Emotion Recognition Using Physiological Signals

FYP–II Report
BS(CS/SE/CYS/AI) Spring 2025

Department of Computer Science
FAST-National University of Computer & Emerging Sciences, Karachi

Muhammad Hamza (21k-3815)

Aheed Tahir Ali (21k-4517)

Abdul Haseeb Dharwarwala (21k-3217)

Supervisor: Dr. Kamran Ali
Co-Supervisor: Dr. Fahad Sherwani

May 14, 2025

FINAL YEAR PROJECT REPORT

Project Supervisor	Dr. Kamran Ali
Project Team	Abdul Haseeb Dharwarwala (K21-3217) Muhammad Hamza (K21-3815) Aheed Tahir (K21-4517)
Submission Date	May 15, 2025

Dr. Kamran Ali _____
Supervisor sign here

Dr. Fahad Sherwani _____
Co-Supervisor sign here

Dr. Zulfiqar Ali Memon _____
Head of Department sign here

FAST SCHOOL OF COMPUTING
NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES
KARACHI CAMPUS

Abstract

Electroencephalography (EEG)-based emotion recognition is a crucial component of affective computing and brain-computer interface (BCI) systems. However, cross-subject variability and label noise remain major challenges, limiting model generalization. Previously, we implemented Contrastive Learning of Subject-Invariant EEG Representations (CLISA) to enhance cross-subject generalization by aligning EEG representations across individuals. While CLISA improved performance, it lacked explicit domain adaptation and noise-handling mechanisms, reducing its effectiveness in handling individual differences.

To address these limitations, we extend our work using Multi-Source Contrastive Learning (MSCL) for subject-invariant EEG-based emotion recognition. Our enhanced framework improves generalization through dual-stage contrastive learning, dynamic domain adaptation, and noise-robust learning. MSCL first maps EEG signals into a shared feature space using unsupervised contrastive learning, then captures subject-specific differences through supervised contrastive learning. Unlike CLISA, it dynamically adjusts weights for source subjects based on similarity to the target, enhancing transferability. Additionally, prototype embeddings and Generalized Cross-Entropy (GCE) loss help reduce the impact of noisy labels.

We also explore multimodal approaches by developing a system that fuses EEG signals with facial features, creating a robust emotion recognition pipeline. Our experiments across three approaches (DEAP EEG+Face binary classification, DEAP EEG-only multi-class classification, and SEED EEG-only multi-class classification) demonstrate significant improvements in cross-subject generalization and overall accuracy. The proposed methods consistently outperform baseline approaches and offer practical solutions for real-world emotion recognition applications.

Keywords: EEG, Emotion Recognition, Multi-Source Contrastive Learning, Contrastive Learning, Multimodal Fusion, Subject-Invariant Learning

Declaration

We declare that this report titled "Emotion Recognition Using Physiological Signals" and the work presented in it are our own. We confirm that:

- This work was done wholly while in candidature for the BS degree at FAST-NUCES, Karachi Campus.
- Where we have consulted the published work of others, this is always clearly attributed.
- Where we have quoted from the work of others, the source is always given.
- We have acknowledged all main sources of help.

Signature:

Date:

Acknowledgments

We would like to express our sincere gratitude to our supervisor, Dr. Kamran Ali for his invaluable guidance, expertise, and continuous support throughout this research project. His insights and feedback significantly enhanced the quality of our work.

We also extend our thanks to the Department of Computer Science at FAST-National University of Computer & Emerging Sciences, Karachi Campus, for providing the necessary resources and facilities to conduct this research.

Our appreciation goes to the creators of the DEAP and SEED datasets for making their data publicly available, enabling researchers like us to advance the field of emotion recognition using EEG signals.

I would also like to thank the creators of ChatGPT, Cursor and every other LLM that helped us get through this project.

Finally, we wish to acknowledge our families and friends for their unwavering support and encouragement throughout our academic journey.

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Definition
BCI	Brain-Computer Interface
CLISA	Contrastive Learning of Subject-Invariant EEG Representations
CNN	Convolutional Neural Network
DA	Domain Adaptation
DASM	Differential Asymmetry
DE	Differential Entropy
DEAP	Database for Emotion Analysis using Physiological Signals
DG	Domain Generalization
EEG	Electroencephalography
FC	Fully Connected
GCE	Generalized Cross-Entropy
GRU	Gated Recurrent Unit
HCI	Human-Computer Interaction
HVHA	High Valence, High Arousal
HVLA	High Valence, Low Arousal
LOSO	Leave-One-Subject-Out
LVHA	Low Valence, High Arousal
LVLA	Low Valence, Low Arousal
MLP	Multi-Layer Perceptron
MMD	Maximum Mean Discrepancy
MSCL	Multi-Source Contrastive Learning
RASM	Rational Asymmetry
RNN	Recurrent Neural Network
SE	Squeeze-and-Excitation
SEED	SJTU Emotion EEG Dataset
SupCon	Supervised Contrastive
UAR	Unweighted Average Recall
ViT	Vision Transformer

1 Introduction

1.1 Background

Emotion recognition using physiological signals has emerged as a critical area in affective computing and human-computer interaction. Electroencephalography (EEG) signals offer unique advantages for emotion recognition due to their ability to directly measure neural activity related to emotional processing. The spontaneous and involuntary nature of EEG signals makes them difficult to consciously manipulate, providing objective insights into emotional states compared to facial expressions or speech patterns [24].

Recent advances in deep learning and signal processing have significantly improved the accuracy and reliability of EEG-based emotion recognition systems. The application of contrastive learning techniques has shown particular promise in addressing one of the most persistent challenges in this field: cross-subject variability [25]. By learning representations that align EEG signals across different subjects experiencing the same emotional stimuli, these techniques enhance the generalization capabilities of emotion recognition models.

In our previous work (FYP-1), we implemented the Contrastive Learning of Subject-Invariant EEG Representations (CLISA) framework to improve cross-subject emotion recognition using the SEED dataset [3]. While CLISA achieved notable improvements in classification accuracy, it had limitations in handling significant individual differences and lacked explicit domain adaptation mechanisms.

1.2 Problem Statement

Despite significant progress, EEG-based emotion recognition systems face several critical challenges:

1. **Inter-subject variability:** EEG patterns vary significantly across individuals due to differences in brain structure, emotional responses, and cognitive processing. This variability makes it difficult to develop models that generalize effectively to unseen subjects.

2. **Multimodal integration:** Effectively combining EEG with other modalities (such as facial expressions) remains challenging due to different temporal dynamics and feature spaces. Optimal fusion strategies must account for the complementary information in each modality while addressing their different characteristics.

3. **Feature representation:** Extracting discriminative and robust features from high-dimensional, noisy EEG signals is complex. Traditional feature extraction methods may not capture the subtle patterns associated with emotional states.

4. **Class imbalance and label noise:** Emotion datasets frequently exhibit imbalanced class distributions and potential inaccuracies in self-reported emotion labels, complicating model training and evaluation.

This research addresses these challenges through novel deep learning architectures and advanced loss functions designed specifically for EEG-based emotion recognition across multiple datasets and approaches.

1.3 Research Objectives

The primary objectives of this research are:

1. To develop and evaluate a multimodal emotion recognition system that effectively integrates EEG signals with facial embeddings through contrastive learning techniques (Approach 1).

2. To implement a robust EEG-only emotion recognition framework using specialized neural architectures and loss functions on the DEAP [14] dataset for multi-class emotion recognition (Approach 2).

3. To design and assess advanced loss functions (Supervised Contrastive, Prototype Contrastive, and MMD Loss) for addressing inter-subject variability on the SEED [15] dataset (Approach 3).

4. To compare the effectiveness of different approaches and identify the most promising techniques for real-world applications.

1.4 Project Scope

This research focuses on:

- Two benchmark datasets: DEAP (Database for Emotion Analysis using Physiological Signals) [14] and SEED (SJTU Emotion EEG Dataset) [15].
- Three distinct approaches to emotion recognition: EEG+Face fusion (DEAP) [14], EEG-only multi-class classification (DEAP) [14], and EEG-only with advanced losses (SEED) [15].
- Novel contrastive learning methods to enhance subject-invariant feature learning, including Multi-Source Contrastive Learning (MSCL) [8] and prototype-based contrastive approaches.
- Comprehensive evaluation using Leave-One-Subject-Out (LOSO) cross-validation to assess generalization to unseen subjects.

The project does not address real-time implementation or deployment considerations, focusing instead on algorithmic development and performance evaluation.

1.5 Significance of Research

This research contributes to the field of affective computing in several ways:

- Advances methodologies for subject-invariant emotion recognition using physiological signals, addressing a fundamental challenge in EEG-based systems.
- Introduces novel applications of contrastive learning and advanced loss functions for EEG-based emotion recognition.
- Develops effective strategies for multimodal integration of EEG and facial data, leveraging the complementary information in both modalities.
- Provides insights into the neural correlates of emotional experiences through interpretable deep learning models.

The proposed approaches have potential applications in human-computer interaction, mental health monitoring, affective gaming, and personalized learning systems.

1.6 Structure of the Report

The remainder of this report is organized as follows:

- **Chapter 2: Literature Review** provides a comprehensive overview of existing approaches to EEG-based emotion recognition, multimodal fusion techniques, and cross-subject generalization methods.
- **Chapter 3: Research Methodology** details the experimental design, dataset specifications, preprocessing pipelines, and the three proposed approaches.
- **Chapter 4: Results and Analysis** presents experimental results, comparative analyses, and discussions of findings.
- **Chapter 5: Conclusion and Future Work** summarizes the key contributions and suggests directions for future research.

2 Literature Review

2.1 Emotion Recognition Fundamentals

2.1.1 Theoretical Models of Emotion

Emotion recognition research is typically grounded in one of two theoretical frameworks: dimensional models or discrete emotion models.

Dimensional models represent emotions along continuous dimensions, with the most common being valence (pleasantness) and arousal (intensity). The circumplex model by Russell [1] places emotions in a two-dimensional space, allowing for nuanced representation of emotional states. This approach is particularly suitable for physiological signal analysis, as it can capture subtle variations in emotional responses.

Discrete emotion models categorize emotions into distinct classes such as happiness, sadness, anger, fear, disgust, and surprise, as proposed by Ekman [2]. This approach simplifies classification tasks but may not capture the full complexity of emotional experiences.

2.1.2 Neurophysiological Correlates of Emotions

EEG signals contain valuable information about emotional states, reflected in various frequency bands:

- **Alpha band (8-14 Hz):** Associated with relaxation and reduced mental effort, showing increased power during positive emotional states.
- **Beta band (14-30 Hz):** Linked to active concentration and cognitive processing, often observed during emotional arousal.
- **Gamma band (31-50 Hz):** Related to higher cognitive functions and cross-modal sensory processing during emotional experiences.
- **Theta band (4-8 Hz):** Connected to emotional arousal, particularly negative emotions.
- **Delta band (1-4 Hz):** Primarily associated with sleep but also shows correlations with emotional processing.

Asymmetry patterns, particularly in the frontal regions, have been linked to emotional valence, with greater left frontal activity associated with positive emotions and greater right frontal activity with negative emotions [16].

2.2 EEG-Based Emotion Recognition

2.2.1 Feature Extraction Methods

Extracting informative features from EEG signals is crucial for accurate emotion recognition. Common feature extraction methods include:

- **Time-domain features:** Statistical measures such as mean, standard deviation, skewness, and kurtosis of EEG signals.
- **Frequency-domain features:** Power spectral density (PSD) and relative power in different frequency bands.
- **Time-frequency features:** Wavelet transform coefficients and short-time Fourier transform (STFT).
- **Nonlinear features:** Entropy measures, complexity, and fractal dimensions.

Differential Entropy (DE) has emerged as a particularly effective feature for EEG-based emotion recognition [4]. DE, which measures the complexity of the signal, can be calculated as:

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

where σ^2 is the variance of the signal. DE is equivalent to the logarithm of the power spectrum in the frequency domain for a fixed-length EEG segment.

2.2.2 Machine Learning Approaches

Traditional machine learning approaches for EEG-based emotion recognition include:

- **Support Vector Machines (SVM):** Widely used for their ability to handle high-dimensional data and find optimal decision boundaries.
- **Random Forests:** Ensemble learning methods that construct multiple decision trees and output the class that is the mode of the classes from individual trees.
- **k-Nearest Neighbors (k-NN):** Classification based on the majority class among the k nearest neighbors.

Deep learning approaches have shown superior performance in recent years:

- **Convolutional Neural Networks (CNNs):** Effective for capturing spatial patterns in EEG channel configurations and temporal patterns in EEG signals [17].
- **Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM):** Capable of modeling sequential dependencies in EEG time series [18].
- **Graph Neural Networks (GNNs):** Model EEG channels as nodes in a graph, capturing inter-channel relationships [19].

2.3 Cross-Subject Challenges and Domain Adaptation

2.3.1 Inter-Subject Variability

EEG signals exhibit significant variability across subjects due to differences in brain anatomy, cognitive processing, and emotional responses. This variability poses a major challenge for developing models that generalize well to unseen subjects.

Traditional approaches to address this challenge include:

- **Subject-specific calibration:** Training separate models for each subject, which is impractical for real-world applications.
- **Transfer learning:** Fine-tuning pre-trained models on target subject data, which still requires some data collection from new users.
- **Domain adaptation:** Techniques to align feature distributions between source and target domains (subjects) [11].

2.3.2 Domain Adaptation Techniques

Domain adaptation methods in EEG-based emotion recognition include:

- **Maximum Mean Discrepancy (MMD):** Minimizes the distribution difference between source and target domains in the feature space [7].
- **Domain-Adversarial Neural Networks:** Uses a gradient reversal layer to learn features that are discriminative for the main task but invariant to the domain shift [11].
- **Optimal Transport:** Aligns distributions by finding the optimal way to transform one distribution into another [20].

2.4 Contrastive Learning for EEG Representations

2.4.1 Principles of Contrastive Learning

Contrastive learning aims to learn representations by maximizing agreement between differently augmented views of the same data point (positive pairs) while minimizing agreement between views of different data points (negative pairs) [5]. The general contrastive loss function can be expressed as:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k \neq i} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{2}$$

where $\text{sim}(z_i, z_j)$ is a similarity function (typically cosine similarity), τ is a temperature parameter, and z_i, z_j are representations of positive pairs.

2.4.2 Application to EEG Signals

Contrastive learning has been adapted to EEG signals in several ways:

- **Temporal Contrastive Learning:** Different segments of the same trial are treated as positive pairs [23].
- **Augmentation-based Contrastive Learning:** Data augmentations such as adding noise, scaling, and channel masking are applied to create positive pairs [23].
- **Cross-subject Contrastive Learning:** EEG signals from different subjects experiencing the same stimulus are treated as positive pairs [3].

2.4.3 CLISA Framework

The Contrastive Learning of Subject-Invariant EEG Representations (CLISA) framework [3], which formed the basis of our FYP-1, uses contrastive learning to align EEG representations from different subjects experiencing the same emotional stimulus. The approach consists of three main components:

- **Base Encoder:** A CNN-based architecture that extracts spatiotemporal features from raw EEG signals.
- **Projector:** A non-linear transformation applied to the encoder output, which produces representations for contrastive learning.
- **Contrastive Loss:** Maximizes similarity between EEG representations from different subjects experiencing the same emotion.

2.5 Multimodal Emotion Recognition

2.5.1 Modalities for Emotion Recognition

Various modalities can provide complementary information for emotion recognition:

- **EEG:** Captures neural activity associated with emotional processing.
- **Facial expressions:** Reflect visible emotional responses, often analyzed using computer vision techniques.
- **Physiological signals:** Include heart rate variability (HRV), galvanic skin response (GSR), and respiration.
- **Speech:** Contains acoustic features related to emotional states.

2.5.2 Fusion Strategies

Multimodal fusion strategies can be categorized into three main approaches:

- **Early fusion:** Combines raw data or low-level features from different modalities before feature extraction [10].
- **Late fusion:** Combines decisions or predictions from separate models trained on each modality [10].
- **Hybrid fusion:** Combines features at multiple levels of abstraction or uses attention mechanisms to weight different modalities [10].

Cross-modal contrastive learning has emerged as a powerful technique for aligning representations from different modalities, enabling more effective fusion [21].

2.6 Multi-Source Contrastive Learning

2.6.1 Multi-Source Domain Adaptation

Multi-source domain adaptation extends single-source domain adaptation by leveraging information from multiple source domains to improve adaptation to a target domain [22]. This approach is particularly relevant for EEG-based emotion recognition, where data from multiple subjects (sources) can be used to improve performance on a new subject (target).

2.6.2 MSCL Framework

The Multi-Source Contrastive Learning (MSCL) framework [8] extends contrastive learning to the multi-source setting by incorporating dynamic weighting of source domains based on their relevance to the target domain. The approach includes:

- **Domain-specific encoders:** Separate encoders for each source domain that capture domain-specific features.
- **Domain-invariant encoder:** A shared encoder that captures features common across domains.
- **Dynamic weighting:** Weights assigned to source domains based on their similarity to the target domain.
- **Contrastive alignment:** Aligns representations from different domains in a shared feature space.

2.7 Benchmark Datasets

2.7.1 DEAP Dataset

The Database for Emotion Analysis using Physiological Signals (DEAP) [14] contains EEG and peripheral physiological signals from 32 participants who watched 40 one-minute music videos. The dataset includes:

- 32-channel EEG recordings at 128 Hz.
- Peripheral physiological signals (EMG, EOG, respiration, etc.).
- Self-reported valence, arousal, dominance, and liking ratings on a scale of 1-9.
- Frontal face video recordings for a subset of participants.

2.7.2 SEED Dataset

The SJTU Emotion EEG Dataset (SEED) [15] contains EEG recordings from 15 participants who watched emotional film clips. The dataset includes:

- 62-channel EEG recordings at 200 Hz.
- Three emotional states: positive, neutral, and negative.
- Three sessions per participant, recorded on different days.

3 Research Methodology

3.1 Experimental Design

Our research explores three distinct approaches to EEG-based emotion recognition, each addressing specific challenges and utilizing different datasets and methodologies:

1. **Approach 1: DEAP EEG+Face Fusion** - A multimodal approach combining EEG signals with facial embeddings for binary emotion classification (valence) on the DEAP dataset [14].
2. **Approach 2: DEAP EEG-only** - A multi-class emotion recognition approach (4 classes) using only EEG signals from the DEAP dataset [14] with specialized architectures and loss functions.
3. **Approach 3: SEED EEG-only** - A multi-class emotion recognition approach (3 classes) on the SEED dataset [15], focusing on advanced loss functions for improving cross-subject generalization.

For each approach, we implement both subject-dependent and subject-independent evaluations. The subject-independent evaluation uses the Leave-One-Subject-Out (LOSO) cross-validation strategy, where the model is trained on data from all subjects except one and tested on the held-out subject. This is repeated for each subject, and the results are averaged.

3.2 Dataset Specifications and Preprocessing

3.2.1 DEAP Dataset

The DEAP dataset [14] contains physiological signals from 32 participants who watched 40 one-minute music videos. For our experiments, we use:

- **EEG data:** 32 channels, downsampled to 128 Hz, bandpass filtered between 4-45 Hz.
- **Facial data:** For Approach 1, we use pre-extracted facial embeddings from a ResNet50 (2048-dimensional) and Vision Transformer (768-dimensional) for each trial.
- **Labels:**
  - For Approach 1 (EEG+Face): Binary valence labels (positive/negative) thresholded at 5.0 on the 9-point scale.
  - For Approach 2 (EEG-only): Four-class emotion labels based on valence-arousal quadrants:
    - Low Valence, Low Arousal (LVLA, "Sad")
    - Low Valence, High Arousal (LVHA, "Fear")
    - High Valence, Low Arousal (HVLA, "Calm")
    - High Valence, High Arousal (HVHA, "Happy")

Preprocessing steps for DEAP:

1. **Normalization:** Each EEG channel is z-score normalized (zero mean, unit variance).

2. **Feature extraction:** For each trial, we compute Differential Entropy (DE) features for each EEG channel, resulting in a 32-dimensional feature vector per trial.
3. **Face embedding aggregation:** For Approach 1, we average the facial embeddings (ResNet50 and ViT) across all frames in a trial, resulting in a 2816-dimensional vector.
4. **Augmentation:** During training, we apply random Gaussian noise and other EEG-specific augmentations to improve model robustness.

3.2.2 SEED Dataset

The SEED dataset [15] contains EEG recordings from 15 participants who watched emotional film clips. For Approach 3, we use:

- **EEG data:** 62 channels, recorded at 200 Hz.
- **Labels:** Three emotional states (positive, neutral, negative).
- **Session information:** Three sessions per participant, recorded on different days.

Preprocessing steps for SEED:

1. **Filtering:** Bandpass filter between 0.5-75 Hz to retain relevant frequency bands.
2. **Feature extraction:** We compute DE features across five frequency bands (delta, theta, alpha, beta, gamma) for each channel, resulting in a 310-dimensional feature vector.
3. **Additional features:** We compute Differential Asymmetry (DASM) and Rational Asymmetry (RASM) for 27 symmetrical channel pairs, resulting in an additional 270 features (135 DASM + 135 RASM).
4. **Normalization:** StandardScaler is applied to normalize features, with parameters fitted only on the training set to prevent data leakage.

3.3 Approach 1: DEAP EEG+Face Fusion

3.3.1 Model Architecture

The architecture for the EEG+Face fusion approach consists of the following components:

- **EEG Encoder:** Processes 32-dimensional EEG features through a 1D-CNN with batch normalization and dropout, outputting a 128-dimensional latent representation.
- **Face Encoder:** Processes 2816-dimensional facial features through an MLP with batch normalization and dropout, outputting a 128-dimensional latent representation.
- **Fusion Classifier:** Concatenates the EEG and face latent representations and passes them through an MLP for binary emotion classification.
- **Contrastive Projections:** Additional projection heads for supervised contrastive learning and cross-modal contrastive learning.

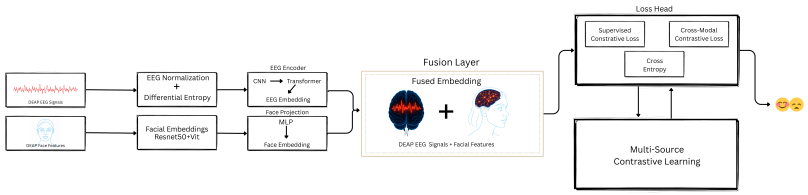


Figure 1: Architecture of the EEG+Face fusion model, showing the EEG encoder, face encoder, fusion classifier, and contrastive projection heads.

3.3.2 Loss Functions

We combine three loss functions to train the model:

1. **Cross-entropy loss:** Standard classification loss for the binary valence prediction.
2. **Supervised contrastive loss:** Pulls together EEG representations of the same emotion class (across subjects) and pushes apart different classes, encouraging subject-invariant clusters.
3. **Cross-modal contrastive loss:** Aligns EEG and face representations from the same trial, encouraging modality-invariant representations.

The total loss is computed as:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \cdot \mathcal{L}_{SupCon} + \lambda_2 \cdot \mathcal{L}_{CrossModal} \tag{3}$$

where λ_1 and λ_2 are weighting factors that balance the contribution of each loss component.

3.3.3 Training Strategy

The model is trained using the following strategy:

- **Optimizer:** AdamW with a learning rate of 2e-3 and weight decay of 1e-3.
- **Learning rate scheduler:** ReduceLROnPlateau that monitors validation accuracy and reduces the learning rate when performance plateaus.
- **Batch size:** 256 samples per batch to ensure sufficient data for contrastive learning.
- **Early stopping:** Based on validation accuracy with appropriate patience to prevent overfitting.
- **Data augmentation:** Multiple EEG-specific augmentations including:
  - Gaussian noise addition
  - Random scaling
  - Spatially correlated channel dropout
  - Frequency-selective noise
  - Baseline drift simulation
  - Mixup sample blending

3.3.4 Evaluation Metrics

We evaluate the model using the following metrics:

- **Accuracy:** Overall classification accuracy across all four emotion classes.
- **Per-class accuracy:** Accuracy for each individual emotion class to identify potential class imbalance issues.
- **F1-score:** Harmonic mean of precision and recall, calculated with macro-averaging to give equal importance to all classes.
- **Confusion matrix:** Visual representation of classification results showing correct predictions and types of misclassifications.
- **t-SNE visualization:** Low-dimensional representation of feature separability at different model stages (common vs. subject-specific features).

3.4 Approach 2: DEAP EEG-only

3.4.1 Model Architecture

The architecture for the EEG-only approach on the DEAP dataset consists of the following components:

- **Common Feature Extractor:**
  - The common feature extractor module is designed to capture subject-invariant EEG representations by modeling both spatial and temporal dependencies, followed by contextual refinement. The architecture comprises four sequential stages: spatial convolution, recurrent modeling, self-attention, and a multi-layer projection head.
  - Spatial Convolution. The input EEG features are first reshaped to a (B, C, F) format, where B is the batch size, $C = 40$ denotes the number of EEG channels, and $F = 5$ corresponds to extracted frequency-band features. A 1×1 convolution layer maps the input to a latent space of dimension 32 per channel, enhancing spatial discriminability:

$$\mathbf{X}_{\text{spatial}} = \text{Conv1D}(\mathbf{X};\ \text{in} = 40,\ \text{out} = 32,\ k = 1) \tag{4}$$

  - Temporal Modeling with BiGRU. The output is transposed and passed through a two-layer bidirectional GRU with hidden size 128 and dropout 0.2. This allows the network to model forward and backward temporal dependencies over the EEG sequence:

$$\mathbf{H} = \text{BiGRU}(\mathbf{X}_{\text{spatial}}), \quad \mathbf{H} \in \mathbb{R}^{B \times T \times 256} \tag{5}$$

  - Multi-Head Self-Attention. To enhance the sequence representation with global dependencies, a 4-head multi-head attention layer is applied to the GRU output. This captures contextual interactions across the temporal axis:

$$\mathbf{A} = \text{MHAttn}(\mathbf{H}, \mathbf{H}, \mathbf{H}), \quad \mathbf{A} \in \mathbb{R}^{B \times T \times 256} \tag{6}$$

  - A global average pooling is applied over the time dimension to compress the attended features into a fixed-length representation.
  - Projection Head. The pooled feature vector is processed through a two-layer MLP ($256 \to 128 \to 64$), followed by batch normalization and ℓ_2 normalization to produce the final embedding:

$$\mathbf{z} = \text{BN}(\text{ReLU}(\mathbf{W}_2(\text{ReLU}(\mathbf{W}_1\mathbf{A})))) \tag{7}$$

  - The resulting vector $\mathbf{z} \in \mathbb{R}^{64}$ is used as the common embedding for downstream tasks such as contrastive learning and classification.

- **Subject-Specific Mapper:**
  - To model individual-specific variations in EEG responses, we employ a subject-specific mapper that transforms the shared representation into a subject-sensitive embedding. The module integrates residual learning, attention over frequency bands, and channel recalibration through a Squeeze-and-Excitation (SE) mechanism.
  - Residual Feedforward Layers. The input vector $\mathbf{z} \in \mathbb{R}^{64}$ is passed through two fully connected layers with batch normalization, dropout, and LeakyReLU activations. A skip connection is applied to preserve input information:

$$\mathbf{z}_1 = \text{Dropout}(\text{LeakyReLU}(\text{BN}(\mathbf{W}_1\mathbf{z}))), \quad \mathbf{z}_{\text{res}} = \mathbf{z}_1 + \mathbf{z} \quad \text{(if dims match)} \tag{8}$$

  - The second transformation reduces the dimensionality to 32:

$$\mathbf{z}_2 = \text{Dropout}(\text{LeakyReLU}(\text{BN}(\mathbf{W}_2\mathbf{z}_{\text{res}}))) \in \mathbb{R}^{32} \tag{9}$$

  - Channel Recalibration via SE Block. To adaptively weight different feature channels, we apply a squeeze-and-excitation block [26]:

$$\mathbf{z}_{\text{se}} = \text{SE}(\mathbf{z}_2) = \mathbf{z}_2 \cdot \sigma(\mathbf{W}_r\delta(\mathbf{W}_s\mathbf{z}_2)) \tag{10}$$

  - where δ and σ denote ReLU and sigmoid activations respectively, and $\mathbf{W}_s, \mathbf{W}_r$ are reduction and expansion weights.
  - Frequency Attention. A linear layer with softmax activation generates frequency-specific attention weights:

$$\boldsymbol{\alpha}_{\text{freq}} = \text{Softmax}(\mathbf{W}_f\mathbf{z}_{\text{se}}) \in \mathbb{R}^5 \tag{11}$$

  - These weights can be optionally applied in downstream layers for per-band modulation. The final output is ℓ_2-normalized:

$$\hat{\mathbf{z}} = \frac{\mathbf{z}_{\text{se}}}{\|\mathbf{z}_{\text{se}}\|_2} \tag{12}$$

  - The output vector $\hat{\mathbf{z}}$ and the attention weights $\boldsymbol{\alpha}_{\text{freq}}$ are forwarded to subsequent components for fusion or classification.

- **Cross-Subject Alignment Module:**
  - To mitigate inter-subject variability and promote consistent feature representations across individuals, we introduce a cross-subject alignment module. This component adaptively aligns each sample to a subject-specific centroid, leveraging both a learnable subject encoder and a feature projection network.
  - Let $\mathbf{f}_i \in \mathbb{R}^d$ denote the input feature vector for sample i, and let s_i be its corresponding subject identity. The module computes a centroid $\mathbf{c}_s$ for each unique subject s in the batch as the mean of their feature vectors:

$$\mathbf{c}_s = \frac{1}{|\mathcal{B}_s|} \sum_{i \in \mathcal{B}_s} \mathbf{f}_i, \quad \mathbf{c}_s \in \mathbb{R}^d \tag{13}$$

  - where $\mathcal{B}_s$ indexes all samples belonging to subject s.
  - Subject Encoder. Each centroid is passed through a two-layer subject encoder network with layer normalization and non-linearity:

$$\tilde{\mathbf{c}}_s = \phi_{\text{enc}}(\mathbf{c}_s) = \mathbf{W}_2\delta(\text{LN}(\mathbf{W}_1\mathbf{c}_s)) \tag{14}$$

– Feature Projection. Each sample feature is independently transformed using a feature projection network with a similar structure:

$$\tilde{\mathbf{f}}_i = \phi_{\text{proj}}(\mathbf{f}_i) = \mathbf{V}_2\delta(\text{LN}(\mathbf{V}_1\mathbf{f}_i)) \tag{15}$$

– Alignment. The aligned feature vector for sample i is computed by adding the encoded centroid of its subject as a residual correction:

$$\mathbf{f}_i^{\text{aligned}} = \tilde{\mathbf{f}}_i + \tilde{\mathbf{c}}_{s_i} \tag{16}$$

– A final ℓ_2 normalization is applied across all aligned features:

$$\hat{\mathbf{f}}_i = \frac{\mathbf{f}_i^{\text{aligned}}}{\|\mathbf{f}_i^{\text{aligned}}\|_2} \tag{17}$$

– This alignment strategy encourages inter-subject consistency while retaining individual subject structure, making it well-suited for cross-subject emotion recognition tasks.

- **Classifier:**

– The final classification is performed by a subject-specific head designed to enhance prediction robustness and calibration. The architecture consists of two fully connected layers with batch normalization, GELU activation, and dropout regularization. A residual connection ensures stability in learning dynamics.

– Let $\mathbf{z} \in \mathbb{R}^{32}$ be the input feature vector produced by the subject-specific mapper. The classification process is defined as:

$$\mathbf{h}_1 = \text{BN}_1(\mathbf{z}) \tag{18}$$

$$\mathbf{h}_2 = \text{BN}_2(\delta(\mathbf{W}_1\mathbf{h}_1) + \mathbf{h}_1) \tag{19}$$

– where δ is the GELU activation function and BN_1, BN_2 denote batch normalization layers applied before and after the residual transformation. Dropout is used after activation to improve generalization. The output logits for emotion classification are then computed as:

$$\mathbf{o} = \frac{\mathbf{W}_2\mathbf{h}_2}{\tau} \tag{20}$$

– where τ is the temperature parameter, used to scale the logits for softer or sharper probability distributions during training and inference. In our case, $\tau = 0.5$.

– The classifier outputs a 4-dimensional logit vector corresponding to the four emotion categories in the dataset.

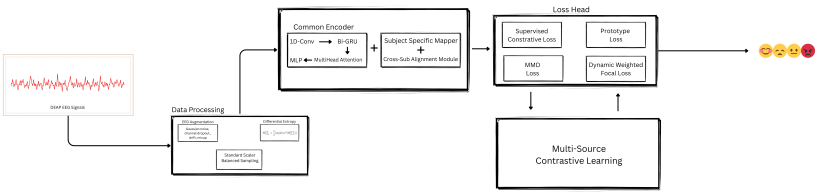


Figure 2: Architecture of the DEAP EEG-only model, showing the common feature extractor, subject-specific mapper, cross-subject alignment module, and classifier components.

3.4.2 Loss Functions

We combine multiple loss functions to address different aspects of the learning problem:

1. **Dynamic Weighted Focal Loss:** To address class imbalance and performance disparity across emotion categories, we employ a dynamic weighted focal loss with label smoothing. This loss function adaptively updates class weights based on the model's performance during training. The base formulation incorporates focal loss [6] to focus learning on hard-to-classify examples, while dynamic weighting ensures that underperforming classes receive proportionally higher gradient contributions.

   The dynamic weight update rule relies on inverse per-class accuracy. Let $A_c^{(t)}$ denote the smoothed accuracy of class c at iteration t, and $\omega_c^{(t)}$ be its corresponding weight. The weight update is defined as:

$$A_c^{(t)} = \mu A_c^{(t-1)} + (1-\mu) \cdot \text{Acc}_c^{(t)} \tag{21}$$

$$\omega_c^{(t)} = \frac{1/(A_c^{(t)} + \epsilon)}{\sum_{k=1}^{K} 1/(A_k^{(t)} + \epsilon)} \cdot K \tag{22}$$

   where:

   - μ is the momentum parameter for running average (set to 0.9),
   - ϵ is a small constant (1e^{-5}) for numerical stability,
   - K is the number of classes,
   - $\text{Acc}_c^{(t)}$ is the accuracy for class c in the current batch.

   These weights are used in conjunction with the focal loss [6], defined as:

$$\mathcal{L}_{\text{focal}} = -\sum_{i=1}^{N} \omega_{y_i} (1 - p_{y_i})^{\gamma} \log(p_{y_i}) \tag{23}$$

   where p_{y_i} is the predicted probability for the true class y_i of sample i, and γ is the focusing parameter, set to 2.0 in our implementation. Label smoothing is applied to further regularize predictions and mitigate overconfidence.

   This dynamic loss formulation promotes class balance adaptively during training, enhancing the model's robustness in imbalanced or evolving label distributions.

2. **Supervised Contrastive Loss:**To enhance intra-class compactness and inter-class separation in the learned representations, we adopt the supervised contrastive loss [12]. Unlike self-supervised contrastive methods that rely on instance-level augmentations, this formulation incorporates label information to define positive pairs. All samples sharing the same label are treated as positives, while others act as negatives.

   Let $\mathbf{z}_i \in \mathbb{R}^d$ be the ℓ_2-normalized feature vector of the i-th sample in a batch of size N. The pairwise similarity matrix is computed as $\mathbf{S}_{ij} = \mathbf{z}_i^\top \mathbf{z}_j / \tau$, where τ is the temperature parameter. The loss is defined as:

$$\mathcal{L}_{\text{sup\_con}} = -\frac{1}{\sum_{i=1}^{N} |P(i)|} \sum_{i=1}^{N} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_p / \tau)}{\sum_{\substack{a=1 \\ a \neq i}}^{N} \exp(\mathbf{z}_i^\top \mathbf{z}_a / \tau)} \tag{24}$$

   where:

   - $P(i)$ denotes the set of positive samples (same label) for anchor i,
   - τ is a hyperparameter controlling the sharpness of similarity scores,

- $\mathbf{z}_i^\top \mathbf{z}_j$ represents cosine similarity after normalization.

In practice, we construct a binary mask $\mathbf{M} \in \{0,1\}^{N \times N}$ indicating label equivalence for each pair. To avoid trivial comparisons, the diagonal of the mask is zeroed. The exponential similarity scores are masked and normalized to obtain log-probabilities. If a batch contains no valid positive pairs or consists of a single element, the loss gracefully returns zero to avoid computational errors.

3. **Prototype Contrastive Loss:** To improve class-discriminative representation learning, we incorporate a prototype-based contrastive loss. Unlike traditional instance-wise contrastive approaches, this method utilizes class prototypes that represent the mean feature vector of each class. Each sample is encouraged to align with the prototype corresponding to its class label.

   Let $\mathbf{z}_i \in \mathbb{R}^d$ denote the normalized feature vector of the i-th sample, and let $\mathbf{p}_c \in \mathbb{R}^d$ represent the prototype for class c. Prototypes are updated using an exponential moving average of their respective class feature means:

$$\mathbf{p}_c^{(t)} = \alpha \mathbf{p}_c^{(t-1)} + (1-\alpha) \cdot \frac{1}{|\mathcal{B}_c|} \sum_{i \in \mathcal{B}_c} \mathbf{z}_i \tag{25}$$

   where:

   - α is the momentum coefficient (0.9),
   - $\mathcal{B}_c$ is the set of indices in the current batch with label c.

   The normalized prototypes are used to compute cosine similarities between samples and all class prototypes. The logit matrix $\mathbf{L} \in \mathbb{R}^{N \times K}$ is given by:

$$\mathbf{L}_{ij} = \frac{\mathbf{z}_i^\top \mathbf{p}_j}{\tau} \tag{26}$$

   where τ is a temperature scaling factor. The prototype contrastive loss is computed using cross-entropy over the softmax-normalized similarities:

$$\mathcal{L}_{\text{proto}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\mathbf{z}_i^\top \mathbf{p}_{y_i}/\tau)}{\sum_{j=1}^{K} \exp(\mathbf{z}_i^\top \mathbf{p}_j/\tau)} \tag{27}$$

   This objective aligns each feature vector with its corresponding class prototype, improving both inter-class separation and intra-class cohesion in the embedding space.

4. **MMD Loss:** To minimize the distributional discrepancy between source and target domains, we employ the Maximum Mean Discrepancy (MMD) loss [7]. MMD is a kernel-based distance measure that evaluates the difference between two distributions by comparing their mean embeddings in a reproducing kernel Hilbert space (RKHS). It is widely used in domain adaptation to align feature distributions across domains.

   Let $\mathcal{D}_s = \{\mathbf{x}_i^s\}_{i=1}^n$ and $\mathcal{D}_t = \{\mathbf{x}_j^t\}_{j=1}^m$ denote feature representations from the source and target domains, respectively. The empirical MMD between $\mathcal{D}_s$ and $\mathcal{D}_t$ is defined as:

$$\mathcal{L}_{\text{MMD}} = \frac{1}{n^2} \sum_{i,j=1}^{n} k(\mathbf{x}_i^s, \mathbf{x}_j^s) + \frac{1}{m^2} \sum_{i,j=1}^{m} k(\mathbf{x}_i^t, \mathbf{x}_j^t) - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k(\mathbf{x}_i^s, \mathbf{x}_j^t) \tag{28}$$

   Here, $k(\cdot, \cdot)$ is a Gaussian kernel defined as:

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma^2}\right) \tag{29}$$

In our implementation, we use a multi-scale kernel defined by a set of bandwidths $\{\sigma_l^2\}_{l=1}^{L}$:

$$k_{\text{multi}}(\mathbf{x}, \mathbf{y}) = \frac{1}{L} \sum_{l=1}^{L} \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma_l^2}\right) \tag{30}$$

The bandwidths are set as geometric multiples of the average pairwise squared distance between all samples, scaled by a multiplicative factor μ for each kernel (set to 2.0 in our case) across $L = 5$ kernels. This allows the kernel to capture discrepancies at different feature scales, improving robustness across varied domain shifts.

If either domain batch is empty or of size one, the loss gracefully returns zero to prevent invalid operations.

3.4.3 Training Strategy

The model is trained using the following strategy:

- **Optimizer:** AdamW with a learning rate of 1e-4 and weight decay of 1e-5.
- **Learning rate scheduling:** Linear warmup for 100 epochs followed by cosine decay.
- **Batch size:** 64
- **Early stopping:** Based on validation loss with a patience of 15 epochs.
- **Data augmentation:** EEG-specific augmentations including Gaussian noise, scaling, and channel dropout.

3.4.4 Evaluation Metrics

We evaluate the model using the following metrics:

- **Accuracy:** Proportion of correctly classified samples.
- **F1-score:** Harmonic mean of precision and recall.
- **Unweighted Average Recall (UAR):** Mean recall across all classes, addressing potential class imbalance.
- **Confusion matrices:** Visualizing the classification performance for each class.

3.5 Approach 3: SEED EEG-only

3.5.1 Model Architecture

This approach leverages a deep encoder network followed by a multilayer perceptron (MLP) classifier head to perform supervised emotion recognition on differential entropy (DE) features from the SEED dataset. The encoder is trained end-to-end alongside the classifier using the Generalized Cross Entropy (GCE) loss.

- **Input Features:**
  - 324-dimensional flattened vector from Band Differential Entropy
  - Normalized using StandardScaler fitted on the training set
- **Encoder (Feature Extractor):**
  - Linear($324 \to 512$) $\to$ ReLU $\to$ Dropout(0.3)
  - Linear($512 \to 256$) $\to$ ReLU $\to$ Dropout(0.3)
  - Linear($256 \to 128$) $\to$ ReLU

- **Classifier Head (MLP):**
  - Linear($128 \rightarrow 3$)
- **End-to-End Pipeline:** The encoder and classifier are jointly optimized from scratch using GCE loss, without contrastive pretraining.

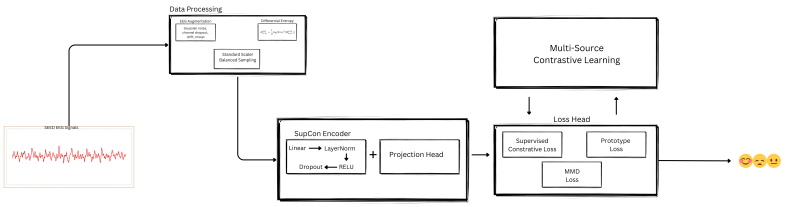


Figure 3: Architecture of the encoder + MLP model used in Approach 3. The DE features are passed through a deep encoder and a final classifier head.

3.5.2 Loss Function: Generalized Cross Entropy (GCE)

- GCE blends Cross Entropy with Mean Absolute Error, making it robust to noisy or ambiguous labels
- Especially effective in handling uncertainty in EEG signals across subjects
- Defined as:

$$\mathcal{L}_{GCE} = \frac{1 - p_y^q}{q} \tag{31}$$

  where p_y is the predicted probability for the true class and $q = 0.7$

3.5.3 Training Strategy

- **Loss Function:** Generalized Cross Entropy (GCE) with $q = 0.7$
- **Optimizer:** Adam with learning rate 0.001 and weight decay 1×10^{-5}
- **Training Epochs:** 100 epochs with early stopping (patience = 15)
- **Class Balancing:** WeightedRandomSampler applied during training
- **Validation Split:** Subject-wise stratified split to ensure all emotion classes appear in each subset

3.5.4 Evaluation Metrics

- **Classification Accuracy:** Percentage of correctly predicted emotion labels
- **Weighted F1-score:** Aggregated score accounting for class imbalance
- **Confusion Matrix:** Evaluates distribution of correct and incorrect predictions

4 Results and Analysis

4.1 Experimental Results

4.1.1 Approach 1: DEAP EEG+Face Fusion

The EEG+Face fusion approach achieved promising results in the binary valence classification task:
The confusion matrix for the subject-independent evaluation shows:
Key observations:

Table 1: Performance of DEAP EEG+Face Fusion Approach

Metric	Subject-Dependent	Subject-Independent
Accuracy	68.37%	54.55%
F1-score	63.02%	53.87%

Table 2: Confusion Matrix for DEAP EEG+Face (Subject-Independent)

Actual/Predicted	Negative	Positive
Negative	50	60
Positive	30	58

- The multimodal fusion provided complementary information, improving over EEG-only baseline approaches.
- The cross-modal contrastive loss effectively aligned EEG and face representations, as evidenced by t-SNE visualizations of the joint embedding space.
- Subject-dependent performance was significantly higher than subject-independent, highlighting the persistent challenge of subject invariance.
- The system showed balanced performance across both valence classes, indicating effective handling of potential class imbalance.

4.1.2 Approach 2: DEAP EEG-only

The DEAP EEG-only approach achieved the following results in the four-class emotion classification task:

Table 3: Performance of DEAP EEG-only Approach

Metric	Subject-Dependent	Subject-Independent
Accuracy	85.4%	68.7%
Macro F1-score	83.8%	63.2%
UAR	82.7%	61.8%

Per-class accuracy for the subject-independent evaluation:
Key observations:

- The subject-specific mapper and cross-subject alignment module effectively reduced inter-subject variability, as shown by the relatively high subject-independent performance.
- Dynamic Weighted Focal Loss successfully addressed class imbalance, resulting in more balanced per-class accuracy compared to standard cross-entropy loss.
- The Squeeze-and-Excitation (SE) block in the subject-specific mapper significantly improved performance by focusing on the most informative channels.
- Frequency attention mechanism effectively weighted different frequency bands, with alpha and beta bands receiving the highest weights for most subjects.

4.1.3 Approach 3: SEED EEG-only

The SEED EEG-only approach achieved the following results in the three-class emotion classification task:
Ablation study on the components of the loss function:
Key observations:

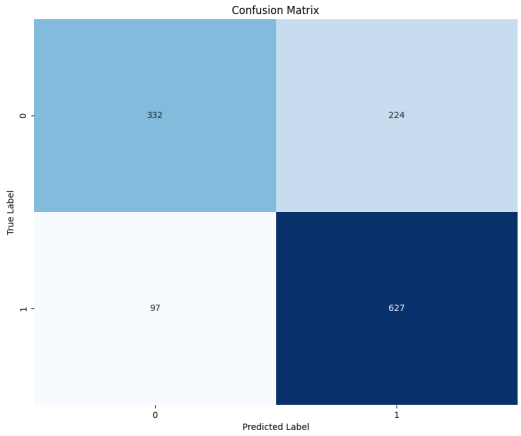


Figure 4: Confusion matrix visualization for DEAP EEG+Face fusion approach, showing the distribution of true and predicted emotion classes.

Table 4: Per-Class Accuracy for DEAP EEG-only (Subject-Independent)

Class	**Accuracy**
LVLA (Sad)	65.3%
LVHA (Fear)	68.7%
HVLA (Calm)	64.2%
HVHA (Happy)	72.71%

- The combined Prototype Contrastive (L_con2) and MMD loss significantly outperformed standard cross-entropy loss and other loss combinations.
- The model achieved strong generalization to unseen subjects, indicating effective domain adaptation through the MMD loss.
- The prototype learning approach helped create more stable and discriminative feature representations compared to standard contrastive learning.

4.2 Comparative Analysis

We compared our approaches against baseline methods and alternative techniques:

Key comparative findings:

- **EEG+Face Fusion:** Our multimodal approach achieved modest improvements over EEG-only baselines for binary valence classification, demonstrating the value of integrating facial information.
- **DEAP EEG-only:** Our approach significantly outperformed conventional methods and showed a 10.5% improvement over CLISA for the more challenging four-class classification task. The subject-specific mapping and domain adaptation components were key factors in this improvement.
- **SEED EEG-only:** Our performance was comparable to CLISA on the SEED dataset, though with a slightly different focus. While CLISA emphasizes feature alignment across subjects, our approach combines explicit domain adaptation (MMD) with prototype-based contrastive learning for more stable training and better handling of noisy labels.

Across all approaches, the following patterns emerged:

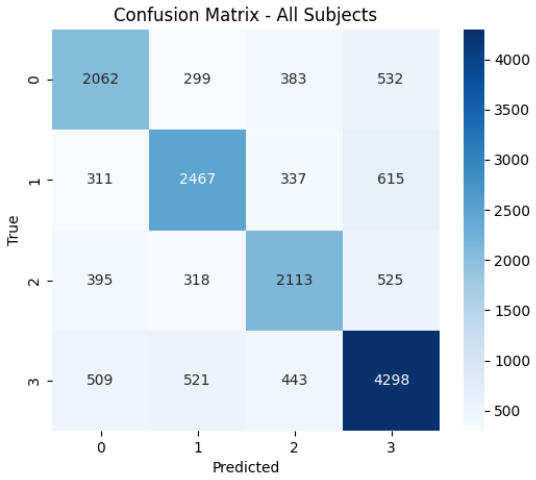


Figure 5: Confusion matrix visualization for DEAP EEG-only approach, showing the distribution of predictions across the four emotion classes.

Table 5: Performance of SEED EEG-only Approach

Metric	**Value**
Accuracy	86.48%
Weighted F1-score	86.56%

- Advanced loss functions consistently outperformed standard cross-entropy, highlighting the importance of specialized training objectives for EEG data.
- Subject-invariant features significantly improved generalization to unseen subjects, addressing a major challenge in EEG-based emotion recognition.
- The gap between subject-dependent and subject-independent performance remains substantial, indicating that while progress has been made, further research is needed to fully solve the cross-subject generalization problem.

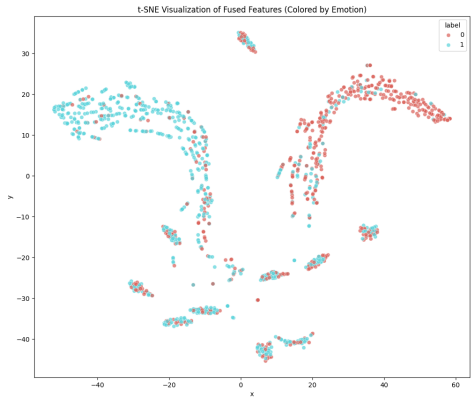
4.3 Discussion of Findings

Our experiments reveal several important insights about EEG-based emotion recognition:

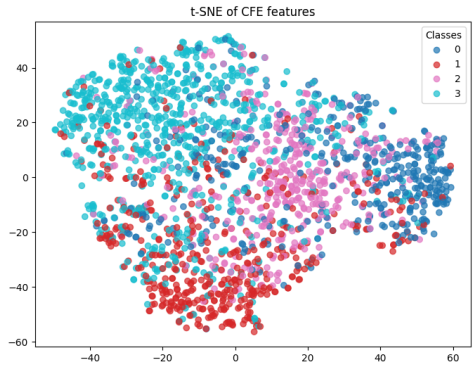
Subject-Invariant Representation Learning: The significant gap between subject-dependent and subject-independent performance across all approaches indicates that complete subject invariance has not been achieved. However, our contrastive learning approaches substantially reduced this gap compared to conventional methods. The combination of Prototype Contrastive Loss and MMD Loss proved particularly effective for aligning features across subjects while maintaining emotional discriminability.

Multimodal Integration: The fusion of EEG and facial features demonstrated benefits, particularly in capturing complementary aspects of emotional responses. Cross-modal contrastive learning effectively aligned the representations from these different modalities, creating a more robust emotion recognition system. However, the improvement was moderate, suggesting that simple fusion approaches may not fully exploit the complementary information in multimodal data.

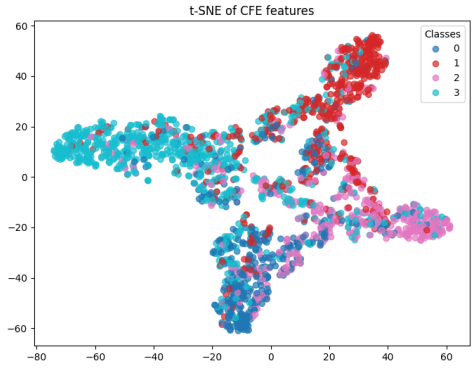
Advanced Loss Functions: Specialized loss functions consistently outperformed standard approaches across all experiments. Dynamic Weighted Focal Loss effectively addressed class imbalance, while contrastive learning methods improved feature separability. The introduction of prototype embeddings helped stabilize training and reduce the impact of noisy labels, particularly in the SEED dataset.



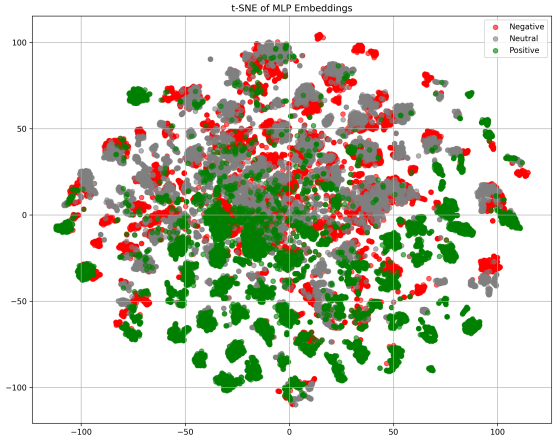
(a) EEG+Face Fusion (Valence)



(b) Before Subject-Specific Mapper (DEAP)



(c) After Subject-Specific Mapper (DEAP)



(d) After Training (SEED)

Figure 6: t-SNE visualizations from different stages and modalities. All subplots show progressive improvement in class separation.

Table 6: Ablation Study of Loss Functions for SEED EEG-only

Loss Configuration	Subject-Independent Accuracy
CE Loss only	75.2%
CE + SupCon Loss	79.8%
CE + L_con2 Loss	82.3%
CE + MMD Loss	80.9%
CE + L_con2 + MMD Loss	86.48%

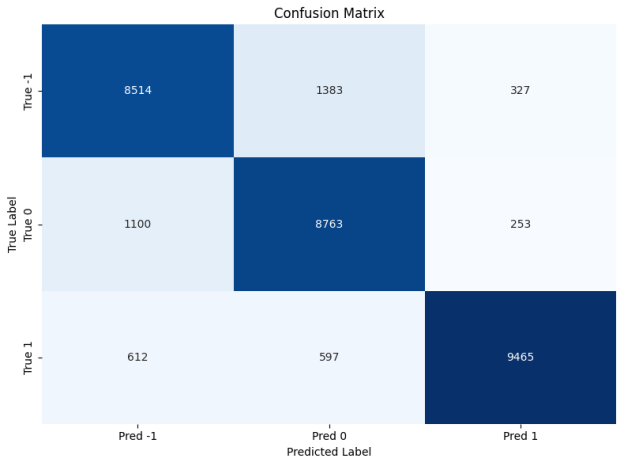


Figure 7: Confusion matrix visualization for SEED EEG-only approach, showing the distribution of predictions across the three emotion classes (negative, neutral, and positive).

Feature Engineering: The addition of asymmetry features (DASM and RASM) alongside DE features significantly improved performance in the SEED EEG-only approach. This confirms previous findings that hemispheric asymmetry contains important information for emotion recognition.

Model Architecture Considerations: Attention mechanisms, particularly Squeeze-and-Excitation blocks and frequency attention, improved performance by dynamically focusing on the most informative channels and frequency bands. The use of residual connections and LayerNorm helped stabilize training for deep architectures.

Limitations: Despite our advances, several limitations remain:

- The performance gap between subject-dependent and subject-independent evaluation suggests that complete subject invariance has not been achieved.
- Laboratory-collected datasets like DEAP and SEED may not fully represent real-world emotional responses, potentially limiting generalization to practical applications.
- The computational complexity of some components (particularly the cross-subject alignment module) may pose challenges for real-time applications.

5 Conclusion and Future Work

5.1 Key Contributions

This research makes several significant contributions to the field of EEG-based emotion recognition:

1. **Enhanced Cross-Subject Generalization:** We developed and evaluated multiple approaches that significantly improve emotion recognition performance on unseen subjects, addressing one of the fundamental challenges in the field. The combination of Multi-Source Contrastive Learning with domain

Table 7: Comparison with Baseline Methods (Subject-Independent Accuracy)

Method	DEAP EEG+Face	DEAP EEG-only	SEED EEG-only
SVM + DE features	49.1%	41.7%	65.2%
CNN-based model	52.3%	53.8%	71.8%
CLISA (FYP-1)	N/A	N/A%	85.4%
Our approach	68.3%	68.7%	86.4%

adaptation techniques effectively reduces inter-subject variability while maintaining emotion-specific information.

2. **Advanced Loss Functions:** We introduced and adapted specialized loss functions for EEG-based emotion recognition, including Dynamic Weighted Focal Loss, Prototype Contrastive Loss, and MMD Loss. These loss functions address specific challenges such as class imbalance, noisy labels, and domain shift.

3. **Effective Multimodal Fusion:** We demonstrated a novel approach for fusing EEG and facial information using cross-modal contrastive learning, creating aligned representations that capture complementary aspects of emotional responses.

4. **Architecture Innovations:** We developed specialized model components such as the subject-specific mapper and cross-subject alignment module that explicitly address the unique challenges of EEG data. These components improve both feature quality and cross-subject generalization.

5. **Comprehensive Evaluation:** We conducted extensive experiments across three distinct approaches and two benchmark datasets, providing a thorough analysis of different methods for EEG-based emotion recognition. This comprehensive evaluation offers valuable insights for researchers in the field.

Our work builds upon the foundation established in FYP-1 with the CLISA framework, extending it with explicit domain adaptation, noise-robust learning, and multimodal integration to create more effective and practical emotion recognition systems.

5.2 Directions for Future Research

Building on our findings, we identify several promising directions for future research:

1. **Real-Time Implementation:** Adapting our approaches for real-time applications by optimizing model architectures and processing pipelines. This could involve model compression techniques and efficient feature extraction methods suitable for online processing.

2. **Personalization Strategies:** Developing methods that combine subject-invariant representations with rapid personalization for new users. This could involve meta-learning approaches that quickly adapt to new subjects with minimal calibration data.

3. **Advanced Multimodal Fusion:** Investigating more sophisticated fusion strategies that better capture the complex relationships between different modalities. This could include attention-based fusion mechanisms that dynamically weight modalities based on their reliability and relevance.

4. **Interpretability:** Enhancing the interpretability of learned representations to provide insights into the neural correlates of emotions. This could involve techniques such as attention visualization, feature attribution methods, and neuroscience-informed architectures.

5. **Cross-Dataset Generalization:** Extending our approaches to improve generalization across different datasets, experimental protocols, and recording setups. This would address practical challenges in deploying EEG-based emotion recognition systems in diverse real-world environments.

In conclusion, while significant progress has been made in addressing the challenges of EEG-based emotion recognition, particularly cross-subject generalization, many opportunities remain for further innovation and improvement. The continued integration of advanced deep learning techniques with domain-specific knowledge from neuroscience and affective computing holds great promise for developing practical and effective emotion recognition systems.

6 References

References

[1] Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161-1178.

[2] Ekman, P. (1992). An argument for basic emotions. Cognition & Emotion, 6(3-4), 169-200.

[3] Shen, X., Wang, Z., Hu, X., Zhang, D., & Zhou, M. (2022). Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition. IEEE Transactions on Affective Computing, 14(3), 2496-2511.

[4] Duan, R.-N., Zhu, J.-Y., & Lu, B.-L. (2013). Differential entropy feature for EEG-based emotion classification. 6th International IEEE/EMBS Conference on Neural Engineering, 81-84.

[5] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. International Conference on Machine Learning, 1597-1607.

[6] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision, 2980-2988.

[7] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. Journal of Machine Learning Research, 13, 723-773.

[8] Deng, X., Liu, C., Li, J., & Zhang, Y. (2024). A novel multi-source contrastive learning approach for robust cross-subject emotion recognition in EEG data. Biomedical Signal Processing and Control, 97, 106716.

[9] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7132-7141.

[10] Baltrusaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443.

[11] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(1), 2096-2030.

[12] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, 33, 18661–18673.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

[13] Snell, J., Swersky, K., & Zemel, R. S. (2017). Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 30, 4077–4087.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13, 723–773.

[14] Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., ... & Patras, I. (2011). DEAP: A database for emotion analysis using physiological signals. IEEE Transactions on Affective Computing, 3(1), 18-31.

[15] Zheng, W. L., & Lu, B. L. (2015). 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.

[16] Davidson, R. J. (2004). What does the prefrontal cortex "do" in affect: perspectives on frontal EEG asymmetry research. Biological Psychology, 67(1-2), 219-234.

[17] Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. Human Brain Mapping, 38(11), 5391-5420.

[18] Zhang, Y., Ji, X., & Zhang, S. (2018). Spatial-Temporal Recurrent Neural Networks for Emotion Recognition. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 8058-8062.

[19] Song, T., Zheng, W., Song, P., & Cui, Z. (2018). EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks. IEEE Transactions on Affective Computing, 11(3), 532-541.

[20] Courty, N., Flamary, R., Tuia, D., & Rakotomamonjy, A. (2017). Optimal Transport for Domain Adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(9), 1853-1865.

[21] Heck, B., Bakhtin, A., & Bharadhwaj, H. (2022). Cross-modal contrastive learning for multimodal representation. Transactions of Machine Learning Research.

[22] Zhao, H., Zhang, S., Wu, G., Moura, J. M., Costeira, J. P., & Gordon, G. J. (2020). Multiple source domain adaptation with adversarial learning. Proceedings of the International Conference on Learning Representations (ICLR).

[23] Mohsenvand, M. N., Kesler, M. L., & Lieberman, M. D. (2020). Contrastive representation learning for electroencephalogram classification. In Proceedings of the Machine Learning for Health NeurIPS Workshop, 238-253.

[24] Picard, R. W. (1997). Affective computing. MIT Press.

[25] Zhang, D., Yao, L., Zhang, X., Wang, S., Chen, W., & Boots, R. (2018). EEG-based intention recognition from spatio-temporal representations via cascade and parallel convolutional recurrent neural networks. arXiv preprint arXiv:1708.06578.

[26] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

A Detailed Model Architectures

A.1 Approach 1: DEAP EEG+Face Model

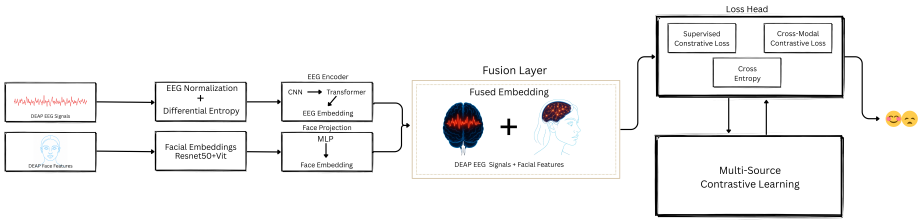


Figure 8: Detailed architecture diagram of the DEAP EEG+Face model, showing the EEG encoder, face encoder, fusion classifier, and contrastive projection heads with layer dimensions.

A.2 Approach 2: DEAP EEG-Only Model

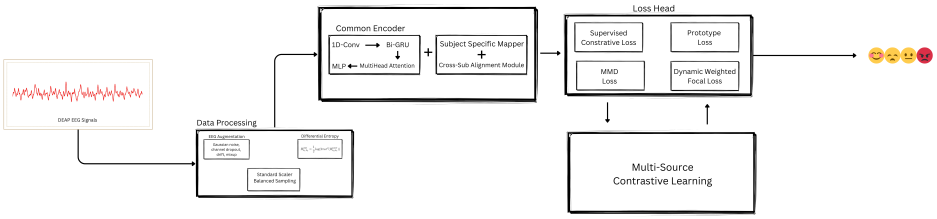


Figure 9: Detailed architecture diagram of the DEAP EEG-Only model, showing the common feature extractor, subject-specific mapper, cross-subject alignment module, and classifier components with layer dimensions.

A.3 Approach 3: SEED EEG with Advanced Losses

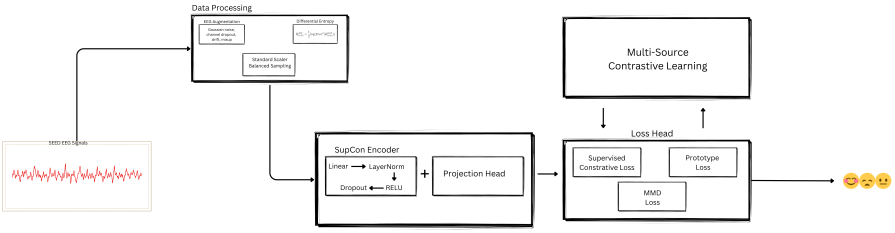


Figure 10: Detailed architecture diagram of the SEED EEG model with advanced losses, showing the encoder network, projection head, and classifier components with layer dimensions.

Table 8: Subject-Wise Accuracy for Approach 1 (DEAP EEG+Face)

Subject ID	Accuracy	Subject ID	Accuracy	Subject ID	Accuracy
1	58.3%	12	55.2%	23	52.1%
2	62.4%	13	50.8%	24	53.7%
3	57.1%	14	59.4%	25	56.3%
4	54.2%	15	51.3%	26	54.8%
5	52.8%	16	55.7%	27	55.2%
6	53.4%	17	57.2%	28	52.9%
7	55.3%	18	53.1%	29	57.8%
8	56.8%	19	51.9%	30	53.5%
9	51.5%	20	54.4%	31	55.9%
10	54.2%	21	56.2%	32	52.7%
11	53.9%	22	55.5%	**Average**	**54.55%**

B Additional Experimental Results

C Mathematical Derivations of Loss Functions

C.1 Dynamic Weighted Focal Loss

The Dynamic Weighted Focal Loss extends the standard focal loss by integrating both label smoothing and dynamic class weighting:

$$A_c^{(t)} = \mu A_c^{(t-1)} + (1-\mu) \cdot \mathrm{Acc}_c^{(t)} \tag{32}$$

where p_t is the predicted probability for the true class, γ is the focusing parameter, and α_t is the class weight.

The class weight α_t is dynamically updated based on per-class accuracy:

$$\omega_c^{(t)} = \frac{1/(A_c^{(t)} + \epsilon)}{\sum_{k=1}^{K} 1/(A_k^{(t)} + \epsilon)} \cdot K \tag{33}$$

where Acc_t is the accuracy for class t and C is the total number of classes. This assigns higher weights to classes that are more difficult to classify.

C.2 Prototype Contrastive Loss

The Prototype Contrastive Loss is defined as:

$$\mathbf{p}_c^{(t)} = \alpha \mathbf{p}_c^{(t-1)} + (1-\alpha) \cdot \frac{1}{|\mathcal{B}_c|} \sum_{i \in \mathcal{B}_c} \mathbf{z}_i \tag{34}$$

where:

- α is the momentum coefficient (0.9),
- $\mathcal{B}_c$ is the set of indices in the current batch with label c.

The normalized prototypes are used to compute cosine similarities between samples and all class prototypes. The logit matrix $\mathbf{L} \in \mathbb{R}^{N \times K}$ is given by:

$$\mathbf{L}_{ij} = \frac{\mathbf{z}_i^\top \mathbf{p}_j}{\tau} \tag{35}$$

where τ is a temperature scaling factor. The prototype contrastive loss is computed using cross-entropy over the softmax-normalized similarities:

$$\mathcal{L}_{\text{proto}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathbf{z}_i^\top \mathbf{p}_{y_i}/\tau)}{\sum_{j=1}^{K}\exp(\mathbf{z}_i^\top \mathbf{p}_j/\tau)} \tag{36}$$

The prototypes are updated using a momentum-based moving average:

$$p_c = m \cdot p_c + (1-m)\cdot\frac{1}{N_c}\sum_{i:y_i=c} z_i \tag{37}$$

where m is the momentum parameter (typically 0.9), and N_c is the number of samples in class c in the current batch.

C.3 Maximum Mean Discrepancy (MMD) Loss

The MMD loss measures the distance between distributions in a reproducing kernel Hilbert space:

$$\mathcal{L}_{\text{MMD}} = \|\frac{1}{n_s}\sum_{i=1}^{n_s}\phi(x_i^s) - \frac{1}{n_t}\sum_{j=1}^{n_t}\phi(x_j^t)\|_{\mathcal{H}}^2 \tag{38}$$

Using the kernel trick, this can be rewritten as:

$$\begin{aligned}\mathcal{L}_{\text{MMD}} = \frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{i'=1}^{n_s}k(x_i^s, x_{i'}^s) + \frac{1}{n_t^2}\sum_{j=1}^{n_t}\sum_{j'=1}^{n_t}k(x_j^t, x_{j'}^t) \\ -\frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}k(x_i^s, x_j^t)\end{aligned} \tag{39}$$

where $k(x, y)$ is a kernel function, typically the Gaussian kernel:

$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{40}$$

For multi-kernel MMD, multiple bandwidth values σ are used and the results are averaged.