ISNNI: 3007-3391 HEC Category: Y

Journal of Computers and Intelligent Systems

Vol. 4, No. 1, 2026, pages 50 - 66

Journal Homepage:

https://journals.iub.edu.pk/index.php/JCIS/

Privacy-Preserving EEG Stress Classification via Hybrid CNN–Transformer Learning and Selective Machine Unlearning

Riaz Ul Amin[1], Yusra Irshad Ali[1], Javed Rashid[1], Muhammad Shoaib Saleem[2], Muhammad Sajjad Saleem[1]

[1]Department of Computer Science, University of Okara, Okara, Punjab 56310, Pakistan.

[2]Department of Mathematics, University of Okara, Okara, Punjab 56310, Pakistan.

ARTICLE INFO

Article History:

Received:	March	01, 2026
Revised:	March	20, 2026
Accepted:	March	25, 2026
Available Online:	March	30, 2026

Keywords:

Keyword **selective machine unlearning**
Keyword **stress recognition**
Keyword **Electroencephalography (EEG)**
Keyword **privacy-preserving learning**
Keyword **neural signal processing**

Classification Codes:

Funding:

This research received no specific grant from any funding agency in the public or not-for-profit sector.

ABSTRACT

Electroencephalography (EEG) for stress detection is an emerging area of interest in affective computing, brain-computer interfaces (BCIs), and mental health monitoring. EEG is beneficial in both determining stress-related patterns of neural activity and measuring related neural activity. This information is useful for health monitoring, workplace safety, and adaptive human-computer interaction. However, due to biometric nature of its data, EEG is a potential source of privacy violations, in consideration of the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA). This paper presents a Convolutional Neural Network (CNN) and Transformer model for classifying stress states of EEG data — *Low*, *Moderate*, and *High* — in a privacy-preserving manner. EEG data from subjects is first processed using band-pass filtering, Independent Component Analysis (ICA), normalization, and class balancing. This yields structured inputs for deep learning models. We propose a selective machine unlearning method to erase data pertaining to specific participants, without a full model retraining, by freezing a number of CNN layers and performing Transformer and classification layer updates. The results show the baseline model achieves 95.67% test accuracy, while the privacy compliant model post unlearning achieves 96.59%. The results show that privacy-preserving mechanisms can be used in conjunction with EEG-based stress monitoring solutions, without reducing the level of predictive accuracy of the solution.

Corresponding Author's Email: riazulamin@gmail.com

Citation:

1. Introduction

Stress and anxiety as well as other stress-related mental illnesses affect millions of people all over the world making objective tracking of stress levels a very important issue globally. Such conditions significantly undermine the quality of life, productivity, and overall well-being, which is why automated stress recognition systems based on mental health assessment, cognitive monitoring, and adaptive human-computer interaction

are developed [1].
EEG offers a valid and objective tool of identifying stress due to its high temporal resolution and the ability of measuring neural activity directly tied to cognitive and emotional processes objectively [2]. However, EEG data is one of the most sensitive biometric data, which is why the privacy concerns occur in the context of various laws, including the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) [3]. The majority of current EEG-based stress recognition technologies are focused on the accuracy of the classification but do not include regulatory compliance or safe post-deployment data elimination [4]. The recent progress of deep learning, especially hybrid CNN–Transformer-based models, has been seen to perform well with spatial-spectral feature extraction and temporal attention mechanisms [5]. Nevertheless, the privacy preservation is not usually discussed in these models. To address this shortcoming, we suggest a privacy-aware CNN–Transformer model with selective machine unlearning to EEG-based stress recognition. The suggested system categorizes EEG signals as either Low, Moderate, or High stress states and allows an effective deletion of the participant specific data without complete retraining. The experiments have shown that it is possible to keep the high level of classification and at the same time comply with the regulations [6].

Contributions:

The major findings of the present work are as follows:

- We present a privacy-sensitive hybrid CNN–Transformer architecture designed to be used in EEG based stress classification, introducing selective machine-unlearning, and thus making it possible to eliminate the participant-specific data without negatively affecting the performance of the model.
- We work on a complete EEG preprocessing pipeline based on band-pass filters, Independent Component Analysis (ICA), signal normalization, and class-balancing measures, which enhances signal quality and offers structured inputs to deep-learning models.
- We introduce an efficient selective machine-unlearning approach by training a CNN, which involves freezing feature-extraction layers, Transformer encoder, and classification layers, making it possible to remove training samples without re-training the entire model.
- The intended framework has been tested in an extensive experiment where the framework was found to have a high stress-classification accuracy (96.59%) among three stress conditions, including Low, Moderate, and High stress, after the unlearning process.

2. Related Work

Stress recognition has evolved at a rapid pace through EEG based techniques as a result of the growing need to objectively measure stress in brain-computer interface systems, safety in the workplace, and driver fatigue detection and adaptive human-computer interaction environments [7], [8]. Stress-related disorders affect cognitive functioning, productivity and overall well-being directly through chronic stress, anxiety and depression [9], [10]. As a result, automated EEG solutions based on deep learning methods have been developed to record discriminative spatial and temporal patterns of stress responses.
The wide usage of CNNs in EEG-based stress classification is due to their powerful abilities of deriving both spatial and spectral representations of neural signals. Some researchers have been able to demonstrate the usefulness of CNN structures in stress detection, but subject-independent classification is still a major issue [7],

[11]. Hybrid models like CNN-RNN [8] and CNN-LSTM [12] have been suggested to better be able to model temporal dependencies in EEG data. These models combine space feature extraction with time sequence modeling, thus enhancing classification accuracy and generalization across subjects in a better way.

More recently, attention mechanisms in Transformer-based architectures have generated interest in EEG analysis because they are able to model long-range temporal dependencies. Large-scale pre-training and transformer models have already been explored and are proved to be more effective and generalize better in EEG-based stress recognition research studies [13], [14]. Irrespective of the progress, most of the deep learning algorithms aim at improving the classification accuracy, without paying much attention to privacy-sensitive measures to handle sensitive EEG signals.

The issue of privacy is urgent due to the need to protect the confidentiality of cognition and neural responses represented by the electroencephalographic (EEG) signals that might reveal individual traits [15]–[17]. Regulatory frame-works, in particular, the General Data Protection Regulation (GDPR) require mechanisms that support the safe excision of personally identifiable data when requested to do so. In such a regulatory setting, machine unlearning has become an interesting and potentially useful technique, allowing specific data influences to be removed from trained models without requiring retraining exhaustion to do so [18]–[23].

Several recent studies have begun to combine EEG-based analysis directly with privacy-preserving learning paradigms. Shao et al. proposed a feature-unlearning method for EEG-based seizure prediction that removes a target patient's influence from a trained model while preserving cross-patient performance [32]. Umair et al. combined a lightweight hybrid-fusion EEGNetv4 architecture with federated learning for EEG-based dementia classification, keeping raw recordings on local clinical sites while reaching 97.1% accuracy [33]. Gupta et al. proposed a cross-silo federated multimodal system for major depressive disorder identification from audio and EEG that similarly avoids centralizing raw biosignals [34]. These studies confirm growing interest in privacy-aware EEG analysis, but, unlike the present work, none of them combine a CNN–Transformer hybrid architecture with a selective, participant-level unlearning mechanism for EEG-based stress classification.

Despite these advances, published models of EEG-based stress recognition place much emphasis on predictive accuracy in many cases at the cost of privacy protection. To address this weakness, this paper combines convolutional neural network-transformer hybrid architecture with selective feature unlearning, which makes it easier to apply stress recognition with privacy-preserving properties to EEG data. The proposed framework achieves better classification performance and at the same time, it enables regulation-compliant data removal mechanisms in Table I.

TABLE I: EEG-Based Stress Recognition Models

Ref.	Summary	Methodology	Input	Hybrid	Acc. (%)
[11]	Used CNN with 3D convolutional layers for spatiotemporal EEG features.	CNN (3D)	EEG	No	88.5
[14]	Introduced GPT-style pretraining for EEG, fine-tuned for stress detection.	Transformer	EEG	No	-
[7]	CNNs for EEG stress classification, highlighting spatial/spectral features.	CNN	EEG	No	87
[8]	CNN-RNN hybrid integrating spatial and temporal EEG features.	CNN-RNN	EEG	Yes	90
[12]	CNN–LSTM hybrid for spatial-temporal EEG features, subject-independent.	CNN–LSTM	EEG	Yes	91
[13]	Survey of Transformer-based models, Advocating CNN–Transformer hybrids.	Transformer	EEG	No	-
Proposed Method	privacy-preserving CNN–Transformer with selective feature unlearning.	CNN–Transformer + Unlearning	EEG	Yes	96.59

3. Methodology

EEG data were obtained in a group of 40 healthy participants (average age 21.5 years; 14 female subjects) using a 32-channel Emotiv EPOC Flex Gel Kit in compliance with the international 10–20 system of electrode placement [24]. Sampling of the signals was done at 128Hz and 256Hz so that the canonical EEG frequency bands could be covered and they include Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma (30–50 Hz). The levels of stress were coded using a one-hot scheme, i.e., Low, Moderate, and High stress were modeled by the vectors: $[1, 0, 0]$, $[0, 1, 0]$, and $[0, 0, 1]$, respectively. The prevailing EEG features of every stress level are presented in Table II. The characteristics of high stress are high variances, increased entropy, high FFT spikes, low inter-channel correlation, high Beta and Gamma activity with a low Alpha activity. The moderate stress is characterized by the moderate variance, steady statistical moments, even entropy and mixed Alpha/Beta activity. Low stress (relaxed or calm conditions) usually display less variance, organized covariance, constant spectral energy and increased Alpha activity.

The architecture of the CNN–Transformer model shown in Figure 1 will be used to classify the degree of stress based on the EEG signal. Convolutional layers provide spatial and spectral characteristics of the multichannel recordings, and transformer layers provide temporal characteristics based on the attention mechanisms, which allows the correct recognition of the stress level. Also, an optional machine-unlearning process is followed to maintain the privacy of users because it can remove data related to the participants but does not need to reset the entire model.

The CNN–Transformer hybrid was chosen over CNN-only or Transformer-only alternatives because the two components address complementary weaknesses. CNN-only architectures extract strong spatial and spectral features from the multichannel EEG grid but rely on a limited receptive field, which restricts their ability to capture dependencies that span an entire epoch or extend across epochs. Transformer-only architectures capture such long-range temporal dependencies through self-attention but typically need larger training sets to learn spatial structure without a convolutional inductive bias, which is a practical constraint given the SAM-40 dataset's 40 participants [24]. Combining convolutional feature extraction with a Transformer encoder lets the model exploit the spatial inductive bias of CNNs while still modeling the long-range temporal attention patterns associated with sustained stress responses.

EEG signals are represented as a matrix (see Equation 1):

$$X \in \mathbb{R}^{C \times N} \tag{1}$$

Here, the variable C denotes the number of recording channels, and N is used to denote the number of temporal observations. Each matrix value, therefore, is the amplitude of channel c at the time instant numbered n and this tabular structure forms the basis substrate of all subsequent preprocessing operations and feature-extraction operations. In order to enhance the fidelity of the electrophysiological recordings, the raw EEG data are then band-pass filtered, the spectral ranges that are most relevant to stress related neurophysiological patterns being filtered out. (Equation 2):

$$X_{\text{filtered}}(c, n) = \sum_{k=0}^{K} h_k X(c, n-k) \tag{2}$$

In this case, h_k is the filter coefficients and K is the filter length. The resulting filtering process isolates the Delta, Theta, Alpha, Beta, and Gamma frequency bands which are correlated to neural responses to

diverse stress conditions.
This is followed by Independent Component Analysis (ICA) that is used to decompose neural activity and artifacts in (Equation 3):

$$X_{\text{filtered}} = AS, X_{\text{clean}} = AS_{\text{clean}}, S_{\text{clean}} = S \setminus S_{\text{artifact}} \qquad (3)$$

In which, the mixing matrix A is represented, S is a set of independent components, and S_{clean} is free of artifact-related components.
This is then followed with normalization procedure to reduce inter-subject variation (Equation 4 and 5):

$$X_{\text{norm}}(c,n) = \frac{X_{\text{clean}}(c,n) - \mu_c}{\sigma_c} \qquad (4) \quad ,$$

$$\mu_c = \frac{1}{N}\sum_{n=1}^{N} X_{\text{clean}}(c,n), \sigma_c = \sqrt{\frac{1}{N}\sum_{n=1}^{N} (X_{\text{clean}}(c,n) - \mu_c)^2} \qquad (5)$$

The normalized signals are divided into epochs so as to define the time patterns in (Equation 6):

$$X_{\text{epoch}}^{(i)} = X_{\text{norm}}[:,(i-1)L+1:iL] \qquad (6)$$

where L is the epoch length and i is the epoch index.
To address class imbalance in training, only the training epochs are adjusted (Equation 7):

$$X_{\text{balanced}}^{(k)} = X_{\text{epoch}}^{(k)} \left\lceil \frac{n_{max}}{n_k} \right\rceil \qquad (7)$$

where n_k is the epoch count in class k and $n_{\max}$ is the maximum epoch count across all training classes. The EEG inputs are clean, normalized, and balanced which helps the hybrid CNN–Transformer model learn the temporal, spectral, and connectivity patterns of various stress levels. The steps of preprocessing described in (Equations 1–7) enable machine unlearning, in the sense of keeping the model performance intact while removing data of certain participants.

Feature Extraction:

Time-domain features: capture certain statistical measures on the EEG signals such as the mean, standard deviation, skewness, and kurtosis (Equations 8–9). These features help capture the EEG signal amplitude at a given time and help describe the statistical behavior of the neural signal at certain time intervals of stress state.

$$\mu_c = \frac{1}{N}\sum_{n=1}^{N} X_{\text{epoch}}(c,n), \sigma_c = \sqrt{\frac{1}{N}\sum_{n=1}^{N} \left(X_{\text{epoch}}(c,n) - \mu_c\right)^2} \qquad (8)$$

$$\mu_c = \frac{1}{N}\sum_{n=1}^{N} X_{\text{epoch}}(c,n), \sigma_c = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(X_{\text{epoch}}(c,n)-\mu_c\right)^2}$$

where μ_c represents the mean amplitude of the EEG signal for channel c, indicating the average neural activity, while σ_c represents the standard deviation, measuring signal variability around the mean.

$$\gamma_{1,c} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{X_{\text{epoch}}(c,n)-\mu_c}{\sigma_c}\right)^3$$

$$\gamma_{2,c} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{X_{\text{epoch}}(c,n)-\mu_c}{\sigma_c}\right)^4 - 3 \tag{9}$$

Here, $\gamma_{1,c}$ denotes skewness, which measures the asymmetry of the signal distribution, while $\gamma_{2,c}$ represents kurtosis, describing the peakedness or tail behavior of the EEG signal distribution.

Frequency-domain features: are computed using the Discrete Fourier Transform (DFT) and Power Spectral Density (PSD) (Equation 10). These features capture the spectral energy distribution of EEG signals across different frequency bands associated with neural oscillations.

$$X_c(f) = \sum_{n=0}^{N-1} X_{\text{epoch}}(c,n)e^{-j2\pi fn/N}$$

$$PSD_c(f) = |X_c(f)|^2$$

$$P_{\text{band},c} = \sum_{f\in \text{band}} PSD_c(f) \tag{10}$$

where $X_c(f)$ represents the Fourier transform of the EEG signal for channel c, converting the time-domain signal into the frequency domain. $PSD_c(f)$ measures the power distribution across frequencies, and $P_{\text{band},c}$ denotes the total spectral power within a specific EEG frequency band (Delta, Theta, Alpha, Beta, or Gamma), reflecting neural oscillatory activity related to stress levels.

Connectivity features: capture inter-channel relationships using covariance and Pearson correlation (Equation 11). These features characterize functional connectivity between different brain regions.

$$\Sigma = \frac{1}{N-1}\left(X_{\text{epoch}} - \mu_X\right)\left(X_{\text{epoch}} - \mu_X\right)^\top$$

$$r_{ij} = \frac{\text{cov}\left(X_{\text{epoch},i}, X_{\text{epoch},j}\right)}{\sigma_i\sigma_j} \tag{11}$$

In this formulation, Σ represents the covariance matrix describing the joint variability between EEG channels, while r_{ij} denotes the Pearson correlation coefficient between channels i and j, indicating the

strength of functional connectivity between different brain regions.
By integrating time-domain, frequency-domain, and connectivity features, each EEG epoch is represented by a rich multidimensional feature set that serves as input to the CNN–Transformer model.

Hybrid CNN–Transformer Model:

The features from the EEG epochs are input to a hybrid CNN–Transformer model after a feature extraction step. While local spatial and spectral patterns in the multi-channel EEG data are captured by the CNN layers, long-range temporal dependencies are modeled by the Transformer layers via attention. CNN feature extraction (Equation 12):

$$F_{\text{CNN}} = f_{\text{CNN}}\left(X_{\text{epoch}}\right) \tag{12}$$

where Xepoch is the input EEG epoch, and fCNN$(\cdot)$ is the function representing the convolutional layers responsible for capturing spatial-spectral feature representations.
Transformer attention mechanism (Equation 13):

$$F_{\text{Trans}} = \text{Attention}(F_{\text{CNN}})$$

$$\alpha_{t,c} = \frac{\exp\left(e_{t,c}\right)}{\sum_{t',c'} \exp\left(e_{t',c'}\right)} \tag{13}$$

$$e_{t,c} = \text{score}(q_t, k_c)$$

Selective Machine Unlearning:

To comply with privacy and other regulatory issues, selective machine unlearning is used to erase the impact of certain participant data while maintaining the quality of the model. The principle is to keep the knowledge acquired from the other data, and unlearn the contribution of the chosen samples from the trained model to the final model.

Privacy, legal, and regulatory constraints may require specific participant information to be omitted in EEG-based stress recognition systems. To avoid unnecessarily high computational costs, an entire model does not need to be retrained in such scenarios. Instead, selective machine unlearning gives the benefit to the model of forgetting specific samples while retaining general EEG representations from other participants. This involves removing the gradients of the sample in question, freezing the layers that generalize to EEG (CNN layers), and retraining the other layers (Transformer and classifier) through data retention.

In the experiments reported here, the forget set S_u corresponded to the complete EEG recordings of 4 of the 40 participants (10% of participants), drawn at random from the training partition without regard to class label so that the class balance of the retained data was preserved. After gradient removal (Equation 14) and CNN-layer freezing (Equation 15), the Transformer encoder and classification head were fine-tuned on the remaining 36 participants' data for 15 epochs using the Adam optimizer with a reduced learning rate of 1×10^{-4} (one-tenth of the original training rate), which was sufficient for the validation loss to re-stabilize before the unlearning-validation step (Equation 17).

Step 1: Gradient removal

The parameters of the model in question have the impact of the sample in question subtracted from them:

$$\theta \leftarrow \theta - \eta \sum_{i \in S_u} \nabla_\theta \mathcal{L}(x_i, y_i) \quad (14)$$

where S_u denotes the set of samples to be unlearned, L represents the loss function, and η is the learning rate. Equation 14 removes the influence of these samples from the learned model parameters.

Step 2: CNN freezing

The general spatial–spectral EEG representations have their CNN layers frozen in Equation 15.

$$\theta_{\text{CNN}} \leftarrow \text{constant} \quad (15)$$

This prevents the catastrophic forgetting of the features learned from the full dataset.

Step 3: Transformer and classifier fine-tuning

The Transformer and classifier (fully connected) layers are fine-tuned using only the data that has been retained:

$$\theta_{\text{Trans}}, \theta_{\text{FC}} \leftarrow \theta_{\text{Trans}}, \theta_{\text{FC}} - \eta \sum_{i \notin S_u} \nabla_\theta \mathcal{L}(x_i, y_i) \quad (16)$$

In Equation 16, the aim is to restore the model's performance and not to account for the influence of the removed samples.

Step 4: Unlearning validation

In order to confirm selective unlearning has not negatively affected global model performance, the predictions before and after unlearning are compared:

$$\Delta y = \left| f_\theta(X_{\text{test}}) - f_{\theta_{\text{unlearned}}}(X_{\text{test}}) \right| \quad (17)$$

A small value of Δy indicates that the model has successfully removed the influence of the targeted data while maintaining overall predictive performance.

Selective machine unlearning therefore provides an efficient and privacy-compliant mechanism for removing specific participant information from EEG-based stress recognition models. By combining gradient removal (Equation 14), CNN layer freezing (Equation 15), Transformer/classifier fine-tuning (Equation 16), and validation (Equation 17), the proposed approach ensures that the model retains general EEG knowledge while forgetting sensitive participant data.

PERFORMANCE EVALUATION:

The performance of the model is measured by the conventional metrics of multi-class classification. Let C denote the number of classes (Low, Moderate, High), and TP_i, FP_i, and FN_i denote the True Positive, False Positive, and False Negative counts for class i, respectively. The mentioned metrics quantify the ability of the model to separate the varying degrees of stress.

Accuracy (Equation 18):

Accuracy indicates the overall ratio of correct predictions against all predictions made. It offers an overarching assessment of the model's ability to classify all the stress categories.

$$\text{Accuracy} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} (TP_i + FP_i + FN_i)} \tag{18}$$

Precision (Equation 19):

Each class has a corresponding level of precision which defines the accuracy of the positive predictions made. It answers the question of how many samples that were predicted to be at a given level of stress are indeed at that level.

$$\text{Precision}_{\,i} = \frac{TP_i}{TP_i + FP_i} \qquad (19)(19)$$

Recall (Equation 20):

Recall assesses the model's proficiency at recognizing samples that belong to a particular class. It reflects the model's ability to identify all pertinent instances relating to each level of stress.

$$\text{Recall}_{\,i} = \frac{TP_i}{TP_i + FN_i} \tag{20}$$

F1-Score (Equation 21):

The F1-score takes precision and recall and combines the two metrics using the harmonic mean. This evaluation is helpful for classification tasks in which the class distribution is not balanced.

$$F1_i = 2 \cdot \frac{\text{Precision}_{\,i}\ \text{Recall}_{\,i}}{\text{Precision}_{\,i} + \text{Recall}_{\,i}} \tag{21}$$

Confusion Matrix (Equation 22):

The confusion matrix is a more detailed illustration of classification results for all levels of stress. Each row shows the actual class. Each column shows the predicted class. This structure allows the investigation of patterns for misclassification for the Low, Moderate, and High stress classes.

$$CM = \begin{bmatrix} CM_{\text{Neg,Neg}} & CM_{\text{Neg,Neu}} & CM_{\text{Neg,Pos}} \\ CM_{\text{Neu,Neg}} & CM_{\text{Neu,Neu}} & CM_{\text{Neu,Pos}} \\ CM_{\text{Pos,Neg}} & CM_{\text{Pos,Neu}} & CM_{\text{Pos,Pos}} \end{bmatrix} \tag{22}$$

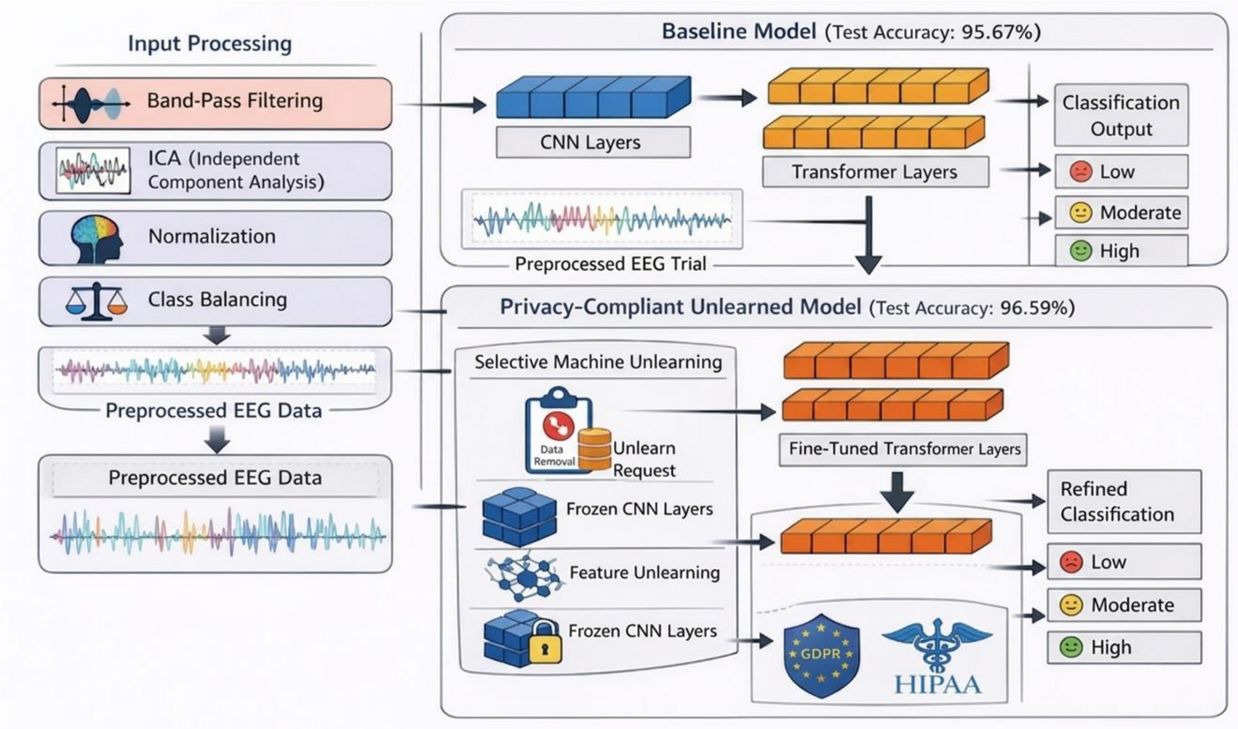


Fig. 1: CNN–Transformer Based Privacy-Preserving EEG Stress Classification Framework

4. EXPERIMENTAL SETUP

The electrode placement was done according to international 10–20 electrode placement protocol and EEG records were recorded in 32 channels of the Emotiv EPOC Flex system, comprising of forty healthy subjects. Sampling of the signals was done at 128 Hz and 256 Hz to sample all the canonical EEG frequency bands which ranged between Delta and Gamma. Recording signals were subjected to band-pass filtering, Independent Component Analysis (ICA) and normalization preprocessing to remove artifactual contamination and suppress inter-subject variability.

After preprocessing, the features included time-domain statistics, power spectral density (PSD) and covariance-based connectivity measures were obtained and were used as inputs to the suggested CNN–Transformer architecture with selective machine unlearning. The tests were conducted in a workstation with an NVIDIA RTX 3090 graphics card and 64 GB of RAM with the help of the PyTorch deep learning system. Established multi-class classification performance metrics were used to evaluate the model performance, such as Accuracy, Precision, Recall, F1-score and Confusion Matrix analysis. A summary of the detailed experimental setup and computing resources is provided in Table III.

To ensure a realistic estimate of generalization to unseen individuals, the 40 participants were partitioned at the subject level into training (32 participants, 80%), validation (4 participants, 10%), and test (4 participants, 10%) sets, with no participant's recordings appearing in more than one partition. This subject-independent protocol was adopted in preference to a subject-dependent (within-participant) split, which allows epochs from the same participant to appear in both training and test data and tends to overestimate classification performance as a result.

TABLE III: Experimental Setup and Computational Resources

Parameter	Specification

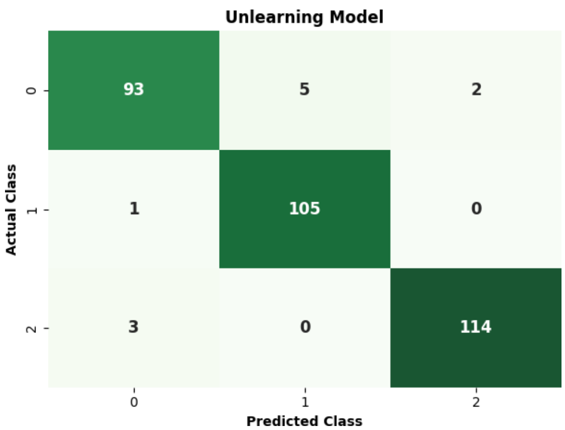
Participants	40 (14 females, mean age 21.5)
Data Split	32 / 4 / 4 participants (80% / 10% / 10%), subject-independent
EEG Device	32-channel Emotiv EPOC Flex
Electrode Placement	International 10–20 system
Sampling Rate	128 Hz, 256 Hz (Delta–Gamma bands)
Classes	Low, Moderate, High
Preprocessing	Band-pass filtering, ICA, normalization
Features	Time-domain, PSD, covariance
Model	CNN–Transformer
Unlearning Strategy	CNN frozen, Transformer fine-tuned
Unlearning Set Size	4 of 40 participants (10%), selected at random from the training partition
Unlearning Fine-tuning Epochs	15 epochs (Adam, LR = 1e-4)
Hardware	RTX 3090 GPU, 64 GB RAM
Framework	PyTorch
Training Setup	Batch size = 64, Epochs = 100, Adam optimizer (LR = 0.001)
Weight Initialization	He initialization (CNN layers); Xavier initialization (Transformer/classifier)
Early Stopping	Patience = 10 epochs on validation loss
Hyperparameter Selection	Grid search over LR {1e-3, 5e-4, 1e-4} and batch size {32, 64, 128} on the validation split
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, Confusion Matrix

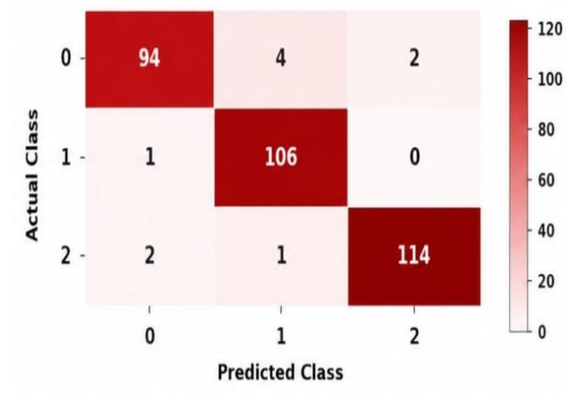
5. Results and Discussions

The stress recognition CNN–Transformer framework proposed was tested in two experimental conditions, the Original (Baseline) Model and the Privacy-Compliant Unlearning Model. The standard evaluation metrics that were used to evaluate model performance included accuracy, precision, recall, F1-score, and confusion matrix analysis. The overall test accuracy of the baseline model was 95.67% and the unlearning model had a slightly better score of 96.59%. This can be attributed to the fact that only noisy or less informative samples can be removed under the unlearning process. EEGs often have artifacts or individual noise that may have negative implications on model generalization. With the removal of the effects of such samples, this model focuses on more stable neural patterns related to stress states. Consistent with this interpretation, the four unlearned participants' recordings showed, on average, higher intra-class variance and a higher proportion of ICA-flagged artifactual epochs than the retained set. In addition, the fine-tuning step in unlearning also allows the model to re-optimize the stored representations, which contributes to the performance improvement seen. These findings demonstrate that selective machine unlearning preserves predictive power, but preserved accuracy alone is not formal proof that the targeted participants' influence was fully removed from the model's parameters. The small Δy values obtained from Equation 17 are consistent with successful forgetting, but a stronger guarantee would require complementary auditing, such as membership-inference testing on the unlearned samples, which we identify as a direction for future work. Recent studies on EEG-based stress recognition demonstrate strong performance using both transformer-based and hybrid deep learning models. For example, CognEmoSense, which utilizes a transformer architecture, achieved 88% accuracy on the DEAP dataset [25], while EmoSTT reported 92.67% accuracy on the SEED dataset [26]. These findings highlight the effectiveness of attention mechanisms in capturing long-range temporal dependencies in EEG signals. In parallel, hybrid deep learning approaches have also shown strong performance in stress classification tasks.

The Brain2Vec model, which integrates CNN–LSTM–Attention mechanisms, achieved 81.25% accuracy on the DEAP dataset [27]. Similarly, CNN–LSTM based models have demonstrated improved performance compared to traditional machine learning methods such as Support Vector Machines and Random Forest, with reported accuracies typically ranging between 85% and 91%, depending on dataset characteristics and preprocessing techniques [28, 29, 30, 31]. Overall, these studies indicate that both transformer-based and hybrid architectures are effective for EEG-based stress analysis. Transformer-based models generally achieve higher performance in capturing complex temporal dependencies, while CNN–LSTM-based models provide stable and consistent results due to their ability to learn both spatial and temporal EEG features. However, differences in datasets and evaluation settings across studies limit direct comparability of results.

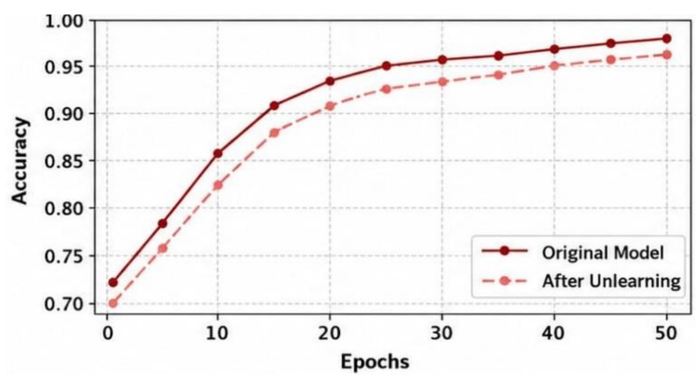
The confusion matrices analysis gives supportive evidence to the reliability of proposed model. The baseline setup (Fig. 2a) is characterized by a high degree of diagonal dominance, which implies that most of the samples by the three stress classes are accurately classified. Simultaneously, the unlearning setup (Fig. 2b) has a class-wise performance that is either equal or slightly better than the one of the baseline, with insignificant misclassifications. These results testify that the boundaries of the choices of the trained network are not significantly disturbed by selective unlearning of information. The supplementary evaluation metrics also support the strength of the model. The precision, the recall, and the F1-scores were estimated at about 0.97 at the baseline configuration, and the scores were also relatively high at 0.967, 0.96, and 0.96, respectively in the unlearning configuration (Fig. 2d).

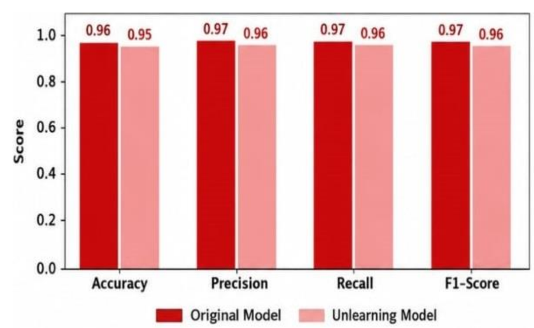




(a) Original Model Confusion Matrix
Matrix

(b) Unlearning Model Confusion





(c) Accuracy vs Epochs (Original vs Unlearning) score Comparison

(d) Accuracy, Precision, Recall, and F1-

Fig. 2: Performance evaluation of the proposed framework: (a) Confusion matrix of the original model, (b) Confusion matrix of the unlearning model, (c) Accuracy versus epochs comparison, and (d) Performance metrics comparison between original and unlearning models.

The constantly high values are the signs of a harmonious predictive performance and reduced rates of false positives and false negatives. Accuracy-versus-epoch plots (Fig. 2c) were used to study the temporal change of training. The two configurations achieve steady and progressive convergence and the unlearning configuration is similar to the baseline trajectory and has a slightly better final accuracy. These observations verify that selective unlearning provides an efficient re-optimization of the learned information and maintenance of stability during training.

Overall, the obtained empirical results validate the notion that the CNN–Transformer framework provides a skillful atonement between the high performance in classification and privacy-preserving features. The consistency of the confusion matrices, evaluation measures, and learning dynamics indicates that the model is capable of maintaining discriminative EEG representations while removing sensitive information where needed.

6. Conclusion

This paper proposed a privacy-sensitive CNN–Transformer architecture on EEG-based stress recognition that also uses selective machine-unlearning. The architecture combines both convolutional layers which extract spatial spectral features with transformer layers which extract temporal dependencies, which allow it to correctly classify Low, Moderate, and High stress states. The experimental assessment proves good results as the baseline model is 95.67 percent accurate, whereas, after selective unlearning, the model is 96.59% accurate. These results suggest that the deletion of data at specific points can be implemented without reducing the accuracy of the model. Deploying this framework at scale raises practical considerations beyond the classification accuracy reported here. The Transformer component adds computational and memory overhead relative to CNN-only baselines, which may constrain use on low-power or edge devices commonly used for continuous stress monitoring. Scaling the selective-unlearning procedure to larger cohorts will also require revisiting the fine-tuning epoch budget and forget-set size used here, since unlearning cost grows with the number of retained samples that must be fine-tuned on. Real-world healthcare deployments would additionally benefit from combining this mechanism with a formal auditing procedure, such as membership-inference testing, to provide stronger guarantees of data removal than accuracy preservation alone. Therefore, the suggested framework provides a viable approach to implementing EEG Based stress recognition systems in privacy sensitive areas including health care and affective computing. Future research will focus on the use of bigger datasets and the construction of adaptive unlearning approaches in order to improve model robustness and generalizability further.

7. References

[1] Z. Huang, Y. Ma, R. Wang, W. Li, and Y. Dai, "A model for eeg-based emotion recognition: Cnn-bi-lstm with attention mechanism," *Electronics*, vol. 12, no. 14, p. 3188, 2023.

[2] N. Simić, S. Suzić, N. Milošević, V. Stanojev, T. Nosek, B. Popović, and A. Bajović, "Enhancing emotion recognition through federated learning: A multimodal approach with convolutional neural networks," *Applied sciences*, vol. 14, no. 4, p. 1325, 2024.

[3] M. A. Becerra, C. Duque-Mejia, A. Castro-Ospina, L. Serna-Guarín,

C. Mej´ıa, and E. Duque-Grisales, “Eeg-based biometric identification and emotion recognition: An overview,” *Computers*, vol. 14, no. 8, p. 299, 2025.

[4] Yusra and R. UlAmin, “A privacy-aware federated hybrid model for multimodal mental health analysis,” *International Journal of Advanced Computer Science and Applications*, vol. 17, no. 5, 2026. [Online]. Available: https://doi.org/10.14569/IJACSA.2026.0170598

[5] Y. Ali and R. U. Amin, “Hybrid deep learning framework for eeg-based anxiety detection and classification in brain–computer interfaces,” *Journal of Computers and Intelligent Systems*, vol. 4, no. 1, pp. 14–21, 2026. [Online]. Available: https://journals.iub.edu.pk/index.php/JCIS/article/view/4571

[6] J. Freeman, M. L. Yell, J. G. Shriner, and A. Katsiyannis, “Federal policy on improving outcomes for students with emotional and behavioral disorders: Past, present, and future,” *Behavioral Disorders*, vol. 44, no. 2, pp. 97–106, 2019.

[7] A. Acharya *et al.*, “Eeg-based stress classification using convolutional neural networks,” *Neurostressology*, vol. 12, pp. 45–56, 2025.

[8] J. Hyung *et al.*, “Cnn-rnn hybrid network for eeg-based stress detection,” *Journal of Affective Computing*, vol. 18, pp. 102–115, 2025.

[9] L. Zhao *et al.*, “Impact of chronic stress on cognitive function,” *Journal of Cognitive Health*, vol. 14, pp. 1–12, 2025.

[10] K. Yu *et al.*, “Stress and well-being: Neural and behavioral insights,” *Frontiers in Psychology*, vol. 16, pp. 101–115, 2025.

[11] S. Kim *et al.*, “3d cnn for spatiotemporal feature extraction in eeg-based stress recognition,” *Computational Neuroscience*, vol. 10, pp. 201–214, 2022.

[12] L. Wang and R. Duan, “Cnn-lstm hybrid model for subject-independent eeg stress detection,” *Machine Brain-Computer Interfaces*, vol. 7, pp. 33–45, 2025.

[13] M. Vafaei *et al.*, “Transformer-based models for eeg stress recognition: A survey,” *IEEE Transactions on Neural Systems*, vol. 35, pp. 101–120, 2025.

[14] P. Lloyd, “Gpt-style pretraining for eeg-based stress detection,” *Neuroinformatics*, vol. 22, pp. 77–90, 2024.

[15] M. Ienca and G. Malgieri, “Ethics and privacy of eeg data in cognitive computing,” *Nature Machine Intelligence*, vol. 4, pp. 2–9, 2022.

[16] Y. Zhang *et al.*, “Eeg-based biometric authentication: Privacy and security review,” *IEEE Access*, vol. 9, pp. 42 000–42 015, 2021.

[17] F. Akbarnia *et al.*, “Privacy risks in eeg-based identification systems,” *Journal of Neural Engineering*, vol. 21, p. 046002, 2024.

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

[19] L. Bourtoule *et al.*, “Machine unlearning,” in *IEEE Symposium on Security and Privacy (SP)*, 2019, pp. 371–388.

[20] C. Guo *et al.*, “Certified data removal from machine learning models,” *Proceedings of the ACM on Privacy and Security*, vol. 3, pp. 1–28, 2020.

[21] X. Liu *et al.*, “Survey on machine unlearning for deep neural networks,” *Journal of*

Intelligent Secure AI, vol. 5, pp. 12–30, 2025.
[22] A. Koloskova *et al.*, "Certified unlearning for deep neural networks," *Neural Networks*, vol. 150, pp. 120–135, 2025.
[23] H. Yang, "Gdpr-compliant machine learning pipelines with integrated unlearning," *Ethics in Technology and Data*, vol. 8, pp. 55–70, 2024.
[24] 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.
[25] H. Dalkic, "Cognemosense: A continual learning and context-aware eeg emotion recognition system using transformer-augmented brain-state modeling," *Journal of Brain Science and Mental Health*, vol. 1, no. 1, pp. 01–09, 2025.
[26] M. Li, P. Yu, and Y. Shen, "A spatial and temporal transformer-based eeg emotion recognition in vr environment," *Frontiers in Human Neuroscience*, vol. 19, p. 1517273, 2025.
[27] M. Mynoddin, T. Dev, and R. Chakma, "Brain2vec: A deep learning framework for eeg-based stress detection using cnn-lstm-attention," *arXiv preprint arXiv:2506.11179*, 2025.
[28] K. V. Reddy, S. T. Dharavath, N. Koushik, and R. S. Kumar, "Deep eeg classifier: Cnn-lstm approach for mental state recognition," *REST Journal on Data Analytics and Artificial Intelligence*, vol. 4, no. 1, pp. 352–359, 2025.
[29] 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
[30] Amin, R.U., 2026. Hybrid Deep Learning Framework for EEG-Based Anxiety Detection and Classification in Brain–Computer Interfaces. *Journal of Computers and Intelligent Systems*, *4*(1), pp.14-21.
[31] Yusra, & UlAmin, R. (2026). Regulatory-Compliant EEG Stress Recognition Using CNN–Transformer Models with Feature Unlearning. International Journal of Advanced Computer Science and Applications, 17(7). https://doi.org/10.14569/IJACSA.2026.0170788
[32] C. Shao, C. Li, R. Song, G. Xu, and X. Chen, "Feature unlearning for EEG-based seizure prediction," IEEE Internet of Things Journal, vol. 12, pp. 10974–10986, 2024.
[33] M. Umair, M. S. Khan, M. Hanif, W. Ghaban, I. Nafea, S. N. Qasem, and F. Saeed, "Privacy-preserving dementia classification from EEG via hybrid-fusion EEGNetv4 and federated learning," Frontiers in Computational Neuroscience, vol. 19, Art. 1617883, 2025.
[34] C. Gupta, V. Khullar, N. Goyal, K. Saini, R. Baniwal, S. Kumar, and R. Rastogi, "Cross-silo, privacy-preserving, and lightweight federated multimodal system for the identification of major depressive disorder using audio and electroencephalogram," Diagnostics, vol. 14, no. 1, p. 43, 2024.