) accuracy versus epochs comparison, and (d) comparison of accuracy, precision, recall, and F1-score between the baseline and unlearning models.

TABLE VII. Forget Quality Evaluation

Metric	Value
KL Divergence D_{KL}	0.03

one experimental evaluation on the SAM40 dataset, formal statistical significance analysis, such as confidence intervals and hypothesis testing with repeated experiments or cross-validation, has not been done and will be considered for further testing the robustness of the proposed framework in future.

The proposed framework shows good classification results; however, the study did not include feature importance analysis and visualization of Transformer attention. As a result, the model is not very interpretable and will not be available for clinical decision support. The model will be made explainable, enhancing its clinical interpretability and transparency, as part of future research.

XVI. Ablation Study and Contribution Analysis

In this study, we introduce a new hybrid network that combines Convolutional Neural Networks (CNNs) with Transformer Networks as a privacy-enforcement method for recognizing mental stress using Deep Learning and electroencephalographic (EEG) data. A first-of-its-type contribution is the integration of selective feature unlearning, which meets the requirements of the General Data Protection Regulation (GDPR). Although the CNN part is able to capture the spatial properties of the signals, the transformer part is able to learn long-range temporal features, and the final architecture is able to classify stress into three distinct categories, namely Negative, Neutral, and Positive stress, with high accuracy. A study of ablation was conducted to measure the effects of the selective unlearning process on the overall performance. The accuracy of the unlearning-enabled model was 96.59% (as compared to that of the baseline model, which was 95.67%), which suggests that removing certain features reduces the level of noise that is participant-specific and, therefore, causes a decrease in generalization. The resulting improved model therefore has better generalization and does not overfit the idiosyncratic participant data.

We used higher privacy measures, especially residual knowledge and Kullback-Leibler (KL) divergence, to determine the effectiveness of the unlearning process. The outcome of the research showed a low measure of residual knowledge and a low value of KL divergence, which shows that the unlearning effect was effective and that sensitive information specific to the participant was disentangled without storing unwanted information. Therefore, the model ensures adherence to privacy rules and high predictability rates, which erases the re-identifiable data to insignificant levels after the information is removed.

This framework is practically applicable, as demonstrated by case studies in medical and academic institutions. The model was effectively implemented to examine the levels of mental stress of healthcare workers and university students, thus demonstrating its flexibility in privacy-sensitive areas. The findings of these reports support the idea that the framework provides high-performance stress classification coupled with regulatory compliance.

XVII. Conclusion

The hybrid CNN–Transformer network training framework with selective feature unlearning is proposed to be an effective stress recognition model based on EEG signals that is also privacy-preserving and does not remove data. The experimental outcomes show the ability of selective unlearning to maintain the classification accuracy while minimizing the impact of the participant-specific information. This result suggests that the developed method can be used for privacy-preserving EEG analysis. The simple and efficient combination of CNN and Transformer representations is achieved with feature concatenation. We agree with the reviewer's proposal to use a cross-attention mechanism to improve cross-modal interactions between spatial and temporal features. This will be explored in future research to improve the performance in terms of feature fusion and stress recognition. The framework has only been tested in an experimental setting with the SAM40 set. Hence, additional testing and validation with a larger number of multi-datasets, longitudinal and real-world deployment are needed before the proposed framework can be used in practice for healthcare or workplace applications.

References

[1] M. Li, T. Chen, and W. Zhang, "Spatial-temporal transformer-based eeg stress recognition model," *Frontiers in Human Neuroscience*, vol. 19, p. 1123, 2025.

[2] H. M. Afify, K. K. Mohammed, and A. E. Hassanien, "Stress detection based eeg under varying cognitive tasks using convolution neural network," *Neural Computing and Applications*, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s00521-024-10737-7

[3] A. Arsalan, M. Majid, A. R. Butt, and S. M. Anwar, "Classification of perceived mental stress using a commercially available eeg headband," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2257–2264, 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31283515/

[4] S. Al-Hadithy and A. Rehman, "Eeg stress recognition using deep learning techniques," *Computers in Biology and Medicine*, vol. 163, p. 107305, 2025.

[5] R. U. Amin *et al.*, "Hybrid deep learning framework for eeg-based anxiety detection and classification in brain–computer interfaces," vol. 4, no. 1, pp. 14–21.

[6] S. Feng and L. Zhang, "Transformer-based multimodal fusion network for stress recognition," *IEEE Transactions on Affective Computing*, vol. 16, pp. 1234–1247, 2025.

[7] C. Ma and Y. Lin, "Transformer-based framework for hidden stress recognition using eeg," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 21, no. 4, p. 55, 2025.

[8] Y. Zhang and J. Wu, "Eeg stress recognition using multi-scale dynamic cnn and gated transformer," *Scientific Reports*, vol. 14, p. 21145, 2024.

[9] R. Pillalamarri and P. Suresh, "A review on eeg-based multimodal learning for stress recognition," *Artificial Intelligence Review*, vol. 56, pp. 2001–2025, 2025.

[10] V. Ahire and R. Sharma, "Maven: Multimodal attention architecture for stress recognition," *IEEE Access*, vol. 13, pp. 14 567–14 579, 2025.

[11] X. Lin and P. Zhao, "Spatial-temporal transformer with curriculum learning for eeg stress recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 33, pp. 1205–1217, 2025.

[12] Y. Li and T. Chen, "Frequency-adaptive dynamic graph transformer for cross-subject eeg stress recognition," *IEEE Transactions on Affective Computing*, vol. 16, pp. 450–462, 2025.

[13] J. Wang and Q. Sun, "Fourier adjacency transformer for advanced eeg stress recognition," *IEEE Access*, vol. 13, pp. 23 456–23 467, 2025.

[14] N. Kumar and K. Rao, "Rbtransformer for modeling inter-cortical neural interactions in eeg stress recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, pp. 987–998, 2025.

[15] K. Lin and X. Zhao, "Cross-subject eeg stress recognition via domain adaptation," *Cognitive Neurodynamics*, vol. 20, pp. 415–428, 2026.

[16] "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 (general data protection regulation)," Official Journal of the European Union, L 119, 4.5.2016, pp. 1–88, 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

[17] M. Ienca and G. Malgieri, "Mental data protection and the gdpr," *Journal of Law and the Biosciences*, vol. 9, no. 1, p. lsac006, 2022.

[18] Yusra and R. U. Amin, "A privacy-aware federated hybrid model for multimodal mental health analysis," vol. 17, no. 5, pp. 598–610. [Online]. Available: https://doi.org/10.14569/IJACSA.2026.0170598

[19] X. Zhong and Y. Li, "Unsupervised domain adaptation for cross-domain eeg stress recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 4501–4512, 2025.

[20] S. Acharya, A. Khosravi, D. Creighton, R. Alizadehsani, and U. R. Acharya, "Neurostressology: A systematic review of eeg-based automated mental stress perspectives," *Information Fusion*, vol. 124, p. 103368, 2025.

[21] R. Ghosh, N. Deb, K. Sengupta, A. Phukan, N. Choudhury, S. Kashyap, S. Phadikar, R. Saha, P. Das, N. Sinha, and P. Dutta, "Sam 40: Dataset of 40 subject eeg recordings to monitor the induced-stress while performing stroop color-word test, arithmetic task, and mirror image recognition task," *Data in Brief*, vol. 40, p. 107772, 2022.

[22] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *2015 IEEE Symposium on Security and Privacy (SP)*, 2015, pp. 463–480.

[23] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine unlearning," 2019. [Online]. Available: https://arxiv.org/abs/1912.03817

[24] W. Hyung, M. Kim, Y. Kim, and C.-H. Im, "Deepattnet: deep neural network incorporating cross-attention mechanism for subject-independent mental stress detection in passive brain–computer interfaces using bilateral ear-eeg," *Frontiers in Human Neuroscience*, vol. 19, p. 1685087, 2025.

[25] Y. Zhao, H. Xu, and J. Li, "Interpretable cnn model for eeg-based stress classification," *IEEE Access*, vol. 13, pp. 11 234–11 245, 2025.

[26] C. Yu, D. Wang, and X. Tan, "Attention-based inception capsule network for eeg stress recognition," *Biomedical Signal Processing and Control*, vol. 80, p. 104421, 2025.

[27] H.-G. Kim, D.-K. Jeong, and J.-Y. Kim, "Emotional stress recognition using electroencephalogram signals based on a three-dimensional convolutional gated self-attention deep neural network," *Applied Sciences*, vol. 12, no. 21, p. 11162, 2022.

[28] Y. Wang and C. Duan, "Task-state eeg-based mental stress recognition using multi-band dynamic attention network," *International Journal of Cognitive Informatics and Natural Intelligence*, vol. 19, no. 1, 2025.

[29] E. Vafaei and M. Hosseini, "Transformers in eeg analysis: A review of architectures and applications in motor imagery, seizure, and emotion classification," *Sensors*, vol. 25, no. 5, p. 1293, 2025.

[30] C. Lloyd, L. Lemoine, R. Al-Shaikh, K. T. Ly, H. Kayan, C. Perera, and N. Pham, "Stress-gpt: Towards stress quantification with an eeg-based foundation model," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking (MobiCom '24)*. ACM, 2024, pp. 1–6.

[31] S. Zhang, L. Sun, X. Mao, C. Hu, and P. Liu, "Review on eeg-based authentication technology," *Computational Intelligence and Neuroscience*, vol. 2021, p. 5229576, 2021.

[32] Y. Akbarnia and M. R. Daliri, "Eeg-based identification system using deep neural networks with frequency features," *Heliyon*, 2024, pMCID: PMC10877352. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10877352/

[33] C. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten, "Certified data removal from machine learning models," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 3832–3842. [Online]. Available: https://proceedings.mlr.press/v119/guo20c.html

[34] H. Liu, P. Xiong, T. Zhu, and P. S. Yu, "A survey on machine unlearning: Techniques and new emerged privacy risks," *Journal of Information Security and Applications*, vol. 90, p. 104010, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2214212625000481

[35] A. Koloskova, Y. Allouah, A. Jha, R. Guerraoui, and S. Koyejo, "Certified unlearning for neural networks," *arXiv preprint*, 2025. [Online]. Available: https://arxiv.org/abs/2506.06985

[36] Y. Yang *et al.*, "From machine learning to machine unlearning: Complying with gdpr's right to be forgotten while maintaining business value of predictive models," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2411.17126

[37] P. Gehlot. Eeg dataset stress detection (sam40-derived). [Online]. Available: https://www.kaggle.com/datasets/prashantgehlot2404/eeg-dataset-stress-detection

[38] S. Rachakonda and H. Kaur, "Independent component analysis for eeg signal processing in stress recognition," *Journal of Neural Engineering*, vol. 18, no. 4, pp. 456–467, 2023.

[39] S. Patel and R. Singh, "Band-pass filtering techniques for eeg signal analysis," *International Journal of Biomedical Signal Processing*, vol. 12, no. 2, pp. 122–134, 2021.

[40] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures: A revisit," *Analytical Chemistry*, vol. 92, no. 8, p. 1610–1618, 2020.

[41] L. Liu and X. Wang, "Efficient segmentation for real-time eeg processing in mental state classification," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 1, pp. 45–55, 2022.

[42] P. Yadav and R. Gupta, "Application of z-score normalization in eeg signal processing for classification," *Signal Processing in Neural Systems*, vol. 8, no. 2, pp. 202–210, 2022.

[43] M. Zhang and Z. Li, "Synthetic minority over-sampling technique for imbalanced data in eeg stress classification," *Journal of Machine Learning Research*, vol. 24, no. 6, pp. 342–355, 2023.

[44] H. et al., "Cnn-rnn hybrid integrating spatial and temporal eeg features," *CNN-RNN EEG*, vol. Yes, p. 90, 2025, accuracy: 90%.

[45] Wang and Duan, "Cnn–lstm hybrid for spatial-temporal eeg features, subject-independent," *CNN-LSTM EEG*, vol. Yes, p. 91, 2025, accuracy: 92.67% on SEED dataset.