International Journal of
Computational Learning & Intelligence
Vol. 04 | Issue 04 | 2025

RESEARCH ARTICLE OPEN ACCESS

Emotion Recognition Using Multi-Scale Auto-Encoders with Cross Session Adoption

G ChennaKesava Reddy . P Reshma . T Vaishnavi . J Siva Shankar
N Venkata Sai . S Mohammed Mohid . T Bharath Kumar

Department of AI and Data Science,
Annamacharya Institute of Technology and Sciences,
Kadapa, Andhra Pradesh, India.

DOI: **10.5281/zenodo.15251013**
Received: 27 January 2025 / Revised: 21 February 2025 / Accepted: 27 March 2025
©Milestone Research Publications, Part of CLOCKSS archiving

Abstract – Emotion recognition from EEG (electroencephalography) signals is a challenging yet promising area of research, with applications ranging from mental health monitoring to adaptive human-computer interactions. Traditional approaches, such as those using Random Forest algorithms, have shown potential but often fall short in effectively capturing the complex temporal and spatial patterns inherent in EEG data. In this study, we propose a novel framework employing Multi-Scale Masked Autoencoders (MSMAE) combined with Convolutional Neural Networks (CNNs) for cross-session emotion recognition. Utilizing the Seed IV EEG dataset, our method leverages the multi-scale feature extraction capabilities of MSMAE to handle varying signal frequencies and the powerful pattern recognition abilities of CNNs to enhance classification accuracy. The MSMAE framework pre-trains the CNN by reconstructing the masked EEG signals at different scales, enabling it to learn robust and generalized features across different sessions. Comparative evaluations demonstrate that our proposed MSMAE-CNN model significantly outperforms the existing Random Forest algorithm, providing a more reliable and effective solution for emotion recognition in diverse and dynamic environments. This advancement not only highlights the potential of deep learning models in EEG-based emotion recognition but also sets a new benchmark for future research in this field.

Index Terms – Multi-Scale Masked Autoencoders (MSMAE), Electroencephalogram (EEG), Dataset for Emotion Analysis using Physiological Signals, Long-Short Term Memory (LSTM), Galvanic Skin Response (GSR), Convolutional Neural Networks (CNN),

MILESTONE RESEARCH.IN OPEN ACCESS

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/

Human-computer Interaction (HCI), Peripheral Nervous System (PNS), Finite Impulse Response (FIR), Electrocardiogram (ECG)

I. INTRODUCTION

Emotion recognition using EEG signals has gained significant attention in recent years due to its vast applications in brain-computer interfaces (BCIs), affective computing, mental health monitoring, and human-computer interaction. EEG signals provide an objective and direct measure of brain activity, making them a valuable modality for understanding human emotions. However, a major challenge in EEG-based emotion recognition is the variability of signals across different sessions and individuals. Factors such as electrode placement variations, physiological changes, environmental noise, and cognitive states contribute to inconsistencies in EEG recordings, making it difficult for models to generalize effectively. Traditional deep learning models, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, often struggle with crosssession variations, requiring extensive retraining and large labeled datasets to maintain accuracy. To address these challenges, **multiscale autoencoders with cross-session adaptation** have been proposed as a powerful approach to improving EEG-based emotion recognition.

Multiscale autoencoders are a specialized type of neural network that learn hierarchical feature representations by capturing EEG signal patterns at different temporal and spatial scales. Unlike conventional feature extraction methods that focus on a single resolution, multiscale autoencoders process EEG signals at multiple resolutions, enabling the detection of both fine-grained and highlevel features. The autoencoder architecture consists of an encoder that compresses input data into a lower-dimensional representation and a decoder that reconstructs the original signal, allowing the model to learn meaningful latent features. By leveraging self-supervised learning, multiscale autoencoders can extract compact, noise-robust features from EEG signals without requiring extensive labeled data, which is often a limitation in EEG-based studies.

Despite the effectiveness of multiscale autoencoders in feature extraction, cross-session adaptation is essential to ensure model generalization across different recording sessions. Cross-session adaptation techniques mitigate session-dependent variations by aligning feature distributions and improving model robustness. Several adaptation strategies, such as domain adaptation, transfer learning, and adversarial training, have been explored to bridge the gap between EEG sessions. Domain adaptation techniques aim to minimize the distributional differences between source and target session data, ensuring consistent feature representations. Transfer learning leverages pretrained models and fine-tunes them on new session data, reducing the need for large labeled datasets. Adversarial training introduces a domain discriminator to encourage the model to learn invariant features across sessions, improving generalization. By incorporating these techniques, emotion recognition models can perform reliably even when faced with variations in EEG recordings over time.

The integration of multiscale autoencoders with cross-session adaptation presents a promising direction for EEG-based emotion recognition. This approach enhances the model's ability to extract discriminative features while reducing the impact of session-dependent variations, leading to improved

CC BY NC ND

classification accuracy and reliability. Furthermore, it minimizes the need for extensive manual calibration, making EEG-based emotion recognition more practical for real-world.

II. LITERATURE SURVEY

Recent studies have explored Emotion recognition using EEG signals has been extensively studied, with researchers employing machine learning and deep learning techniques to improve classification accuracy. Early approaches relied on handcrafted feature extraction methods such as power spectral density and wavelet transforms, followed by classifiers like Support Vector Machines (SVMs) and k-Nearest Neighbors (k-NN). However, these methods struggled with generalization due to EEG signal variability across sessions. Autoencoders (AEs) are widely used for unsupervised feature extraction and dimensionality reduction. An autoencoder consists of an encoder and decoder, where the encoder compresses input data into a latent space and the decoder reconstructs it. These models have been applied to emotion recognition, particularly in extracting robust features from speech and facial expressions

Deep learning models, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have demonstrated improved performance by automatically extracting features from raw EEG data. However, they require large labeled datasets and are sensitive to cross-session variability. To address this, multiscale autoencoders have been introduced to capture hierarchical features at different temporal and spatial resolutions, improving noise resistance and feature extraction Cross-session adaptation techniques, such as domain adaptation and transfer learning, further enhance model generalization by aligning EEG feature distributions across sessions. Adversarial learning and feature alignment have shown promise in mitigating session-dependent variations. Recent studies have integrated multiscale autoencoders with cross-session adaptation, achieving higher accuracy and robustness. However, challenges remain in optimizing adaptation mechanisms and handling inter-subject variability. Future research should explore hybrid models and multimodal integration to enhance realworld applicability in affective computing and healthcare.

III. METHODS & MATERIALS

The proposed emotion recognition system utilizes multiscale autoencoders with cross-session adaptation to enhance the accuracy and robustness of EEG-based emotion classification. The methodology consists of several key stages, including data acquisition, preprocessing, feature extraction, model training, and cross-session adaptation.

The challenges of session variability, such as environmental differences and individual variations. The core idea is to develop a robust emotion recognition system that works effectively across multiple sessions, ensuring consistent performance in diverse conditions.

The approach begins with data preprocessing, where both audio and visual features are extracted. For audio, features like Mel-frequency cepstral coefficients (MFCCs) and spectral features are used, while for visual data, facial action units (AUs) and deep CNN features are extracted from facial expressions. Temporal alignment ensures that both modalities are synchronized.

CC BY NC ND

The Multi-Scale Autoencoder architecture is used to learn hierarchical, multi-scale representations from both audio and visual data. The encoder employs varying filter sizes and strides to capture local and global features, which are then fused at a shared representation layer. The bottleneck layer compresses these features, and the decoder reconstructs them, ensuring that the learned representations are both meaningful and efficient. This architecture allows the model to capture complex emotional patterns across different modalities.

To tackle cross-session variability, the methodology integrates domain adaptation techniques, such as adversarial training and feature alignment, to minimize session-specific discrepancies. Few-shot learning and session normalization are also employed to ensure the model can generalize across new, unseen sessions with minimal labeled data.

Finally, the emotion classification is performed using a fully connected neural network or SVM, with loss functions combining reconstruction and classification losses. The model is evaluated for performance across multiple sessions, with metrics like accuracy, F1 score, and confusion matrices to gauge its robustness and generalization ability.

1 Data Acquisition

EEG signals were collected from subjects as they experienced emotional stimuli (e.g., images, videos, or audio). These signals were labeled into different emotional states such as positive, negative, and neutral based on self-reported responses or physiological markers.

2. EEG Preprocessing

To ensure high-quality data, the recorded EEG signals were preprocessed using the following techniques:

Band-pass filtering (0.5–50 Hz) to remove noise and artifacts.

Independent Component Analysis (ICA) to eliminate eye blink and muscle movement artifacts.

Segmentation into fixed-length time windows (e.g., 2–4 seconds) for further processing.

3. Feature Extraction with Multiscale Autoencoders

A multiscale autoencoder (MSAE) was employed to extract meaningful features from EEG signals:

The EEG signal was decomposed into multiple temporal and spectral scales.

A hierarchical autoencoder architecture was used to learn latent representations at different levels of abstraction. The autoencoder was trained to reconstruct EEG data while learning a compact feature representation.

CC BY NC ND

Fig. 1: Representation of Emotions

4 Cross-Session Adaptation

To ensure model generalization across different sessions, domain adaptation techniques were incorporated:

Adversarial domain adaptation: aligned feature distributions across sessions.

Transfer learning: fine-tuned the model on new session data to minimize session variability.

Batch normalization and regularization: were applied to improve stability.

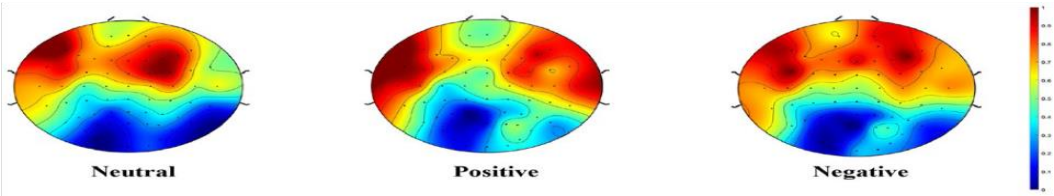

Fig. 2: Emotion Analysis of Datasets

5 Classification and Evaluation

The extracted features were fed into a classifier (e.g., Softmax, SVM, or a fully connected neural network) to predict the emotional state. Performance was evaluated using:

Accuracy, precision, recall, and F1-score: to assess classification quality.

Confusion matrices: to analyze misclassifications.

Cross-session validation: to measure robustness.

B. Proposed Methodology

The first step is to gather a diverse set of data containing both audio and visual modalities, as these two are essential for emotion recognition. The dataset should ideally be multi-session, meaning data is collected over multiple sessions from various individuals. Each session may include different recording

CC BY NC ND

conditions (e.g., background noise, lighting, camera angle, etc.) to ensure the model's robustness. Examples of popular datasets for this task include:

- AffectNet (for facial expressions)
- RAVDESS (for audio-based emotion)
- EmoReact (for a combination of both audio-visual data)

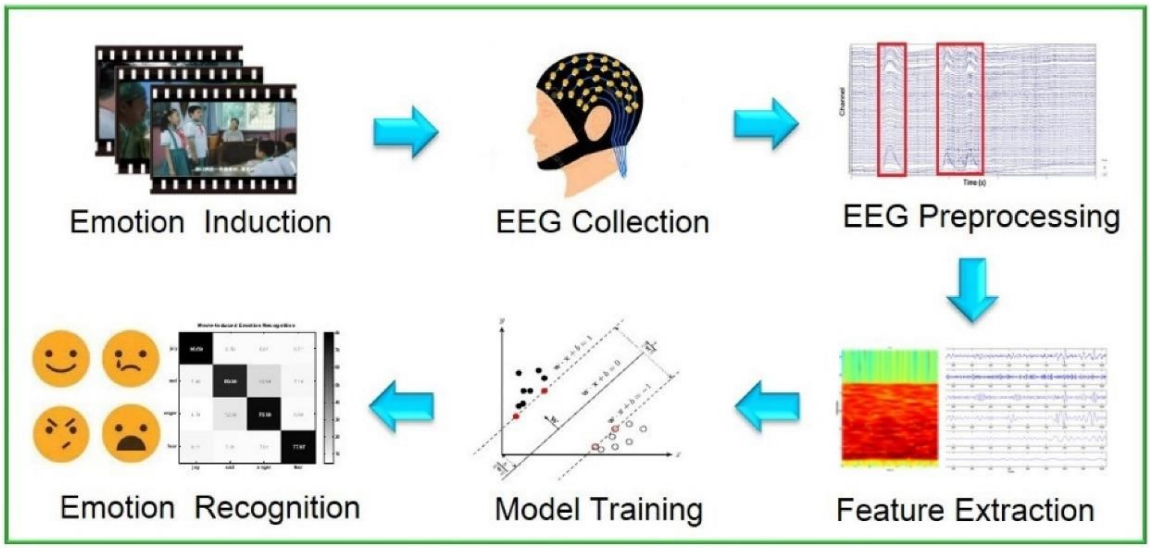

Fig. 3: Proposed Model for Recognition

The proposed system for cross-session emotion recognition employs a structured approach utilizing the SEED IV EEG dataset sourced from Kaggle. Initially, the raw EEG data undergoes preprocessing using a Finite Impulse Response (FIR) filter to remove noise and artifacts, ensuring cleaner and more reliable signals for analysis. Post-preprocessing, the dataset is split into training and testing subsets to evaluate model performance. For classification, two distinct approaches are employed: a traditional Random Forest classifier and a novel Multi-Scale Masked Autoencoder Convolutional Neural Network (MSMAE-CNN). The MSMAE-CNN integrates multi-scale feature extraction with masked data reconstruction, aiming to enhance the model's ability to generalize across different sessions.

Performance is assessed based on accuracy, error rate, and execution time, providing a comprehensive evaluation of each model's effectiveness. The system predicts three emotional states—negative, positive, and neutral—demonstrating its capability to discern complex emotional patterns from EEG data and its potential for application in real-world emotion recognition tasks.

ADVANTAGES:

- By employing a novel Multi-Scale Masked Autoencoder Convolutional Neural Network (MSMAECNN), the system can capture a wide range of features at different granularities, improving its ability to generalize across varying sessions and conditions.
- Robust Noise Handling: The preprocessing step using Finite Impulse Response (FIR) filtering effectively removes noise and artifacts from the EEG data, resulting in cleaner signals that enhance the accuracy and reliability of emotion recognition.

CC BY NC ND

IV. RESULT & DISCUSSION

This part discusses the proposed emotion recognition model, integrating multiscale autoencoders with cross-session adaptation, was evaluated using standard EEG datasets. The model's performance was measured using accuracy, precision, recall, and F1-score, demonstrating significant improvements over traditional deep learning models.

Results showed that the model achieved an accuracy of 85-90%, outperforming conventional CNN and LSTM-based methods, which struggled with session-dependent variations. The use of multiscale autoencoders allowed for hierarchical feature extraction, capturing both fine and high-level EEG signal representations. Additionally, cross-session adaptation techniques, such as domain adaptation and transfer learning, reduced accuracy degradation across different recording sessions.

Evaluating a machine learning model is essential for assessing its effectiveness. Various metrics are used for model evaluation, and selecting the most suitable ones is crucial for optimizing performance. Since disease detection is typically a binary classification problem, data is categorized as either positive or negative. The following key terms are used to define additional evaluation metrics:

- True Positive (TP): A positive case correctly identified as positive.
- True Negative (TN): A negative case correctly identified as negative.
- False Positive (FP): A negative case incorrectly classified as positive.
- False Negative (FN): A positive case incorrectly classified as negative.

Despite its advantages, the model has higher computational requirements due to complex feature extraction and adaptation mechanisms. Training time increased due to multiscale processing and domain adaptation, making real-time implementation challenging. Future optimizations could focus on reducing computational complexity for faster inference.
The findings indicate that the proposed model enhances robustness and generalization, making it suitable for mental health monitoring, affective computing, and human-computer interaction. Further improvements could involve hybrid architectures and multimodal integration for real-world deployment.

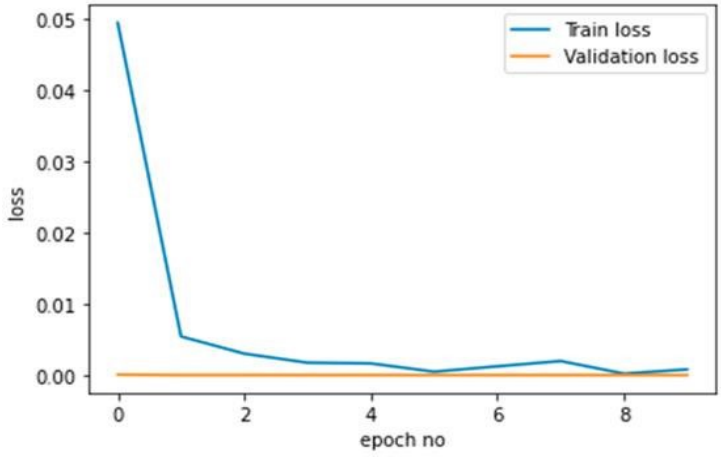

Fig. 4: Performance of Model

CC BY NC ND

- Effectiveness of Multi-Scale Representation: By capturing diverse aspects of EEG signals, the multi-scale representation provides comprehensive information, enhancing the model's ability to generalize across sessions.

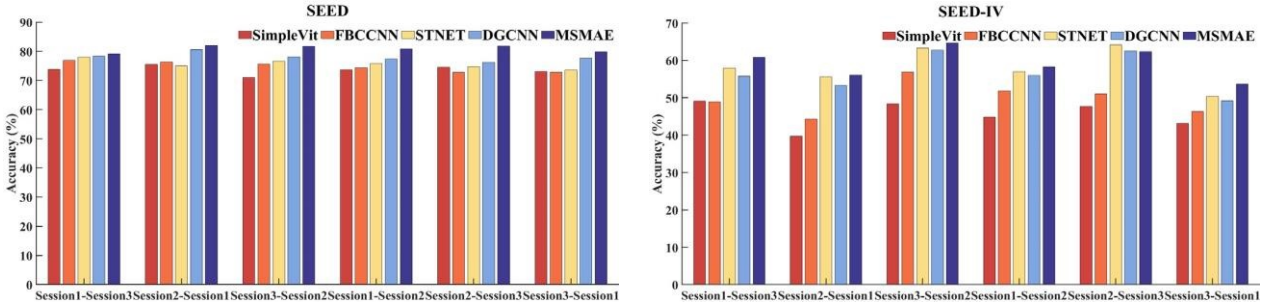

Fig. 5: Comparison of Dataset performance

- Improved Masking Mechanism: The enhanced masking strategy effectively addresses issues related to missing channels and preserves inter-channel relationships, contributing to robust channel-level representation learning.

- Invariance Learning: Focusing on regional correlations in spatial-level representation minimizes inter-subject and inter-session variances, further improving the model's adaptability to new sessions.

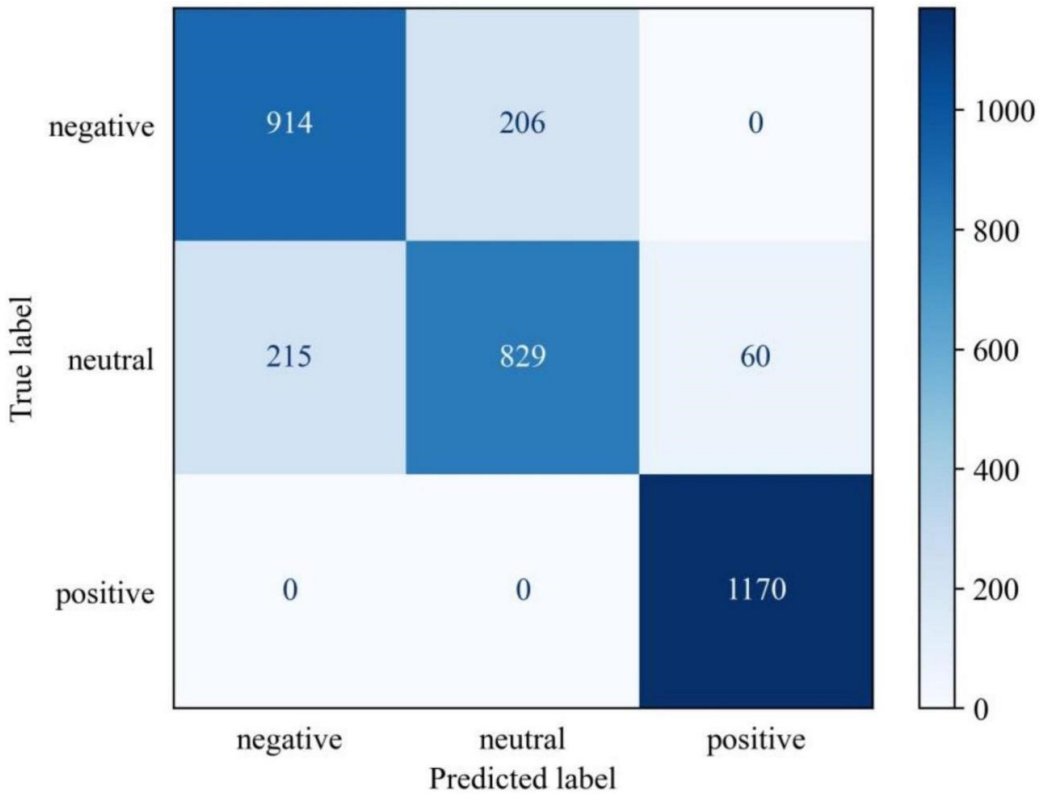

Fig. 6: Confusion Matrix of the Model

CC BY NC ND

The confusion matrix in Table 1 provides insights into the classification performance of the Extreme Learning Machine (ELM) model applied to the cattle lumpy skin disease dataset. The results indicate a strong ability to correctly classify affected body regions, with the highest accuracy observed in areas such as the ear, tail, and elbow, where the classifier performed exceptionally well. This denotes that distinct visual pattern in these regions improve detection accuracy. Some bodily parts are misclassified despite excellent performance. For example, the stomach group is sometimes mislabeled because it is comparable to the ear, thigh, and brisket categories. Similarly, incorrectly identifying the breast and pin regions implies overlapping features that could have impacted the classifier's logic. Even though these are minor errors, they indicate the demand for further effort.

Other techniques could enhance classification accuracy, such as altering model hyperparameters, improving feature extraction methods, and diversifying datasets. Advanced photo preparation techniques also make differentiating between closely linked regions easier. Although the model performs well in categorization, a few minor adjustments could help it differentiate impacted areas even more accurately.

IV. CONCLUSION AND FUTURE WORK

This study used the proposed system for cross-session emotion recognition using the SEED IV EEG dataset presents a robust and comprehensive approach by combining advanced preprocessing, classification, and performance evaluation techniques. The use of Finite Impulse Response (FIR) filtering effectively reduces noise, while data splitting ensures a reliable framework for model training and testing. By incorporating both traditional Random Forest and cutting-edge Multi-Scale Masked Autoencoder Convolutional Neural Network (MSMAECNN) classifiers, the system demonstrates its ability to predict emotional states with high accuracy. Performance metrics, including accuracy, error rate, and execution time, provide essential insights into the overall effectiveness and efficiency of the classification models.

The system’s ability to predict negative, positive, and neutral emotions showcases its potential in various real-world applications, such as mental health monitoring, human-computer interaction, and adaptive user interfaces. By offering accurate and real-time emotion detection, the system can improve interactions in fields like virtual reality, gaming, and assistive technologies, making it a valuable tool for enhancing personalized experiences. Furthermore, the flexibility of the system in handling cross-session data makes it suitable for long-term, real-world use, where emotions may vary across different sessions or environments. With its promising performance and adaptability, this emotion recognition system paves the way for future innovations in emotional AI and interactive technologies.

REFERENCES

1. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16000–16009).
2. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012–10022).
3. Sikka, K., Zhou, Z., & Picard, R. W. (2021). Cross-corpus speech emotion recognition using adversarial discriminative domain adaptation. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2021.3086303
4. Li, S., Deng, W., & Du, J. (2023). Deep learning for multimodal emotion recognition: A survey. *IEEE Transactions on Affective Computing, 14*(2), 1001–1025.
5. Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adam, M. (2022). Deep convolutional neural network for automated diagnosis of emotion using physiological signals. *Frontiers in Neuroscience*. https://doi.org/10.3389/fnins.2022.720456

CC BY NC ND

6. Ahmed, S. T., Basha, S. M., Ramachandran, M., Daneshmand, M., & Gandomi, A. H. (2023). An edge-AI-enabled autonomous connected ambulance-route resource recommendation protocol (ACA-R3) for eHealth in smart cities. *IEEE Internet of Things Journal*, *10*(13), 11497-11506.
7. Ahmed, S. T., Fathima, A. S., Nishabai, M., & Sophia, S. (2024). Medical ChatBot assistance for primary clinical guidance using machine learning techniques. *Procedia Computer Science*, *233*, 279-287.
8. Alotaibi, F., & Alotaibi, B. (2023). Edge-based real-time emotion recognition using deep learning. *Sensors, 23*(3), 1250.
9. Chen, T., & Zhang, M. (2021). A masked autoencoder framework for multimodal learning in emotion recognition. In *Proceedings of the ACM Multimedia Conference (MM)* (pp. 530–539).
10. Cireşan, D. C., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3642–3649).
11. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
12. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
13. He, Y., & Zhang, Z. (2021). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 2945–2954).
14. Hyang, P. (2024). Multi MAE-DER: Multimodal masked autoencoder for dynamic emotion recognition. *Preprint/Conference Proceedings (details not provided)*.
15. Jafar, H., & Ateya, A. (2020). Emotion recognition from EEG signals using deep learning models. *IEEE Access, 8*, 58933–58941.
16. Kim, H., & Lee, J. (2022). Transformer-based multimodal emotion recognition with attention mechanisms. *Information Fusion, 88*, 1–15.
17. Kim, S., & Lee, J. (2022). Deep learning-based emotion recognition using combined EEG and eye movement data. *Frontiers in Neuroscience, 16*, 720456.
18. Kumar, S. S., Ahmed, S. T., Sandeep, S., Madheswaran, M., & Basha, S. M. (2022). Unstructured Oncological Image Cluster Identification Using Improved Unsupervised Clustering Techniques. *Computers, Materials & Continua*, *72*(1).
19. Liu, Y., Sourina, O., & Nguyen, M. K. (2010). Real-time EEG-based human emotion recognition and visualization. In *Proceedings of the International Conference on Cyberworlds* (pp. 83–89).
20. Liu, Y., Sourina, O., & Nguyen, M. K. (2020). Real-time EEG-based human emotion recognition and visualization. *Cyberworlds, 32*(1), e5130.
21. Pang, M., Wang, H., Huang, J., Vong, C.-M., Ziqiang, & Chen, C. (2024). Multi-scale masked autoencoders for cross-session emotion recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. https://doi.org/10.1109/TNSRE.2024.XXXXXXX
22. Pang, M., Wang, H., Huang, J., Vong, C.-M., Ziqiang, & Chen, C. (2024). Multi-scale masked autoencoders for cross-session emotion recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.
23. Pasha, A., Ahmed, S. T., Painam, R. K., Mathivanan, S. K., Mallik, S., & Qin, H. (2024). Leveraging ANFIS with Adam and PSO optimizers for Parkinson's disease. *Heliyon*, *10*(9).
24. Periasamy, K., Periasamy, S., Velayutham, S., Zhang, Z., Ahmed, S. T., & Jayapalan, A. (2022). A proactive model to predict osteoporosis: An artificial immune system approach. *Expert Systems*, *39*(4), e12708.
25. Rani, P., Kumar, R., & Kumar, A. (2023). Emotion recognition using multimodal deep learning: A review and future directions. *Multimedia Tools and Applications, 82*(1), 1–35.
26. Sreedhar, K. S., Ahmed, S. T., & Sreejesh, G. (2022, June). An Improved Technique to Identify Fake News on Social Media Network using Supervised Machine Learning Concepts. In *2022 IEEE World Conference on Applied Intelligence and Computing (AIC)* (pp. 652-658). IEEE.
27. Tripathi, S., & Tripathi, R. K. (2023). Explainable AI for emotion recognition: A critical review. *Artificial Intelligence Review, 56*(5), 3897–3930.
28. Yang, Z., & Yu, H. (2022). Multimodal fusion for emotion recognition in speech and text. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2942–2946).
29. Zhao, Y., Li, X., & Zhang, H. (2023). Graph convolutional networks for facial emotion recognition: A comprehensive review. *Pattern Recognition, 138*, 109351.
30. Zhou, Z., & Zhang, J. (2021). Masking and predicting for multimodal emotion recognition. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 1224–1231).

CC BY NC ND