International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 7, Number 2, Year 2025
DOI: https://doi.org/10.62051/ijcsit.v7n2.02
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Application of Time Series Model LT-MAE in EEG Emotion Recognition

Jianhao Ma *

The School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454003, China

*Corresponding Author

ABSTRACT

Electroencephalography (EEG) signals are non-linear and non-stationary. Traditionally, they are segmented into time windows for feature extraction under the assumption of independence and identical distribution, ignoring temporal connections and distribution discrepancies. Additionally, generating high-quality annotations for dynamic emotions is labor-intensive and time-consuming. To address these issues, we propose LT-MAE, a self-supervised learning model. It segments EEG signals into continuous time steps and uses long short-term memory network (LSTM) to learn context representations across channels. Emotion distributions are learned using an enhanced mask autoencoder with classification and reconstruction tasks. This approach assesses emotional changes over continuous time steps to determine long-term emotional inclinations. Experiments on SEED-IV and DEAP datasets show that LT-MAE learns a broader time-step emotion distribution in the coding space, improving emotion detection accuracy and mitigating labeling inaccuracies due to finite-time granularity. Unlike traditional methods that assume independent and identically distributed data, LT-MAE captures temporal connections and distribution discrepancies. By leveraging LSTM and enhanced mask autoencoder techniques, it provides more accurate emotion recognition while reducing reliance on costly annotations. In conclusion, LT-MAE offers an effective solution to limitations in conventional EEG feature extraction. Using self-supervised learning, it enhances the understanding of temporal sentiments and improves emotion recognition accuracy, addressing challenges related to annotation quality and granularity.

KEYWORDS

Emotion recognition; Mutli-channel EEG signal; Self-supervised; Time series

1. INTRODUCTION

Emotion is a complex psychological and physiological state experienced by individuals in response to external stimuli. Positive emotions contribute to human well-being and productivity, while negative emotions may lead to health issues [1]. Emotion recognition has widespread applications across various industries, such as safe driving [2], mental health monitoring [3] and social security [4]. Accurate emotion identification has the potential to enhance people's quality of life. EEG is a non-invasively activated physiological signal less influenced by subjective human factors [5], making it wellsuited for emotional state classification.

In the context of processing Electroencephalography(EEG) sequence data point by point, temporal dependencies are captured by modeling each sample point separately. However, high-frequency time series often exhibit significant temporal redundancy among neighboring points, allowing information from one data point to be inferred from its neighbors. In order to extract the time-domain, frequency-

domain, time-frequency domain, and nonlinear dynamic characteristics of the time steps based on experience, the majority of current studies employ the method of time window segmentation to extract features of EEG signals. Liu [6] used a sliding window of 1 second to segment the baseline EEG signal into limited sub-segments, and based on the time-invariant assumption, the sub-segments followed the annotation of the entire EEG signal and recognized emotions through the attention mechanism and convolutional capsule network. In order to increase the amount of data, [7] used non-overlapping time windows with a length of 2 seconds to segment the original signal and extract differential entropy as a feature. The label is processed by the clustering method so that the model can recognize six kinds of emotions. [8] enhanced the data using a sliding window of 2 seconds and a step size of 0.5 seconds. 117 new samples are created from the 60-second EEG sample, and the labels of the original samples are passed down to the new samples. Additionally, construct an ERENet network using multi-band electroencephalogram topology maps to identify emotional states.

Deep learning can discover the distributional properties of complex data by combining shallow features to form abstract high-level features. Many complex deep learning networks, such as recurrent neural networks (RNNs), graph neural networks (GNNs), and convolutional neural networks (CNNs), have been applied to recognize emotions. [9] proposed a multidimensional graph convolutional network, MSFR-GCN, which exploits the asymmetry of neuron activity in the left and right brain for emotion recognition. [10] proposed a method that combines EEG signal Dynamic Brain Functional Network features for multidimensional emotion recognition, which introduces an attentional mechanism to fuse time-space, time-frequency, and frequency domain features. which are then transported to BiLSTM to recognize emotional states. [11] extracted features from three dimensions: spatial, frequency, and temporal. Non-Euclidean spatial data were processed using Graph Convolutional Network (GCN), key frequency information was captured by frequency band attention module, and deep temporal features were extracted by using cascading convolutional layers, which effectively realized multi-scale feature fusion.

Despite significant advancements in EEG-based emotion recognition, several challenges persist. Firstly, manual feature extraction can transform complex EEG signals into clear-cut time steps, thereby enhancing data granularity. However, finer-grained time steps necessitate finer-grained emotional annotations. However, finer-grained time steps necessitate finer-grained emotional annotations, but acquiring a large volume of accurately annotated time series data is difficult in real-world circumstances. Secondly, the assumption of independence overlooks the substantial contextual relationships among individual time steps within the time series. Additionally, the time-invariant hypothesis is not always satisfied in the time series due to the latency of emotion and the fact that emotion will change with time following activation. The time steps also don't always follow an independent and identical distribution because of many factors like subjects, time periods, collection environments, and more. Therefore, there is a need to develop more appropriate and effective models to address emotion classification tasks.

In response to the aforementioned problems, a new self-supervised model called LT-MAE for time series is proposed. Firstly, non-overlapping time windows are used to partition the EEG signal into continuous time steps and extract its initial features. The Long Short-Term Memory (LSTM) network is then employed to capture its contextual information. Subsequently, a random masking approach is applied to divide the data into visible and masked portions. The Masked Autoencoder (MAE) module is utilized to perform classification and reconstruction tasks, and the representation coding space is reconstructed to reflect the distribution of time steps under various emotional states. Finally, a mask encoder is used to map the training signal into the representation space as a whole, and EEG emotion recognition is conducted by analyzing emotional changes in consecutive time steps. The main contributions of this study include:

A self-supervised emotion recognition model, LT-MAE, is proposed to model complex EEG signals as training signals composed of multiple non-overlapping time steps to preserve their temporal order. Simultaneously, the dependence between the emotion labels of individual time steps and the overall

emotion label is eliminated, thereby alleviating the problem of limited emotion annotation in existing datasets.

An MAE architecture is designed to achieve self-supervised learning through two auxiliary tasks. The MAE encoder generates similar representation encodings for time steps with similar emotions, while simultaneously differentiating the representation encodings between different emotions. This allows the model to better understand the emotion distribution of time steps.

The MAE encoder maps continuous time steps to the representation encoding space in the downstream supervised emotion recognition task. By analyzing the emotional fluctuations over a period of time, determine the emotional tendency within those time steps. It is evident that this type of overall emotion recognition, which incorporates multiple time steps, is easier to understand and more aligned with the meaning of emotion labels in datasets that are widely utilized for emotion recognition.

2. RELATED WORK

2.1. Emotion Classification Overview

Psychologists employ two distinct techniques for modeling emotions. One approach is to categorize emotions, while another is to label emotions using many dimensions. [12] classified emotions into six separate and measurable states, while [13] proposed a wheel model comprising eight primary emotions: joy, trust, fear, surprise, sadness, disgust, rage, and anticipation. Discrete emotion models rely on verbal expressions of feelings rather than quantitative analysis, making it challenging to study complex emotions. [14] introduced a two-dimensional emotional space using valence and arousal, with varying valence and arousal levels representing different emotions. For instance, rage exhibits negative valence and high arousal, whereas sadness exhibits negative valence and low arousal. [15] expanded the emotional model from two dimensions to three, introducing an additional axis called dominance. Spanning from obedience to dominance, this axis reflects individuals' ability to control a particular emotion, addressing the problem of identifying similar emotions in the two-dimensional model. Dimensional emotion space facilitates the calculation of differences and connections between various emotions, although the emotions it represents may not be as intuitive and interpretable.

2.2. Feature Extraction

Power Spectral Density (PSD) and Differential Entropy (DE) are popular methods for evaluating EEG signals [16, 17] Power spectral density is a physical parameter that represents the connection between a signal's power and frequency. This distribution of energy across frequency bands can determine which bands the energy of the EEG signal is concentrated in, thus providing insights into the subject's mental state at that moment. For example, if the power spectral density of the EEG signal indicates that the energy is predominantly concentrated in the Delta band below 10 Hz, the patient may be relaxed, bored, or drowsy.

DE is the continuous variant of Shannon entropy, and it is used to assess the complexity of continuous random variables. Studies by [18] and [19] have shown that differential entropy features perform well in EEG emotion recognition tasks, which are defined as follows:

$$h(X) = -\int f(x) \log f(x)\, dx \tag{1}$$

For an EEG signal within a time window [a, b], if X follows a Gaussian distribution, its differential entropy is calculated as follows:

$$h(X) = \frac{1}{2} \log 2\pi e\sigma^2 \tag{2}$$

Differential entropy, if e and π serve as constants, can be approximated by the logarithm of the power spectral density for a fixed-length EEG signal sequence within a specific frequency band. However, unlike power spectral density, differential entropy is capable of more evenly discerning between low-frequency and high-frequency energy in EEG signals.

In the Electroencephalography emotion dataset $\mathbb{D} = [(\mathrm{X}^1, \mathrm{y}^1), (\mathrm{X}^2, \mathrm{y}^2), \cdots, (\mathrm{X}^n, \mathrm{y}^n)]$, $\mathrm{X}^i = [\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_T] \in \mathbb{R}^{m \times T}$ represents a multi-channel EEG time series composed of m electrode channels, denotes its corresponding labels, and the total number of samples in the dataset is n. Assuming the length of the time series is T and the size of the time window is set to σ a zero-fill alignment operation is performed on T to obtain a uniform length of $X' \in \mathbb{R}^{m \times \lceil T/\sigma \rceil}$. Non-overlapping time windows are then used to divide the original time series into consecutive time steps, ensuring that no duplicate information is contained in adjacent time steps $s_{i:j} = \{x_i, x_{i+1}, \cdots, x_{i+\sigma}\}$, as illustrated in Fig.1 Feature extraction is subsequently conducted within each time window to convert the continuous EEG signals into discrete time steps.

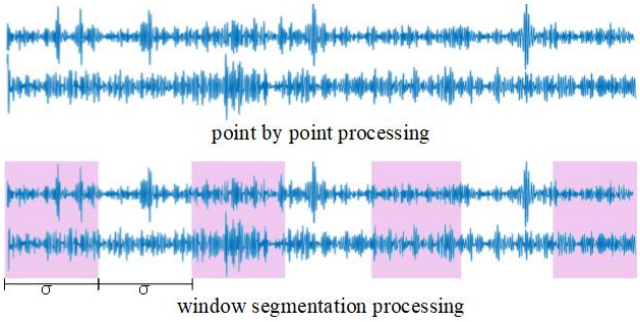


Figure 1. Two approaches of processing EEG data.

3. METHOD

3.1. Self-supervised Representation Learning

The LT-MAE model learning process is illustrated in Fig.2. It employs a double-layer LSTM to extract cross-channel contextual features $Z = [Z_1, Z_2, \cdots, Z_{\lceil T/\sigma \rceil}] \in \mathbb{R}^{d \times \lceil T/\sigma \rceil}$, where d represents the dimension of the hidden layer. To gain a deeper understanding of the training signal representation Z, a random mask ratio of v / m is utilized to create corrupted inputs, with Z_v and Z_m denoting the visible and shielded areas, respectively. Random masking ensures that the likelihood of being masked at each time step is consistent, thereby enhancing signal variety. Random masking ensures that each time step has an equal chance of being masked when creating the training signal, which enhances the signal's diversity. First, to maintain consistent sequence timing, a position code $\mathrm{P} \in \mathbb{R}^{d \times \lceil T/\sigma \rceil}$ is added to the training signal Z, resulting in Z=Z+P. Then, the viewable area Z_v is input into the L_v layer transformer module E_θ to obtain the global context representation $E_\theta^{L_v} = \{e_1^{L_v}, e_2^{L_v}, \cdots, e_v^{L_v}\}$ of Z_v. The encoder E_θ is an ordinary transformer module with multi-head attention and a feed-forward network. Subsequently, instead of self-attention, cross-attention is utilized in the regular transformer module to capture the representation of the masked region Z_m, forming a decoupled encoder module P_φ. Concurrently, a set of emotive code words for the masked area Z_m, denoted by $\tilde{Z}$, is generated by the tokenizer module T. Finally, these emotional code words can replace the masked region Z_m as the query for attention computation, with the visible position representation $E_\theta^{L_v}$ serving as input to assist P_ϕ in generating the corresponding keys and values. The decoupled encoder's m-th layer output $P_\phi^{L_m} = \{p_1^{L_m}, p_2^{L_m}, \cdots, p_m^{L_m}\}$ represents the context representationZ_{pred} of the masked region Z_m predicted by the emotional codeword. In the representation learning of the visible and masked areas, the encoder E_θ is solely responsible for learning the representation of the visible area, while

the decoupled encoder module P_φ exclusively predicts the representation encoding of the masked area, without altering the representation encoding of the visible area. This decoupling mechanism prevents the decoupled encoder module P_φ from learning the visible region, enabling the encoder E_θ to capture more meaningful data.

3.2. Emotional Codeword Classification Task

The window slicing procedure divides the entire time series into numerous small time steps, and feature extraction for each short time step can produce richer semantics. Inspired by product quantization, time steps can be quantized with a new discrete view; that is, time steps are assigned their own emotional codewords. These emotional codewords can be used as supervision signals to help address the problem of restricted emotional annotation. Fig.2 (a) depicts the process.

In order to achieve this, a tokenier module $\mathcal{T}$ is designed to assign the closest emotion code word to each time step. The assignment process is depicted in (3). The fundamental idea is to compute similarity. To do this, we create a codebook matrix $C \in R^{K\times d}$, which roughly represents the states of K different abstract emotions in the representation space. The similarity measure function $sim(\cdot)$ calculates the similarity between the time step z_i and the codebook vector c_j. The emotion codewords is the only one-hot code that has the best approximation of the codebook vector c_j subscript, and the emotion codeword set s_i is created by combining S.

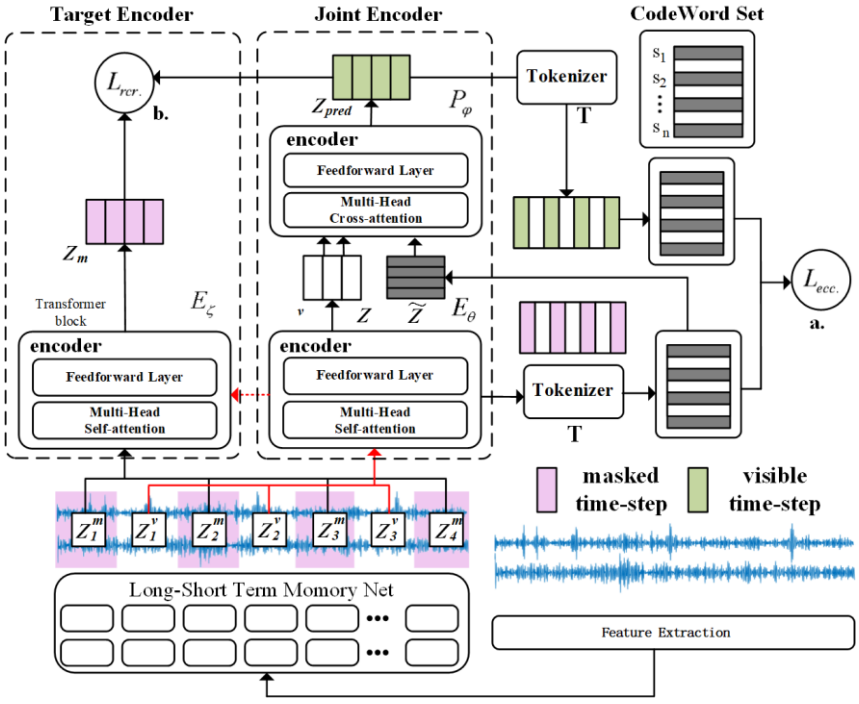


Figure 2. Architecture of the proposed LT-MAE model

$$s_i = onehot\left(\underset{j}{argmax}\, sim\left(z_i, c_j\right)\right), j \in K, c_j \in C \tag{3}$$

The Fig. 2(a) depicts the optimization process. Initially, a tokenizer assigns emotion codewords to the training signal by time steps, establishing the emotion code word distribution $q(s_i \mid z_i)$ of the training signal in the representation space. Next, the emotion code word within the masked area is selected as the supervisory signal, leading to the reconstruction of the emotion code word distribution

$p\left(s_i \mid f_i^{L_m}\right)$ within the masked area. Lastly, the emotion code word's classification task is optimized using the cross-entropy loss, whose loss is $L_{ecc.}$. It is able to be stated as:

$$L_{ecc.}(q, p) = -q\left(s_i \mid z_i^m\right) \log p\left(s_i \mid f_i^{L_m}\right) \tag{4}$$

3.3. Representation Coding Reconstruction Task

Because the complexity of the representation space in the time dimension is substantially lower than that of the original space, it is less sensitive to noise and can pay more attention to the important features in the original space, resulting in a superior reconstruction work. To produce an encoded representation of the masked region, we utilize an autoencoder module E_ζ that is analogous to E_θ. E_ζ adopts an identical hyperparameter configuration to E_θ and shares a substantial portion of the model parameters. For clarity, we designate the encoder ${E_\theta}$ in the visible area and the decoupled encoder P_φ as the Joint Encoder module, while the encoder module E_ζ in the masked area is denoted as the Target Encoder module, as depicted in Fig.2 (b). Based on this siamese Network structure, the joint encoder module and the target encoder module can generate distinct views of the masked area. In order to optimize the reconstruction task, the respective representations under the two views—that is, the distribution of the target representation $E_\zeta^{L_v}(Z_m)$ and the predicted representation $P_\varphi^{L_m}(\tilde{Z})$ are aligned using the mean square error loss $L_{rcr.}$.

$$L_{rcr.} = \left\| E_\zeta^{L_v}(Z_m) - P_\varphi^{L_m}(\tilde{Z}) \right\|_2^2 \tag{5}$$

3.4. Multi-tasking Optimization

LT-MAE is trained with multiple tasks, and the model's loss function is presented in (6). The hyperparameters α and β regulate the weights of the two loss functions. Classification and reconstruction tasks are integrated for optimization, and the joint encoder, target encoder, and tokenizer are all updated. Provide better input for downstream emotion recognition tasks.

$$L = \alpha L_{ecc.} + \beta L_{rcr.} \tag{6}$$

3.5. Emotion Recognition In Continuous Time Steps

In the downstream supervised emotion recognition task, only a portion of the joint encoder module E_θ and the tokenizer module T are retained. Fig.\ref{fig3} illustrates the process of integrating LSTM, E_θ, and the tokenizer module T for emotion recognition. First, the LSTM model is employed to automatically extract cross-channel contextual features. The contextual information is mapped to the learned representation space via E_θ to unify the distribution of each time step. Emotional codewords are then assigned using the tokenizer module T. Next, the emotional codewords are utilized as a reference to approximate the emotional changes over time. Finally, a classifier is employed to determine the emotional tendency during this time period, optimizing the emotion recognition task using the cross-entropy loss function. This type of overall emotion recognition, which incorporates many time steps, is clearly more interpretable and consistent with the meaning of labels in currently popular emotion datasets.

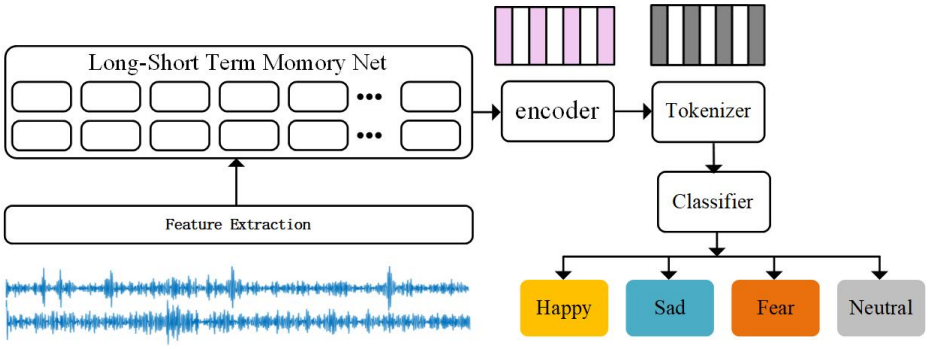


Figure 3. Architecture of the proposed emotion recognition model.

4. RESULTS

4.1. Datasets and Experimental Setup

"SJTU Emotion EEG Dataset IV (SEED-IV)" [20] is a free and open-source EEG emotion data set offered by Shanghai Jiao Tong University. This dataset captures the EEG signals of 15 people as they watch 72 movie snippets over three time periods. The movie snippets were categorized into four emotional states: happy, sad, fearful, and neutral. Each video clip only elicited one feeling in the subjects. The patients' EEG data were collected while they watched the movie using a 62-lead ESI NeuroScan System. The sample frequency was set to 200 Hz, with a 1–50 Hz bandpass filter employed for preprocessing.

The multi-channel EEG emotion data set "Database for Emotion Analysis using Physiological Signals (DEAP)" was produced by [21] This data set captures the physiological signals generated by thirty-two subjects throughout a forty-minute period of stimulation caused by one-minute music videos. Following the international 10-20 system to the letter, 32 active electrode channels provide EEG readings at a sampling rate of 512 Hz. Lastly, each participant rated their own four dimensions (arousal, valence, liking, and dominance) using a self-assessment approach.

In this experiment, non-overlapping time windows are utilized to segment the original signal, and electrode-by-electrode feature extraction is performed within the time window to extract features such as power spectral density and differential entropy. SEED-IV requires additional alignment operations due to the non-uniform durations of its 72 movie clips. Each segment, after alignment, includes 259 time steps. All samples were separated randomly into ten subsets, two of which were chosen as training sets, and the remaining subsets served as training sets. The experiments were carried out using cross-validation. Each experiment followed the same data division, feature extraction process, software and hardware environment, and assessment metrics until all subsets were evaluated. The model is built with the Pytorch framework and runs on the GeForce RTX 4050 GPU. The experiment's batch size is 16, and training is done using the AdamW optimizer with a learning rate of 0.0001.

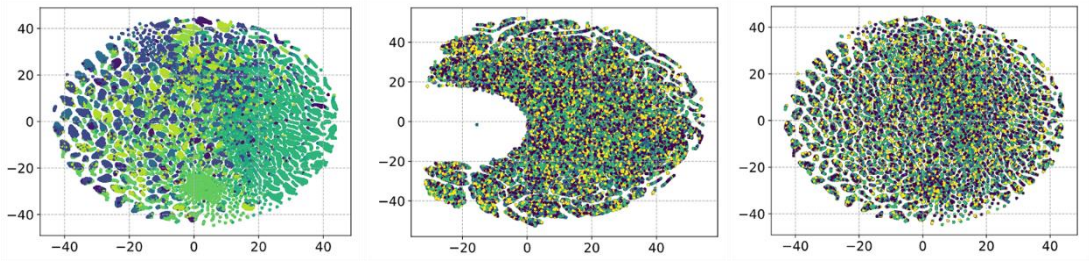
4.2. Time Step Sentiment Distribution

To demonstrate the role of the MAE encoder module in aligning the emotion distribution of time steps, three sets of comparative experiments (T1, T2, and T3) were conducted. All trials employed identical hyperparameters. In T1, the label of the time step inherits that of the entire EEG signal. In T2, although the time step still inherits the label of the full EEG signal, an encoder is used for representation learning. In T3, representation learning is performed, and the emotional codeword assigned by the tokenizer module T is used as the emotional label for the time step. The results are visually represented in a two-dimensional view using the t-sne method. Fig.4 and Fig.5 depict the outcomes from the two respective datasets. Fig.4 (a) and Fig.5(a) a show that time steps under various

emotion labels overlap and that time steps under the same emotion label are chaotically dispersed in the feature space. Fig.4(b) and Fig.5(b) show a more ordered emotion distribution after mapping to the representation space by the encoder, although distinct borders between emotion labels are not apparent. Using emotional codewords as emotion labels results in a more ordered distribution in the representation space, as seen in Fig.4(c) and Fig.5(c). Experimental results reveal that the time steps in the original feature space do not follow the same distribution, and the emotion label of the entire EEG signal cannot be used to estimate the emotion of the time step. The LT-MAE model's self-supervised MAE and marker modules may map time steps to the representation space, learn more reliable distribution patterns with less annotations, and thereby enhance emotion identification accuracy.

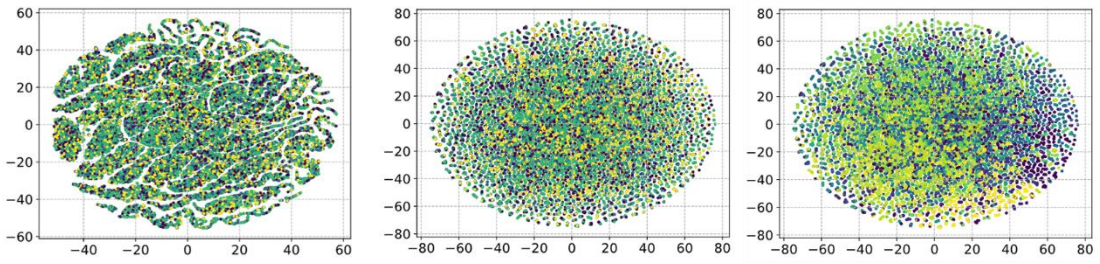
4.3. Time Window And Random Dropout Rate

Due to the non-short-term stationarity of EEG signals, there is currently no universally appropriate time window length for them. Most studies divide their time windows into 0.5-4 seconds to demonstrate the robustness of LT-MAE in varied time windows. 5 different window lengths ranging from 0.5 to 4 seconds were used for experiments (S1-S5). In addition, random masking is employed as a data augmentation technique, and the degree to which it masks data influences the MAE encoder's effectiveness. Control experiments P1-P4 have masking rates of 0.2, 0.4, 0.6, and 0.8, in that order. The accuracy of LT-MAE in completing four-category emotion recognition tasks under various datasets and masking rates is demonstrated in Table 1 and Table 2 Bold text indicates the best accuracy for the same time window. According to the experimental data, the optimal window length is 1 s, and the random dropout rate of 0.4 is a more reasonable choice. Increasing the time window from 0.5s to 1s enhances the information in the time window while decreasing the relationship between adjacent time windows, making the recognition task more difficult and improving recognition accuracy marginally. After the time window exceeds 1 second, a substantial amount of information collects within it, and the relationship between adjacent time windows is progressively diminished, resulting in a significant reduction in recognition accuracy. Regarding the random masking rate, the encoder can gather more information and attain higher accuracy by using a lower masking probability. A higher masking rate decreases the stability of the model by making the reconstruction effort more difficult. However, due to the presence of the marker, a specific number of masked areas are required to generate more precise emotional codewords. It can be seen that the LT-MAE model can still achieve good accuracy even when nearly half of the data is blocked, which reflects from the side that the MAE encoder has excellent representational learning capabilities.



(a) (b) (c)

Figure 4. Emotional distribution of time steps in SEED-IV.



(a) (b) (c)

Figure 5. Emotional distribution of time steps in DEAP.

Table 1. Experimental results under different time Windows and shielding rates in SEED-IV.

No.	window length	dropout rate =0.2 (P1)		dropout rate =0.4 (P2)		dropout rate =0.6 (P3)		dropout rate =0.8 (P4)	
		ACC.	F1-score	ACC.	F1-score	ACC.	F1-score	ACC.	F1-score
S1	0.5	83.33%	83.43%	85.19%	85.20%	84.26%	84.25%	80.55%	80.70%
S2	1	91.20%	91.02%	91.67%	91.39%	90.28%	90.07%	92.59%	92.45%
S3	2	64.35%	64.55%	75.92%	75.75%	73.15%	72.74%	73.61%	73.38%
S4	3	63.42%	63.20%	72.68%	72.62%	70.83%	70.69%	63.42%	62.85%
S5	4	66.20%	64.88%	68.98%	68.41%	68.05%	66.93%	51.85%	51.46%

Table 2. Experimental results under different time Windows and shielding rates in DEAP.

No.	window length	dropout rate =0.2 (P1)		dropout rate =0.4 (P2)		dropout rate =0.6 (P3)		dropout rate =0.8 (P4)	
		ACC.	F1-score	ACC.	F1-score	ACC.	F1-score	ACC.	F1-score
S1	0.5	81.94%	82.04%	79.62%	79.49%	82.41%	82.51%	80.55%	80.55%
S2	1	89.55%	89.09%	91.11%	90.51%	87.99%	87.24%	87.01%	86.76%
S3	2	62.50%	62.13%	68.98%	69.04%	75.00%	75.07%	76.85%	76.34%
S4	3	56.01%	54.57%	71.29%	70.73%	70.37%	70.14%	64.35%	64.23%
S5	4	67.12%	66.39%	62.50%	62.34%	67.59%	67.04%	54.16%	52.87%

4.4. Clustering

The work of the marker module in LT-MAE is very similar to clustering. Firstly, clustering divides all data into a set number of 'clusters' based on specific standards, ensuring that data samples within the same cluster are as similar to one another as possible. Samples in distinct clusters may be as dissimilar as is feasible. Secondly, clustering techniques can be applied to unlabeled datasets, and the resulting 'cluster' index can be utilized in self-supervised learning. Can serve as a form of supervision signal. However, unlike clustering, the tokenizer module is simply used to learn a codeword to characterize the data, whereas the MAE encoder module is responsible for bringing similar data together and distinguishing between distinct data, and the gradients used by the two are not the same. This may cause the "cluster" index to be incompatible with the representation learned by the autoencoder. To demonstrate this, comparison experiments C1, C2, and C3 were set up. C1 and C2 employed K-means clustering and DBSCAN clustering to generate "cluster" indexes as supervision signals, respectively, while C3 used the tokenizer module to generate supervision signals with a window length of 1s and a shielding rate of 0.4. Table 3 shows the experimental results. Because clustering and LT-MAE cannot use the same gradient, they must be conducted alternately, which increases the model's training duration. The experimental results are presented in Table 3 Because clustering and LT-MAE cannot use the same gradient, they must be conducted alternately, increasing the model's training duration. During training, both clustering and autoencoders undertake data zoom-

in and push-out operations at the same time, making it harder for the model to converge, as illustrated in Fig.6, hence affecting identification accuracy.

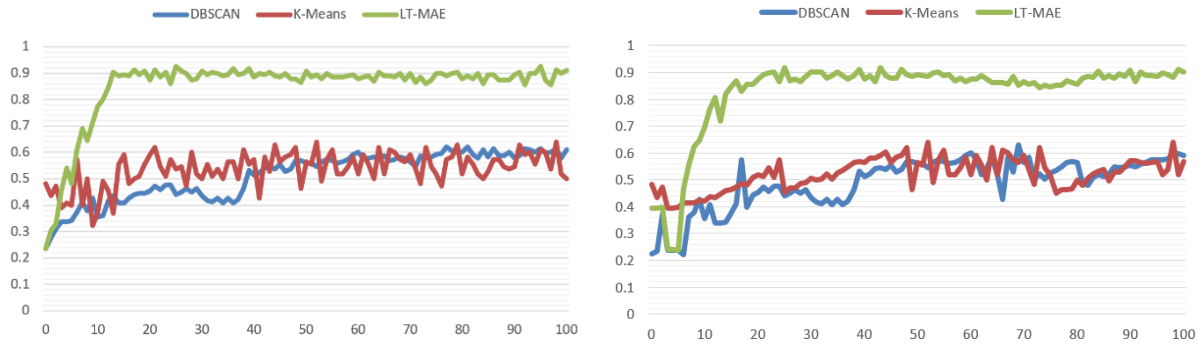


Figure 6. The training process uses the "cluster" index and codeword as a supervisory signal.

Table 3. The influence of clustering methods and tokenizer on accuracy.

No.	Method	SEED		DEAP	
		ACC	F1-score	ACC	F1-score
C1	k-means	63.89 %	61.48 %	58.33%	57.44%
C2	DBSCAN	61.57 %	61.53 %	56.48%	54.40%
C3	Tokenizer	91.20 %	91.39%	91.11%	90.51%

4.5. Subject-Independent Experiments

EEG signal distributions can vary throughout persons due to physiological variations. Subject-independent tests were carried out using the SEED-IV dataset with 15 people in order to validate the universality of the representation space learned by LT-MAE through self supervision. One of the 15 participants was chosen as the test set, while the other 14 subjects were utilized as the training set for emotion detection, which was compared to other commonly used models, as shown in Table 4, where bold denotes the best experimental results. In the subject-independent emotion recognition challenge, LT-MAE had an average recognition accuracy of 80.93%. Owing to emotional datasets' limited capacity as opposed to subject-mixed emotion recognition tasks, recognition accuracy varies significantly between subjects and declines overall. Out of the fifteen participants in the SEED-IV dataset, subjects two and eleven displayed irregularities in their identification accuracy, and the LT-MAE model was not able to infer their unique time step distribution patterns from the other subjects' samples. Compared with other models, LT-MAE has better performance in most cases.

Table 4. Subject-independent emotion recognition in SEED-IV.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
SVM	39.42	78.97	53.37	31.13	47.24	48.32	68.99	70.91	60.82	58.17	53.37	40.26	54.81	68.75	81.37
TCA	72.35	64.49	67.84	65.11	64.13	75.09	67.84	71.38	74.65	69.08	68.73	60.87	67.40	70.94	65.28
JDA	67.31	76.41	73.23	68.02	66.52	72.97	67.84	80.65	84.63	81.98	73.14	64.49	67.67	76.06	81.98
DAN	81.05	79.90	82.28	79.87	78.13	78.28	76.61	75.00	73.35	72.80	73.79	72.20	70.37	76.35	74.58
DANN	77.11	79.38	79.50	81.85	77.17	77.81	76.65	73.74	70.38	70.31	71.15	67.87	66.65	74.58	72.31
LT-MAE	90.28	45.83	91.67	94.44	80.56	77.78	87.50	75.00	84.72	87.50	63.89	93.06	76.39	86.11	79.17

4.6. Comparison With Existing Methods

To further confirm the effectiveness of the LT-MAE network, the sentiment recognition approach suggested in this paper will be compared to other recognition methods that employ the SEED-IV and DEAP datasets. Because accuracy is the most widely used evaluation criterion, recognition accuracy was employed as an indicator to assess the model. The experimental results are shown in Table.\ref{tb5}, frequency domain features offer a deeper insight into EEG data, enabling researchers to better comprehend the characteristics and patterns of brain activity. The majority of studies focus on extracting features in the frequency domain [21-24]. Due to the link between brain waves of different frequency ranges and various cognitive processes and emotional states, reference [22] further extracted frequency domain features of five frequency bands, achieving an accuracy of 82.32% in subject-indepenent subject emotion recognition. Certain research focus primarily on the spatial information present in EEG signals. They create feature maps by placing electrodes in specified locations, map the features that are retrieved into those locations, and then use deep learning networks to automate learning [23-26]. To extract spatial information from EEG signals, attention mechanism based networks and graph neural networks are frequently employed in addition to CNN. Reference [25] obtained an accuracy of 76.43 \% on the SEED-IV dataset using a spatiotemporal graph attention network with a variable voltage encoder. References [27] and references [28] employed comparable feature extraction approaches as this paper. PSD features are primarily utilized to collect and evaluate frequency domain data, whereas DE features are more focused on representing the signal's complexity and nonlinear characteristics. The experimental results demonstrate that the suggested LT-MAE model has a higher accuracy than alternative approaches, demonstrating the efficacy of the LT-MAE network.

5. CONCLUSION

Addressing existing challenges in the field of EEG emotion recognition, this paper introduces a self-supervised emotion recognition model called LT-MAE. The model enhances data through window partitioning, feature extraction, and random masking. It trains a masked autoencoder to learn the emotional distribution of time steps in the representation space, assigning similar emotional codewords to time steps with similar encodings to facilitate downstream identification tasks. In the downstream emotion recognition task, emotional tendencies are inferred by analyzing emotional changes across sequential time steps. Aggregating data from multiple time steps for overall emotion recognition is clearly more interpretable and compatible with the meaning of labels in currently popular emotion datasets.

Table 5. The influence of clustering methods and tokenizer on accuracy.

Datasets	Method	ACC. (%)
SEED-IV	STFT+LSTM [25]	74.26
	STGATE [27]	76.43
	Vae-capsnet [22]	77.14
	DFF-Net [26]	82.32
	JCSFE [29]	83.89
	Our method	91.67
DEAP	Hjorth-activity [28]	69
	EEG-GCN [30]	81.86
	STGB [23]	84.91
	SSDR [24]	88.60
	CNN-BiLSTM-MHSA [31]	89.33
	Our method	91.11

REFERENCES

[1] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. A review of emotion recognition using physiological signals. Sensors, 18(7):2074, 2018.

[2] Silvia De Nadai, Massimo D'Incà, Francesco Parodi, Mauro Benza, Anita Trotta, Enrico Zero, Luca Zero, and Roberto Sacile. Enhancing safety of transport by road by on-line monitoring of driver emotions. In 2016 11th System of Systems Engineering Conference (SoSE), pages 1--4, 2016.

[3] Rui Guo, Shuangjiang Li, Li He, Wei Gao, Hairong Qi, and Gina Owens. Pervasive and unobtrusive emotion sensing for human mental health. In 2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops, pages 436--439, 2013.

[4] Bruno Verschuere, Geert Crombez, Ernst Koster, and Katarzyna Uzieblo. Psychopathy and physiological detection of concealed information: A review. Psychologica Belgica, 46(1-2), 2006.

[5] Jing Cai, Ruolan Xiao, Wenjie Cui, Shang Zhang, and Guangda Liu. Application of electroencephalography-based machine learning in emotion recognition: A review. Frontiers in Systems Neuroscience, 15:729707, 2021.

[6] Shuaiqi Liu, Zeyao Wang, Yanling An, Jie Zhao, Yingying Zhao, and Yu-Dong Zhang. EEG emotion recognition based on the attention mechanism and pre-trained convolution capsule network. Knowledge-Based Systems, 265:110372, 2023.

[7] Liumei Zhang, Bowen Xia, Yichuan Wang, Wei Zhang, and Yu Han. A Fine-Grained Approach for EEG-Based Emotion Recognition Using Clustering and Hybrid Deep Neural Networks. Electronics, 12(23):4717, 2023.

[8] Ziyi Lv, Jing Zhang, and Estanislao Epota Oma. A Novel Method of Emotion Recognition from Multi-Band EEG Topology Maps Based on ERENet. Applied Sciences, 12(20):10273, 2022.

[9] Deng Pan, Haohao Zheng, Feifan Xu, Yu Ouyang, Zhe Jia, Chu Wang, and Hong Zeng. MSFR-GCN: A multi-scale feature reconstruction graph convolutional network for EEG emotion and cognition recognition. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2023.

[10] Yan Wu, Tianyu Meng, Qi Li, Yang Xi, and Hang Zhang. Study on multidimensional emotion recognition fusing dynamic brain network features in EEG signals. Biomedical Signal Processing and Control, 100:107054, 2025.

[11] Liwen Cao, Wenfeng Zhao, and Biao Sun. Emotion recognition using multi-scale EEG features through graph convolutional attention network. Neural Networks, 184:107060, 2025.

[12] Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. Journal of Personality and Social Psychology, 53(4):712, 1987.

[13] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American Scientist, 89(4):344--350, 2001.

[14] Peter J Lang. The emotion probe: Studies of motivation and attention. American Psychologist, 50(5):372, 1995.

[15] Albert Mehrabian. Comparison of the PAD and PANAS as models for describing emotions and for differentiating anxiety from depression. Journal of Psychopathology and Behavioral Assessment, 19:331--357, 1997.

[16] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Transactions on Autonomous Mental Development, 7(3):162--175, 2015.

[17] Li-Chen Shi, Ying-Ying Jiao, and Bao-Liang Lu. Differential entropy feature for EEG-based vigilance estimation. In 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 6627--6630, 2013.

[18] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for EEG-based emotion classification. In 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), pages 81--84, 2013.

[19] Wei-Long Zheng, Jia-Yi Zhu, and Bao-Liang Lu. Identifying stable patterns over time for emotion recognition from EEG. IEEE Transactions on Affective Computing, 10(3):417--429, 2017.

[20] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. IEEE Transactions on Cybernetics, 49(3):1110--1122, 2018.

[21] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. IEEE Transactions on Affective Computing, 3(1):18--31, 2011.

[22] Huayu Chen, Junxiang Li, Huanhuan He, Shuting Sun, Jing Zhu, Xiaowei Li, and Bin Hu. VAE-CapsNet: A common emotion information extractor for cross-subject emotion recognition. Knowledge-Based Systems, 113018, 2025.

[23] Jingjie Yan, Chengkun Du, Na Li, Xiaoyang Zhou, Ying Liu, Jinsheng Wei, and Yuan Yang. Spatio-temporal graph Bert network for EEG emotion recognition. Biomedical Signal Processing and Control, 104:107576, 2025.

[24] Behrooz Zali-Vargahan, Asghar Charmin, Hashem Kalbkhani, and Saeed Barghandan. Deep time-frequency features and semi-supervised dimension reduction for subject-independent emotion recognition from multi-channel EEG signals. Biomedical Signal Processing and Control, 85:104806, 2023.

[25] Xiangkun Yu, Zhengjie Li, Zhibang Zang, and Yinhua Liu. Real-Time EEG-Based Emotion Recognition. Sensors, 23(18):7853, 2023.

[26] Wei Lu, Haiyan Liu, Hua Ma, Tien-Ping Tan, and Lingnan Xia. Hybrid transfer learning strategy for cross-subject EEG emotion recognition. Frontiers in Human Neuroscience, 17, 2023.

[27] Jingcong Li, Weijian Pan, Haiyun Huang, Jiahui Pan, and Fei Wang. STGATE: Spatial-temporal graph attention network with a transformer encoder for EEG-based emotion recognition. Frontiers in Human Neuroscience, 17:1169949, 2023.

[28] Raja Majid Mehmood, Muhammad Bilal, S Vimal, and Seong-Whan Lee. EEG-based affective state recognition from human brain signals by using Hjorth-activity. Measurement, 202:111738, 2022.

[29] Yong Peng, Honggang Liu, Junhua Li, Jun Huang, Bao-Liang Lu, and Wanzeng Kong. Cross-session emotion recognition by joint label-common and label-specific EEG features exploration. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31:759--768, 2022.

[30] Yue Gao, Xiangling Fu, Tianxiong Ouyang, and Yi Wang. EEG-GCN: Spatio-temporal and self-adaptive graph convolutional networks for single and multi-view EEG-based emotion recognition. IEEE Signal Processing Letters, 29:1574--1578, 2022.

[31] Zhangfang Hu, Libujie Chen, Yuan Luo, and Jingfan Zhou. EEG-based emotion recognition using convolutional recurrent neural network with multi-head self-attention. Applied Sciences, 12(21):11255, 2022.