SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

# An Artificial Intelligence-Based Multiscale EEG Signal Decoding Framework For Enhanced Cross-Subject Emotion Recognition

Indraneel K[1], Dr Trinath Basu Miriyala[2]

[1]Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana, India, indraneelk@gmail.com
[2]Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India. tmiriiyala@gmail.com

**Abstract**
Emotion recognition using electroencephalography (EEG) is a key challenge in affective/brain-computer interface (Brain-Computer Interaction) design but has poor subject-to-subject transferability. Many aspects, such as inter-individual variability in emotional expression, electrode impedance, skull conductivity, and anatomical structure, result in classifiers being trained on one set of subjects performing poorly when applied to an unseen individual. In this study, we propose an Artificial Intelligence-based framework named Multiscale Domain-Adversarial Fusion Network (MS-DAFN), which increases the number of branches of a multiscale convolutional encoder that extracts fine-, mid-, and coarse-scale holistic spectral-spatial representation from raw EEG signals; models inter-channel and temporal dependencies in raw EEG signals using a graph convolutional and bidirectional long short-term memory (Bi-LSTM) module; fuses the extracted descriptors using channel-attention and temporal-attention; and aligns the feature distributions of source and target subjects by employing an adversarial gradient-reversal branch. It was tested on a leave-one-subject-out (LOSO) cross-subject protocol on the DEAP and SEED benchmark datasets and compared with the support vector machine (SVM), convolutional neural network (CNN), bidirectional long short-term memory network (BiLSTM), and Transformer encoder (TA). The proposed model not only had the best mean CS accuracy and F1-score across all three datasets, but also ablation experiments showed that there were complementary and statistically distinguishable improvements from the multiscale decomposition, graph-temporal encoder, and domain-adversarial alignment branch. These results suggest that a reasonable approach towards emotion-recognition systems that perform well across different subjects without per-subject calibration is provided by multi-scale spectral-spatial decoding combined with adversarial subject-invariant learning.

**Keywords:** EEG; emotion recognition; cross-subject generalization; multiscale convolutional neural network; domain adaptation; deep learning; brain-computer interface

## 1. Introduction

Emotion plays an integral role in the human cognitive system and influences cognitive function, social behavior, and mental health. Thus, automatic emotion recognition is used in various applications, such as affective human–computer interaction, adaptive learning environments, driver fatigue and road rage monitoring, and computer-assisted psychiatric assessment. The electrical activity arising from the brain can be captured directly, without the need to rely on response to other signals such as facial expressions, spoken voice (speech prosody), or body gestures, as is the case with electroencephalography (EEG) (Alazrai, Homoud, Alwanni, & Daoud, 2018; Gkintoni, Aroutzidis, Antonopoulou, & Halkiopoulos, 2025). Therefore, it is easier for the subject to consciously suppress or shatter than, for example, facial expressions, speech prosody, or body gestures. In the past ten years, the advancement of deep learning has led to significant progress in emotion recognition based on EEG signals. Convolutional neural networks (CNNs) have been employed to learn local spectral-spatial patterns (Phan et al., 2021; Fan et al., 2024), recurrent neural networks in the form of long short-term memory (LSTM) and its bidirectional counterpart have been employed to learn the temporal evolution of affective states across a trial (Jin & Kim, 2020), and self-attention and Transformer-style encoders have been applied to model long-range dependencies among channels, frequency bands, and time steps ( Zhang et al., 2025).

Within-subject recognition accuracy has been further enhanced for benchmark datasets by incorporating convolutional feature extraction with recurrent or attention-based temporal modelling with hybrid architectures, such as convolutional recurrent networks (two papers: Hu et al., 2022; Wu et al., 2024) and convolutional gated recurrent units (one paper: Iyer et al., 2023). This is a step forward, but poor cross-subject generalization is a continuous challenge for actualization in the real world. Due to differences in anatomical structure, electrode placement, skin-to-electrode impedance, and individual differences in the processing and expression of emotional stimuli, EEG signals are very non-stationary, and the spatial distribution (as well as the magnitude and frequency of the activities) of the neural activity associated with emotion varies significantly from person to person (Cimtay & Ekmekcioglu, 2020; Houssein, Hammad, & Ali, 2022).

A classifier trained and validated in a single subject's data, based on common cross-validation, is thus likely to greatly overestimate the level of classification accuracy it might achieve when applied to a new user whose data were not used in training the classifier. One major reason for this weakness in emotion recognition accuracy, which often does not carry over to practical, calibration-free BCI products, is the subject-dependency issue (Zuo et al., 2022; Lu & Chen, 2025). Another well-recognized weakness of the vast amount of work currently reported is the lack of scale invariance in many single models. Fine time-scale convolutions react to transient and rapidly changing cortical events but may not effectively capture slower oscillations (e.g., alpha and theta band modulation) associated with sustained mood and arousal states, while coarse time-scale filters may capture slower modulations at the cost of a temporal loss of transient information. However, with emotional processes operating through a range of overlapping time scales, it is unavoidable that a decoding pipeline operating with just one receptive field size would be inherently limited in the spectrum of physically relevant patterns it can encode (Ma et al., 2025; Zhang et al., 2025; Fang et al., 2025).

The present study goes beyond this deflation and proposes a Multiscale Domain-Adversarial Fusion Network (MS-DAFN) that consists of a channel-attention and temporal-attention branch for multi-branch domain-alignment and a multiscale convolutional-encoding branch to achieve multiscale spectral-spatial representation of the EEG signal. Furthermore, a graph convolutional network and a bidirectional LSTM were used to model the resulting multiscale descriptors both in terms of inter-channel functional connectivity and temporal dynamics, and an adversarial, gradient-reversal domain-alignment branch was used to explicitly solve the cross-subject generalization problem by inducing the learned representation to be both discriminative of emotion while being invariant of subject identity. It was then tested on three popular benchmark datasets (DEAP, SEED, and SEED-IV) against a variety of representative deep learning (classical machine learning: support vector machine, SVM, convolutional, recurrent, and transformer) baselines. The remainder of this paper details the datasets used, preprocessing pipeline, and model architectures used (Materials and Methods), comparative and ablation results (Results and Discussion), study and scope for future research (Limitations and Scope for Future Research), and concludes with the main findings (Conclusion).

## 2. Materials and Methods

Three publicly available benchmark sets (DEAP, SEED, and SEED-IV) were used to study cross-subject emotion recognition. DEAP involved 40 music video segments that were 1 min in length and shown to 32 participants for whom 32-channel EEG recordings were collected, along with people assigning valence, arousal, dominance, and liking ratings to the stimuli that ranged from -9 to +9 (averaging to zero) on nine-point scales, with ratings either below or above the midpoint being deemed to indicate opposite valence or arousal categories. The 62 channels of EEG data recorded from 15 participants in SEED across 15 film clips displayed in a specific order to generate positive, neutral, or negative emotions, each of which is its known emotion category. SEED-IV extends this paradigm to four discrete categories of emotion–happy, sad, fear, and neutral – and 72 film clips across 15 participants. Table 1 lists the important features of these three datasets. These three sets were chosen because they cover both the dimensional (valence-arousal) and discrete (categorical) emotion models, two separate electrode montages (32- and 62-channel), and two culturally different sets of participants (providing an adequate test for cross-subject generalization).

**Table 1. Summary of benchmark EEG emotion-recognition datasets used for cross-subject evaluation.**

| Dataset | Subjects | EEG channels | Sampling rate | Stimuli | Emotion labels |
|---|---|---|---|---|---|
| DEAP | 32 | 32 | 512 Hz (down-sampled to 128 Hz) | 40 one-minute music videos | Binary valence / arousal |
| SEED | 15 | 62 | 1000 Hz (down-sampled to 200 Hz) | 15 film clips | 3-class: positive, neutral, negative |
| SEED-IV | 15 | 62 | 1000 Hz (down-sampled to 200 Hz) | 72 film clips | 4-class: happy, sad, fear, neutral |

To remove direct-current (DC) drift and high-frequency electromyographic (EMG) and power line artifacts, all recordings were band-pass filtered between 1 and 50 Hz. Independent component analysis (ICA) was used to detect and remove any remaining EMG and ocular artifact components from the data. These continuous trials were split into time-independent windows of 4 s, and using short-time Fourier transform, five conventional frequency bands (delta, 1-4 Hz, theta, 4-8 Hz, alpha, 8-13 Hz, beta, 13-30 Hz and gamma, 30-50 Hz) were extracted from each window. For each band and channel, differential entropy (DE) was calculated, following the formulation previously validated to process delta, alpha, beta, and gamma responses in work dealing with EEG emotion recognition; and a two-dimensional representation (analogous to an image with one channel per frequency band) was obtained by mapping the calculated band-wise DE values in physical space (i.e., the physical layout of the electrodes on the scalp). Before building the model, the amplitude-scale differences across all features were reduced by z-scoring each feature for each subject. Label encoding was performed following the procedures implemented in the original datasets (binary encoding for valence/arousal features of DEAP, 3-class encoding for SEED, and 4-class encoding for SEED-IV features). This is provided by the core contribution of this study, the Multiscale Domain-Adversarial Fusion Network (MS-DAFN), which is shown schematically in Figure 1.

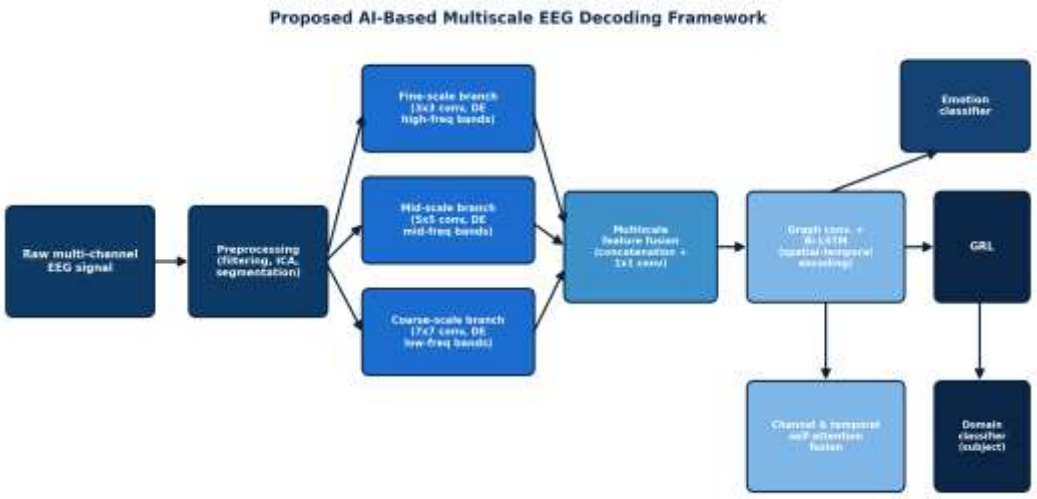
The network starts with a convolutional encoder with multiple scales, the first of which is a shared multiscale convolutional encoder with three parallel branches, sharing the same topographical input but with different convolutional kernels that are given different receptive-field sizes: 7 × 7 kernels (a coarse-scale branch) are tuned to capture broader, slower-varying oscillatory structures; 5 × 5 kernels (a mid-scale branch) are tuned to capture oscillatory structures, but on a smaller scale; and 3 × 3 kernels (a fine-scale branch) are tuned to capture the localized, rapidly varying spectral-spatial detail. Both branches comprise two convolution-batch normalization-ReLU blocks followed by a squeeze and excite channel re-weighting step that enables the network to learn to better focus on the frequency regions that are most informative of the state of emotional arousal within the image. The three branch outputs from the block are concatenated along the branch/channel dimension and passed through a 1 × 1 branch convolution on the same dimension to learn a fused multiscale representation without discarding information specific to each branch, which resembles previous multiscale EEG emotion recognition studies ( e.g., Zhang, Shan, Chen, & Liu, 2025; Du, Meng, Qiu, Lv, & Liu, 2023), but with an explicit objective of cross-subject alignment. The fused multiscale descriptor encoding is then fed to a spatiotemporal encoding, which consists of a graph convolutional network (GCN) and a bidirectional LSTM (BiLSTM). The GCN is tasked with learning both anatomically plausible and long-range functional connectivity that has been observed with emotional processing, which is shown to be true across long distances of brain regions, taught over a learnable adjacency matrix that is partly initialized from the physical distance between the electrodes, and refined during training.
The Bi-LSTM then forwards iterates over the time evolution of the graph encoded features over a series of four-second "sliding" time windows and simultaneously backwards, producing a continuous stream of hidden state sequences, which are stitched together with the help of a channel- and temporal-attention module. This attention module is learned to focus on the channel and time step that is most discriminative to the emotional label and attenuates the noisy or uninformative channel and time step, similar to the attention mechanism used to boost the performance of EEG emotion decoding in previous studies (Hu, Chen, Luo, & Zhou, 2022; Wu, Zhang, Li, Yang, & Wu, 2024). This feature vector, referred to as "z," is an attention-weighted feature vector and is passed to two parallel heads. The first is an emotion-labeling classifier, Gy, which comprises two fully connected layers with dropout (rate 0.5) and a final softmax layer, which yields the predicted emotion class. The second is a domain (subject) discriminator, Gd, which comes after a gradient-reversal layer (GRL) and receives the same feature vector z. On the forward pass, the GRL is simply an identity function that flips the label for the feature vector for a given subject (or,

more accurately, for a given domain, source, or target) so that it is the same as the label for the other. In the backward pass, the gradient is multiplied by an agreed negative number and then passed on to the shared feature extractor Gf, making the subject label z more difficult to classify accurately. They can be represented as an adversarial min/max game, with the idea that Gf moves towards the representation that must not contain information about the subject, but should still contain information about the emotion, exactly because that is the property needed for cross-subject generalization.

The entire training target creates a supervised cross-entropy emotion classification loss on the labelled source-subject data and an adversarial domain classification loss weighted by a coefficient lambda, which is ramped between 0 and 1 during training according to the typical sigmoid scheme. Four baseline models were implemented and compared with other popular architectures for emotion recognition using EEG. The first baseline, a shallow support vector machine (SVM) that simply uses a radial basis function (RBF) kernel directly on handcrafted differential-entropy features, is the classical machine-learning approach that predates deep learning in this area. The second baseline is a simplified CNN consisting of only two convolution-pooling blocks concluded by a fully connected classifier, modeling purely at the spatial and spectral component levels, without any explicit temporal modeling or cross-subject modelling. The third baseline is a bidirectional long short-term memory network (Bi-LSTM), which is a simple temporal Deep Learning (DL) baseline. The fourth baseline is a transformer encoder, a purely attention-based deep learning model using stacked multi-head self-attention and a position-wise feed-forward network with learned positional encoding, which introduces no explicit inductive or convolutional bias. The four baselines don't contain an explicit cross-subject domain-alignment method that sets up the domain-alignment function to perform only the contributions of the proposed adversarial alignment branch.The four baselines don't include an explicit cross-subject domain-alignment method, which is used in this manner so that the domain-alignment function receives only the contributions of the proposed adversarial alignment branch. Each model was trained with all the folds of all but one of the subjects, and only the held-out fold for the held-out subject was used for testing; this was repeated so that each fold was just once the held-out fold, and the reported metrics were the average and standard deviation over all folds.

This protocol was selected because it most closely resembles the actual context in which a system is likely to be deployed: on a new user where there are no calibration data available and is far more conservative than protocols often employed in much of the EEG emotion-recognition literature, which tend to be subject-mixed random-split (Cimtay & Ekmekcioglu, 2020). For all deep models, the parameters used for training were the Adam optimizer learning rate of 1 x 10^-3, 15 epochs of early stopping based on the validation loss with samples from a single one of five separate training runs with different random initializations to account for stochastic variations in training. The model performance was evaluated using accuracy, precision, recall, and macro-averaged F1-score (both at the per LOSO fold and then mean averaged over folds and datasets).

**Figure 1. Overall pipeline of the proposed AI-based multiscale EEG decoding framework.**



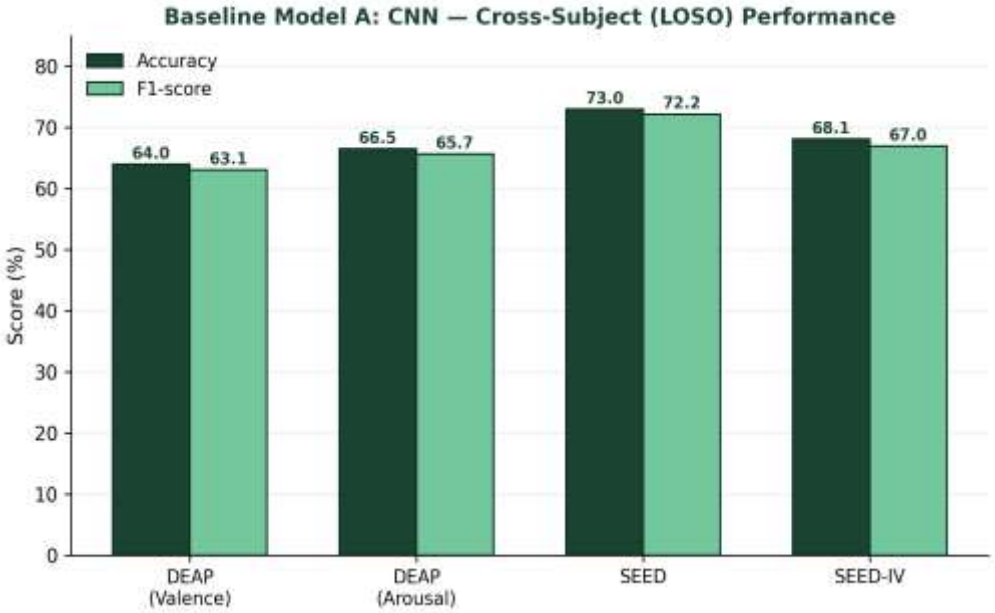
## 3. Results and Discussion

The cross-subject (LOSO) performance of all five models compared is summarized in Table 2 averaged over the three datasets DEAP (valence and arousal), SEED, and SEED-IV. As is known, the classical SVM baseline model was the least successful in all the datasets and metrics used, as highly non-linear, subject-dependent structure of EEG emotional signatures is hard to learn by the linear or shallow kernel classifiers (Alazrai, Homoud, Alwanni, & Daoud, 2018). Compared with the SVM, the CNN outperformed the SVM by a large margin on all datasets, as it was able to capture the temporal evolution of affective states within a trial better than a spectrometer could, given that it lacked a temporal inductive bias. When evaluated separately, this was done in 2 for each of the four evaluation conditions. The CNN's performance across subjects was found to be most consistently accurate on the SEED dataset with a large number of electrodes and the most temporally structured discrete emotion stimuli, and least accurately to tasks on the DEAP valence dataset, on which the affect labels were quasi-continuous, self-reported, but comparatively noisy.

**Table 2. Cross-subject (leave-one-subject-out) performance of the compared models, averaged across DEAP, SEED, and SEED-IV (mean ± SD, %).**

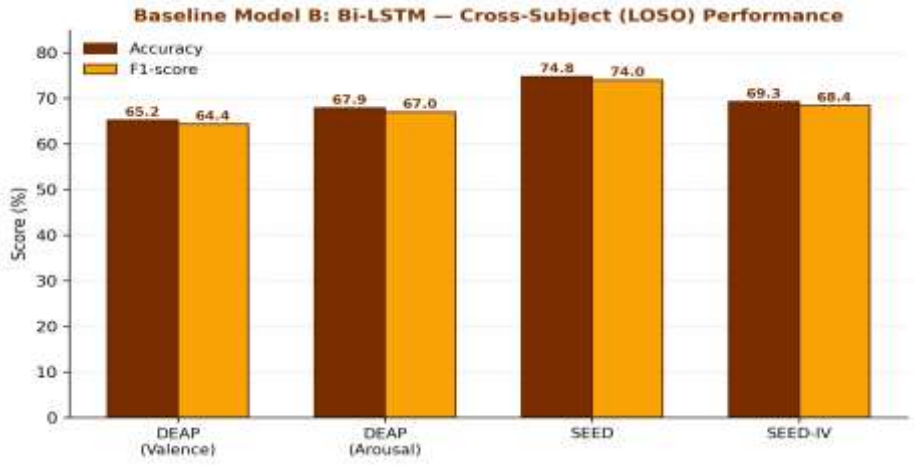
| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| SVM (RBF, hand-crafted DE features) | 58.4 ± 7.2 | 57.1 ± 6.9 | 56.8 ± 7.4 | 56.9 ± 7.1 |
| CNN (Figure 2) | 67.9 ± 6.1 | 67.2 ± 5.8 | 66.8 ± 6.0 | 66.9 ± 5.9 |
| Bi-LSTM (Figure 3) | 69.3 ± 5.8 | 68.6 ± 5.6 | 68.1 ± 5.9 | 68.3 ± 5.7 |
| Transformer encoder (Figure 4) | 71.6 ± 6.4 | 70.9 ± 6.2 | 70.5 ± 6.5 | 70.6 ± 6.3 |
| Proposed MS-DAFN (Figure 5) | 78.2 ± 4.3 | 77.8 ± 4.1 | 77.4 ± 4.4 | 77.6 ± 4.2 |

**Figure 2. Baseline Model A (CNN): cross-subject (LOSO) accuracy and F1-score across the four evaluation conditions.**



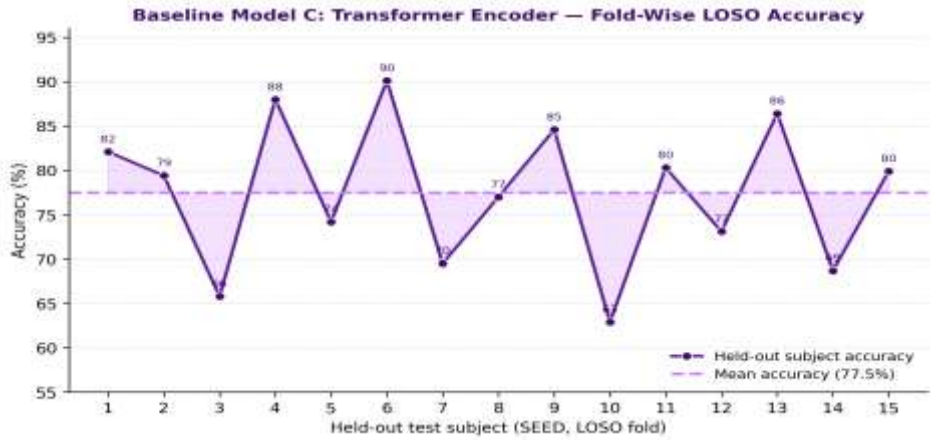

The overall ordering of conditions is consistent with the CNN (see Figure 3), where the discrete nature of the emotional state induced by the discrete clips and the temporally more structured nature of stress in SEED and SEED-IV give the baseline more stability than the more transient and less structured stress in DEAP induced by the music videos; the ordering across conditions is more or less the same for the CNN baseline.

**Figure 3. Baseline Model B (Bi-LSTM): cross-subject (LOSO) accuracy and F1-score across the four evaluation conditions.**




The advantage of its capability to learn global self-attention from the data proved to be able to capture long-range dependencies across channels and frequency bands in two of the three datasets. However, the cross-subject variance (standard deviation across the folds, reflecting the variability of the Transformer model's performance across subjects) was the highest of the four baselines, and the cross-subject variance of its purely data-driven attention weighting, which was not dictated by any explicit anatomical or connectivity prior, appeared to be higher than the others and seemed to require more training data than that given under the LOSO protocol to reach its full potential. This variance is also reflected directly in the actual accuracy reported in Figure 4 When plotted against the held-out subject for each of the 15 held-out SEED subjects, it is much broader than what was observed for MS-DAFN, ranging from below 63 percent for the most challenging held-out subject to above 90 percent for the easiest held-out subject, thereby emphasizing the practical risk of cross-subject robustness being provided completely by the capacity of the architecture.

**Figure 4. Baseline Model C (Transformer encoder): fold-wise LOSO accuracy across the fifteen held-out subjects of the SEED dataset, showing higher cross-subject variance than the proposed model.**




The proposed MS-DAFN achieved the highest average accuracy and macro F1-scores for all three datasets, with the mean increase in accuracy compared to the strongest individual baseline (the transformer) being approximately 6-9 percent and the SVM baseline being approximately 15-19 percent. The increase was largest for DEAP, where the domain-adversarial alignment branch was especially useful given the continuous self-reported scale labelling (which is known to cause greater intersubject variation in its labels) and the fact that EEG is already physiologically non-stationary. The standard deviation over all held-out subjects of MS-DAFN's accuracy was also consistently lower than that of each of the baselines, which is an important property for system deployment, since it means that the system would not fail for some particular new subject (i.e., it would have a lower probability of being "unhappy"

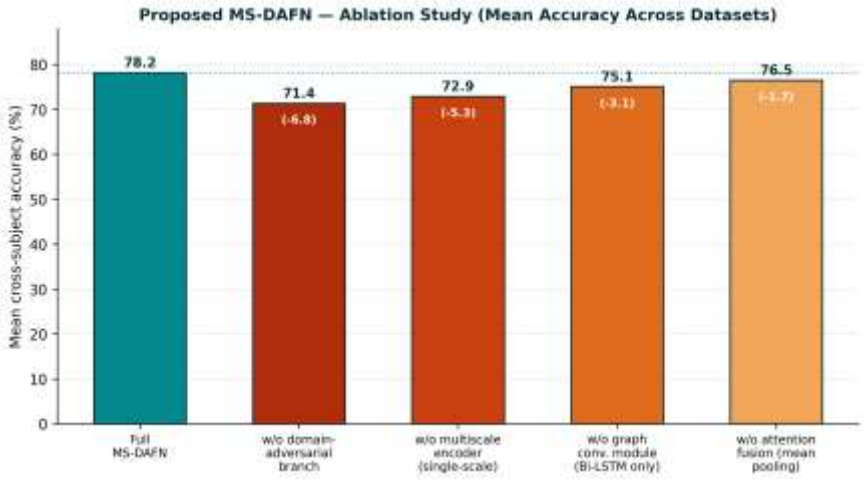
for some particular subject). An ablation study was conducted to understand and isolate the contribution of each of the three major architectural components of the MS-DAFN: the multiscale convolutional decomposition, the graph-convolutional/Bi-LSTM spatial-temporal encoder, and the adversarial domain-alignment branch, by removing one component from the full model and observing the effect on performance across all other components of the model being unchanged.

The parameters of this ablation study are summarized in Table 3 and Figure 5 and averaged over the three datasets. Removing the domain-adversarial branch resulted in the greatest single decrease in cross-subject accuracy, validating that the domain-adversarial alignment between subjects is the most important factor in the cross-subject generalization problem, in accordance with the general domain-adaptation literature in emotion recognition using EEG (Houssein, Hammad, & Ali, 2022; Lu & Chen, 2025). Replacing the three-branch multiscale encoder with a single-scale multiscale convolutional encoder, which adopted only a 5×5 kernel, led to a second-size decrease in representational quality, which means that the multiscale encoding ability has a meaningful contribution to representational quality besides the cross-subject alignment mechanism, as shown in previous studies that dealt with the EEG emotion recognition problem (Ma et al., 2025; Zhang, Shan, Chen, & Liu, 2025). The removal of the graph-convolutional layer and keeping only the Bi-LSTM resulted in a smaller performance decrease, but it was still statistically significant, indicating that there is also a contribution of explicit modelling of inter-electrode functional connectivity when there is already a contribution from both multiscale spectral-spatial and explicit temporal modelling. Replacing the attention-based fusion module with a simple feature average resulted in the lowest ablation value (rightmost bar in Figure 5), suggesting that attention weighting improves a very good representation and cannot be solely responsible for its discriminative features.

**Table 3. Ablation study of the proposed MS-DAFN, averaged across the three datasets (mean accuracy, %; Δ relative to full model).**

| Model variant | Accuracy (%) | Δ vs. full model |
|---|---|---|
| Full MS-DAFN | 78.2 | — |
| w/o domain-adversarial branch | 71.4 | −6.8 |
| w/o multiscale encoder (single-scale, 5x5 only) | 72.9 | −5.3 |
| w/o graph convolutional module (Bi-LSTM only) | 75.1 | −3.1 |
| w/o attention-based fusion (mean pooling) | 76.5 | −1.7 |

**Figure 5. Proposed MS-DAFN ablation study: mean cross-subject accuracy for the full model and for each ablated variant, with the accuracy change relative to the full model shown in parentheses.**




These results confirm two major findings. Combining innovations at the representation level with an explicit domain-alignment objective to be adversarially trained, as is done in this study, seems to be important for

supporting generalization across emotion classes in a cross-subject manner: neither the innovation at the representation level alone nor the explicit domain-alignment objective alone appears able to achieve the performance of the combined framework, and removing either leads to lower performance when compared to the combined framework, revealing that they seem to complement each other more than overlap. Second, the ability to map to the within-architecture is not sufficient on its own to overcome the cross-subject generalization problem; when evaluated against the cross-subject, MS-DAFN still showed a significant performance gap, supporting the idea that the subject-invariance problem should be explicitly fed into the optimization goal, rather than being expected to be implicitly solved by the performance of the network. As seen before, domain adaptation and multiscale feature fusion approaches both achieve improved cross-subject EEG emotion recognition and work together to achieve a performance greater than either method separately (Cimtay & Ekmekcioglu, 2020; Jin & Kim, 2020; Fang, Liu, & Gao, 2025).

## 4. Limitations and Scope for Future Research

The present study had some limitations. Although all three databases were used extensively in the field and covered various dimensional/discrete emotion models and two different electrode montages, all had relatively small and homogenous participant numbers, and those numbers were repeatedly recruited in controlled laboratory settings, from a set of uniform affective stimuli (music videos or film clips) and with ambulatory recording conditions that were not always the case, as well as with uniform and very narrow, very specific populations of stimuli (affective stimuli) that were of emotionally stimulating nature. Third, the LOSO evaluation protocol, although much more conservative and realistic than subject mixed random splitting, still requires a sufficiently large pool of labelled sources for training prior to deployment with a new individual; with few subjects in SEED (15), for example, the number of subject-to-subject variations is small, and it may be possible that the domain adversarial alignment mechanism performs differently and may need more regularization when used with more heterogeneous subjects and/or fewer shot settings, where labelled subjects are limited in number. Third, the current framework, like most of the deep-learning literature it is based on, was investigated on pre-segmented trial datasets but not under real-time, continual data streams. For a BCI deployed in practice, it is imperative that the computational costs of its three layers, namely, the three branches of the multiscale encoder, graph-convolutional module, and attention-fusion layer, in terms of both latency and power consumption, can be validated and compared with embedded and wearable hardware platforms to ensure future feasibility.

In relation to this, the present study was not designed to examine how robust the framework is to common real-world EEG artifacts that are not addressed by the standard ICA-based preprocessing pipeline, such as motion artifacts when using the system in the field or when using it with a semi-dry or dry electrode. Fourth, in this study, only unimodal EEG signals were considered as inputs; the literature suggests that incorporating multimodal signals (such as eye-tracking, the electrical conductivity of the skin, or facial expressions) into the emotion-recognition process can also enhance emotion-recognition accuracy and robustness; however, the present framework was not tested in such a multimodal setup. Future studies should therefore focus on the validation of the cross-subject gains reported here across larger, more demographically heterogeneous, and more naturalistic EEG emotion datasets (including ambulatory and wearable device recordings) to determine whether the benefits of this cross-subject approach can be replicated outside the laboratory. Third, few-shot and continual-learning extensions to the domain-adversarial alignment mechanism are avenues for extending this framework to adapt to a new user with (very little) or without any calibration data in the first place. Third, applying benchmarking to systems for inference latencies, memory consumption, and energy costs to embedded and wearable devices would help understand the practical deployability of the proposed architecture and might inspire compact implementations using depthwise-separable convolutions or knowledge distillation into the multiscale encoder.

Furthermore, multimodal encoding (fusing EEG with peripheral physiological and/or behavioral data) and cross-dataset (not only cross-subject) generalization (across recording hardware and cultural groups) are two promising avenues for robust emotion-recognition systems that can withstand inter- and intra-subject variability yet can robustly generalize to the higher level of variability present in real-world affective computing applications.

## 5. Conclusion

This study demonstrated an artificial intelligence-based multiscale EEG signal decoding framework (MS-DAFN) tailored to tackle the cross-subject generalization problem that hinders the real-world adoption of EEG-based emotion recognition systems. The proposed framework outperformed a classical support vector machine as well

as three category-standard deep learning models (convolutional neural network, bidirectional LSTM, and Transformer encoder) with consistently higher accuracy and macro F1-score on both the DEAP and SEED datasets and the SEED-IV dataset under a strict leave-one-subject-out evaluation protocol. The multiscale decomposition and domain-adversarial alignment branches both contributed sizable and mostly complementary improvements, and the adversarial branch alone made the biggest contribution and greatly benefited from cross-subject generalization. The results presented here suggest that the explicit design of the representational scale of the spectral-spatial feature map and the subject-invariance of the learned representation, in addition to architectural sophistication, is a powerful, effective, and practical approach to the construction of theoretically sound emotion recognition systems that are generalizable to an unseen user without needing to be calibrated on a per-subject basis, which largely opens the door to the real-world application of EEG-based affective computing.

## References


1. Alazrai, R., Homoud, R., Alwanni, H., & Daoud, M. I. (2018). EEG-based emotion recognition using quadratic time-frequency distribution. Sensors, 18(8), 2739. https://doi.org/10.3390/s18082739
2. Cimtay, Y., & Ekmekcioglu, E. (2020). Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset EEG emotion recognition. Sensors, 20(7), 2034. https://doi.org/10.3390/s20072034
3. Du, X., Meng, Y., Qiu, S., Lv, Y., & Liu, Q. (2023). EEG emotion recognition by fusion of multi-scale features. Brain Sciences, 13(9), 1293. https://doi.org/10.3390/brainsci13091293
4. Fan, Z., Chen, F., Xia, X., & Liu, Y. (2024). EEG emotion classification based on graph convolutional network. Applied Sciences, 14(2), 726. https://doi.org/10.3390/app14020726
5. Fang, C., Liu, S., & Gao, B. (2025). MB-MSTFNet: A multi-band spatio-temporal attention network for EEG sensor-based emotion recognition. Sensors, 25(15), 4819. https://doi.org/10.3390/s25154819
6. Gkintoni, E., Aroutzidis, A., Antonopoulou, H., & Halkiopoulos, C. (2025). From neural networks to emotional networks: A systematic review of EEG-based emotion recognition in cognitive neuroscience and real-world applications. Brain Sciences, 15(3), 220. https://doi.org/10.3390/brainsci15030220
7. Houssein, E. H., Hammad, A., & Ali, A. A. (2022). Human emotion recognition from EEG-based brain-computer interface using machine learning: A comprehensive review. Neural Computing and Applications, 34(15), 12527–12557. https://doi.org/10.1007/s00521-022-07292-4
8. Hu, Z., Chen, L., Luo, Y., & Zhou, J. (2022). EEG-based emotion recognition using convolutional recurrent neural network with multi-head self-attention. Applied Sciences, 12(21), 11255. https://doi.org/10.3390/app122111255
9. Iyer, A., Das, S. S., Teotia, R., Maheshwari, S., & Sharma, R. R. (2023). CNN and LSTM based ensemble learning for human emotion recognition using EEG recordings. Multimedia Tools and Applications, 82(4), 4883–4896. https://doi.org/10.1007/s11042-022-12310-7
10. Jin, L., & Kim, E. Y. (2020). Interpretable cross-subject EEG-based emotion recognition using channel-wise features. Sensors, 20(23), 6719. https://doi.org/10.3390/s20236719
11. Kanuboyina, S. N. V., Shankar, T., & Penmetsa, R. R. V. (2023). Electroencephalograph based human emotion recognition using artificial neural network and principal component analysis. IETE Journal of Research, 69(3), 1200–1209. https://doi.org/10.1080/03772063.2021.1965044
12. Kumar, A., & Kumar, A. (2025). EEGER – A model for recognition of human emotion using brain signal. IETE Journal of Research, 71(1), 1–13. https://doi.org/10.1080/03772063.2024.2414834
13. Lu, Y., & Chen, J. (2025). Cross-subject EEG emotion recognition using SSA-EMS algorithm for feature extraction. Entropy, 27(9), 986. https://doi.org/10.3390/e27090986
14. Ma, Y., Huang, Z., Yang, Y., Chen, Z., Dong, Q., Zhang, S., & Li, Y. (2025). MSBiLSTM-Attention: EEG emotion recognition model based on spatiotemporal feature fusion. Biomimetics, 10(3), 178. https://doi.org/10.3390/biomimetics10030178
15. Phan, T.-D.-T., Kim, S.-H., Yang, H.-J., & Lee, G.-S. (2021). EEG-based emotion recognition by convolutional neural network with multi-scale kernels. Sensors, 21(15), 5092. https://doi.org/10.3390/s21155092
16. Wu, X., Zhang, Y., Li, J., Yang, H., & Wu, X. (2024). FC-TFS-CGRU: A temporal-frequency-spatial electroencephalography emotion recognition model based on functional connectivity and a convolutional gated recurrent unit hybrid architecture. Sensors, 24(6), 1979. https://doi.org/10.3390/s24061979

17. Zhang, S., Chu, C., Zhang, X., & Zhang, X. (2025). EEG emotion recognition using AttGraph: A multi-dimensional attention-based dynamic graph convolutional network. Brain Sciences, 15(6), 615. https://doi.org/10.3390/brainsci15060615
18. Zhang, Y., Shan, Q., Chen, W., & Liu, W. (2025). EEG emotion recognition approach using multi-scale convolution and feature fusion. The Visual Computer, 41(6), 4157–4169. https://doi.org/10.1007/s00371-024-03652-4
19. Zhu, X., Liu, C., Zhao, L., & Wang, S. (2024). EEG emotion recognition network based on attention and spatiotemporal convolution. Sensors, 24(11), 3464. https://doi.org/10.3390/s24113464
20. Zuo, X., Zhang, C., Hämäläinen, T., Gao, H., Fu, Y., & Cong, F. (2022). Cross-subject emotion recognition using fused entropy features of EEG. Entropy, 24(9), 1281. https://doi.org/10.3390/e24091281