Deep Learning Architectures for Enhanced Emotion Recognition from EEG and Facial Expressions

by

Sareh Soleimani

Thesis submitted to the university of Ottawa

in partial fulfillment of the requirements

for the Doctorate of Philosophy in Electrical and Computer Engineering

School of Electrical Engineering and Computer Science

Faculty of Engineering

University of Ottawa

uOttawa

L'Université canadienne
Canada's university

Abstract

Human emotion plays a central role in human experiences that are associated with decision-making, interactions, and cognitive processes. Therefore, emotion recognition has become an important area of research in the field of affective computing for human-computer interactions (HCI). Particularly, there is a growing need for automatic human emotion recognition systems for different applications including robotics, games, surveillance, and healthcare.

In this thesis, we aim to improve automated human emotion recognition methods using machine learning and deep learning approaches. We contribute three solutions for human emotional state recognition. First, we propose a hybrid emotion prediction model that extracts frequency and time domain information from Electroencephalogram (EEG) signals. The model is a cascade of deep learning networks consisting of a pre-trained convolutional neural network (CNN) and residual blocks of recurrent networks. The former extracts spatial features from the signal while the latter learns the temporal dynamics of multi-channel EEG signals and introduces shortcuts across neural layers to enhance the deep network's training efficiency. The proposed model is compared with existing state-of-the-art methods and achieves 0.61 and 0.63 accuracy on the validation and 0.65 and 0.68 accuracy on the test dataset for valence and arousal emotional dimensions, respectively for DEAP dataset.

Second, we propose a novel framework we call the Contrastive Learning GAN-based Graph Neural Network to recognize emotion from EEG signals. The framework integrates self-supervised learning with supervised learning to capture high-quality EEG representations and overcome inter-subject and intra-subject emotion variabilities. We compare the proposed model with recent state-of-the-art emotion recognition models on the DEAP and MAHNOB datasets. The results show that the proposed model achieves a higher recognition performance over previous models with 0.64 and 0.66 as emotion classification accuracies on the test set of the DEAP dataset, 0.66 and 0.71 emotion classification accuracies on the test set of the MAHNOB-HCI dataset for the valence and arousal emotional dimensions, respectively. The proposed model also achieves 0.74, 0.74 and 0.74, 0.78 emotion classification accuracies on the training set of the DEAP and MAHNOB-HCI datasets for valence and arousal emotional dimensions, respectively. We conduct an in-depth examination of how each component of the proposed model contributes to enhancing the emotion recognition accuracy.

Third, we propose a novel Transformer-based bimodal model using EEG and facial expression to perform emotion recognition. We deploy transformer encoders to integrate information across different frequency and data channel regions. We evaluate the proposed model using the DEAP and MAHNOB-HCI datasets. Our experimental results demonstrate that the proposed model surpasses existing techniques, achieving 0.66, 0.72 and 0.65, 0.66 in addition to 0.69, 0.73 and 0.61, 0.68 bimodal accuracies on the DEAP and MAHNOB-HCI training and testing datasets for the valence and arousal emotional dimensions, respectively.

Acknowledgements

First and foremost, I would like to convey my deepest appreciation to my supervisor, Dr. Hussein Al Osman, whose unwavering support, insightful feedback, and expert guidance have been invaluable throughout this journey. Dr. Al Osman is not only a dedicated supervisor but also a kind-hearted individual who consistently stood by my side during the most challenging periods of my life. I believe I am truly fortunate to have had Dr. Al Osman as my PhD supervisor. Beside my supervisor, I would like to thank the esteemed members of my thesis committee, Dr. Mohammad Forouzanfar, Dr. Shervin Shirmohammadi, Dr. Shichao Liu and Dr. Abdulmotaleb El Saddik for their time, dedication, and constructive criticism, which have significantly contributed to the refinement and quality of this work.

I must also express my gratitude to the financial supporters of this endeavor, including the NSERC CREATE BEST program and the University of Ottawa for granting me admission scholarships. Their support and funding have been crucial in the completion of this work.

I extend my deepest appreciation to my parents, who have afforded me opportunities and experiences that have shaped the person I am today. I also dedicate this Ph.D. thesis to these two extraordinary individuals who have been my constant pillars of strength and inspiration – my beloved mom and dad. I also express my gratitude to my brother and sisters for providing tremendous support with their presence in town, bringing smiles to my face during moments of challenge throughout my journey.

Lastly, I wish to express my gratitude to dear god, who has provided me with the inner strength and resilience to accomplish this work.

Table of Contents

List of Figures

List of Tables

List of definitions

QoE	Quality of Experience
HIF	Human Influential Factor
HCI	Human Computer Interface
EEG	Electroencephalogram
CNN	Convolutional Neural Network
CL	Contrastive Learning
GAN	Generative Adversarial Network
GNN	Graph Neural Network
BCI	Brain Computer Interface
fMRI	functional Magnetic Resonance Imaging
MEG	Magnetoencephalography
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
ECG	Electrocardiogram
EOG	Electrooculography
ML	Machine Learning
DL	Deep Learning
EMD	Empirical Mode Decomposition
PSD	Power Spectral Density
LBP	Local Binary Pattern
SVM	Super Vector Machine

LDA	Linear Decrement Analysis
RF	Random Forest
DNN	Deep Neural Network
SGD	Stochastic Gradient Decent
DA	Data Augmentation
AE	Autoencoder
AR	Auto Regression
FFT	Fourier Transform
KNN	K-Nearest Neighbor
TCN	Temporal Spatial Convolutional Layer
DE	Differential Entropy
BiLSTM	Bidirectional LSTM
DASM	Differential Asymmetry
RASM	Rational Asymmetry
DCAU	Differential Causality
GCNN	Graph Neural Network
NLP	Natural Language Processing
CWGAN	Conditional Wasserstein GAN
CBEGAN	Conditional Boundary Equilibrium GAN
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
VA	Valence-Arousal

ReLU	Rectified Linear Unit
CE	Cross Entropy
MLP	Multilayer Perceptron
BSS	Blind Source Separation
FD	Frontal Dimension metric
HVHA	High Valence/High Arousal
HVLA	High Valence/Low Arousal
LVHA	Low Valence/High Arousal
LVLA	Low Valence/Low Arousal
GPU	Graphics Processing Unit

Chapter 1. Introduction

1.1 Motivation

The recent improvements in machine learning and deep learning technologies have enabled solutions to challenges in diverse fields such as image classification, speech recognition, text-to-speech synthesis, and other learning-centric domains [1, 2]. These improvements have also motivated researchers to investigate novel technologies that may improve the human condition or shed light on human behavior. One such technological area is affective computing, where human emotions are recognized automatically by computers. Human emotion plays a central role in human experiences associated with decision-making, interactions, and cognitive processes [3]. Automated emotion recognition technologies can be incorporated into diverse applications in the medical, educational, and entertainment fields [4].

In this thesis, we focus on Electroencephalography- (EEG) based emotion recognition. EEG corresponds to the electrical activity of the brain and is captured via electrodes placed on the scalp. EEG-based emotion recognition is particularity suitable for the Quality of Experience (QoE) assessment of multimedia applications in controlled laboratory settings, where data collection and measurement can be standardized. Such assessment allows providers to offer end-users content that aligns with their quality expectations [5]. Based on the definition in [6], QoE is the output of the quality judgment process which involves the identification of emotional and perceptual quality features. This process occurs inside the brain of the individuals consuming the media so, it is not directly observable [7]. Thus,

brain activity signals such as EEG can provide significant information about the quality judgment process. In contrast, other widely researched emotion recognition strategies that leverage facial expressions, body gestures, and vocal expressions can be susceptible to external influences or intentional masking by the user. For instance, an individual might control facial expressions or vocal intonations to conceal true feelings, but the brain's electrical activity, as captured by EEG, offers a more direct and less filtered insight into their emotional state.

EEG is the most used Brain-Computer Interface (BCI) [8] that provides a direct measurement of the cerebral cortex of the brain. EEG is advantageous for its relatively low-cost and high temporal resolution compared to other neuroimaging technologies, such as functional Magnetic Resonance Imaging (fMRI) and Magnetoencephalography (MEG) [9]. In recent years, due to the rapid development of dry electrode technology, which decreases the invasiveness of this BCI, EEG has become even more suitable for emotion recognition [10][11]. Therefore, EEG-based emotion recognition models have received substantial attention lately [12-15].

With the advancement of data-driven deep learning and machine learning models, automatic emotion recognition can be performed by processing various modalities of data typically collected through sensors. These models learn from subjective ground truth which typically includes surveys, questionnaires with rating scales, or interviews to gauge the users' emotional state to perform emotion predictions based on the data they receive. The evaluation of the model is also performed using subjective ground truth.

While there has been considerable research on deep affective computing models and algorithms, challenges persist in this field. In this thesis, we introduce three novel solutions to address several of these issues. The next section delves into these challenges in detail.

1.2 Challenges and research problem

1.2.1 Model Generalizability

In most of the learning methods presented in the literature for emotion estimation, the reported results pertain to n-fold cross-validation, without verification on a testing dataset [16-19]. Assessing the performance on testing datasets is necessary to ensure that the models are not overfitting and to ensure generalizability. Moreover, they do not clarify how they split the data during each cross-validation iteration. For instance, random partitioning may leave parts of the data for some of the subjects in the training and validation portions. Without further verification using a testing dataset post-cross-validation, we cannot ascertain whether these models were overfitting. Therefore, throughout our work, to ensure the reliability of our proposed emotion recognition models, we implement several benchmark state-of-the-art emotion recognition models and compare their performance on a testing dataset to assess how they generalize.

1.2.2 Feature Extraction

EEG and peripheral physical and physiological signal modalities are stored as time-domain series, but we can either analyze them through their time-dependent or frequency-dependent components, or a mix of both. Therefore, researchers have constructed various models based on these components using machine learning and deep learning techniques. However, there is still debate on what set of input features would effectively increase the

performance of EEG-based emotion recognition models. Therefore, we will compare the performance of different sets of input features based on time-domain and frequency-domain characteristics to evaluate our first proposed solution.

Researchers have also found that there is a long-term dependency in EEG data for emotion recognition [20], a Recurrent Neural Network (RNN) [21] was used to capture temporal dependencies in sequential data. There are two widely used types of RNNs, namely Long-Term Short Memory (LSTM) and Gated Recurrent Unit (GRU). A conventional LSTM can model long sequences while circumventing the vanishing and exploding gradient problems that may plague RNNs by leveraging its gating mechanism. A residual LSTM provides additional spatial skip connections from lower layers. It is hypothesized that such spatial shortcut paths improve the network's training [22]. Moreover, using pre-trained CNNs has been proven to be an effective deep feature extraction technique for addressing overfitting issues when training a deep network on a small dataset [23]. Our first solution proposes a hybrid model comprised of a pre-trained CNN and block-based residual LSTM-RNN to reduce overfitting issues and take both temporal and spatial relations of EEG channels into consideration for emotion recognition. We also perform a brief comparison among frequently used pre-trained CNNs to assess which one performs best for the proposed solution. The proposed solution is also compared with the same model comprised of conventional LSTM layers without the use of residual blocks to demonstrate the improvement of the proposed model.

1.2.3 Inter-Subject and Intra-Subject Emotion Variability

Most current EEG-based emotion recognition models do not address inter-subject and most importantly, intra-subject variability which has posed great challenges for emotion

recognition [24-26]. EEG signals exhibit significant inter-subject variability in response to the same stimulus, leading to reduced robustness of trained classifiers for emotion recognition across different individuals. Furthermore, the problem of intra-subject variability further exacerbates the lack of robustness and generalizability many EEG-based emotion classifiers exhibit. Therefore, we leverage contrastive learning in our second solution to reduce the effects of such variabilities on the performance of the models.

1.2.4 Topological Structure of EEGs

In most of the existing work, the topological structure of EEG channels is not effectively considered, which may limit the model's ability to learn discriminative EEG representations. The topological structure of EEG channels is complex. Therefore, CNNs which are mostly used in the literature for EEG-based emotion recognition, are unable to model this complexity. Hence, to model the topological structure of EEGs for EEG-based emotion recognition, we introduce a Graph Neural Network for our second solution.

1.2.5 Dataset Size Limitation

Most existing models are trained on limited datasets, given the difficulty and cost associated with data collection. Conversely, the lack of data makes it hard to perform emotion recognition, especially with deep learning models which require a relatively large amount of training data. Hence, we incorporate data augmentation into our second solution to improve the robustness of our model.

1.2.6 Parallelism and Long-Range Dependencies

Since the introduction of the Transformer architecture [27], initially designed for NLP (Natural Language Processing) tasks, remarkable advancements have been achieved in

diverse domains, including computer vision and reinforcement learning. The Vision Transformer (ViT) [28] emerged as an innovative adaptation of the Transformer architecture presented in [27], specifically designed for computer vision tasks, notably image classification.

Transformers have been successfully used for emotion recognition due to their ability to capture contextual relationships and dependencies within sequential data. Indeed, Transformers are advantageous as they can process input data in parallel rather than sequentially, making them highly efficient for training and inference. This parallelism is a significant advantage over sequential models like recurrent neural networks (RNNs). Moreover, Transformers are designed to capture long-range dependencies in data, which is crucial for tasks involving contextual understanding and generation of sequences, such as language translation and text summarization. Therefore, for our third solution, we propose a novel Transformer-based deep architecture for emotion recognition.

In the first two solutions provided in this thesis, we only made use of EEG signals as an important modality for affect recognition, especially for QoE assessment applications [7]. However, emotion recognition models can process multiple modalities using different fusion techniques. Multimodal affect recognition is beneficial as it gives us the flexibility of emotion classification even when signals from a single modality are not possible to retrieve. The most common multimodal approach is bimodal emotion recognition incorporating EEG and facial expressions as the input modalities [29]. Therefore, in the third solution, we propose a bimodal deep model using EEG and facial data to perform emotion recognition.

1.3 Research Methodology

The main goal of this thesis is to present methods for processing modalities of information through deep learning models to perform emotion classification. Typically, an emotion recognition architecture is comprised of the following components as depicted in Figure 1.1:

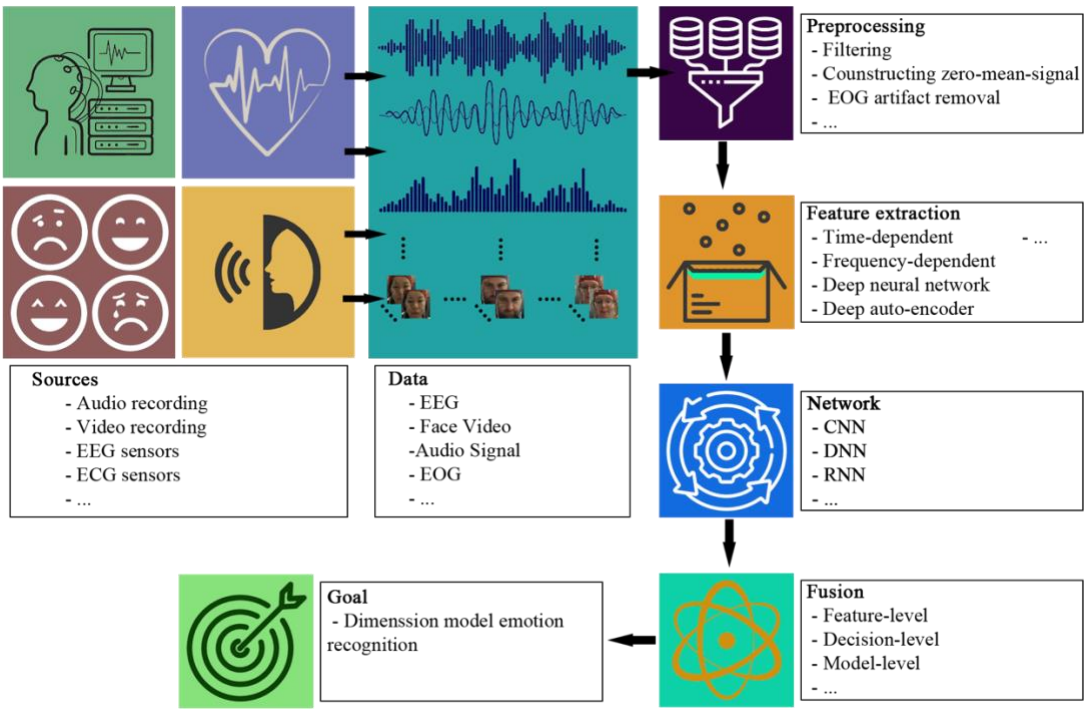


Figure 1.1. General architecture of an emotion recognition model.

1. Collection of multimodal neurophysiological and peripheral physical and physiological data from different sources (e.g., camera recording, EEG sensors, ECG sensors, etc.) followed by labeling signals based on emotional dimensions or emotional keywords. In our work, we solely use publicly available datasets.

2. Then, raw data undergo further pre-processing steps to remove unwanted components of the data that are being measured. in this thesis, we have only used EEG and facial data. Where, EEG data requires preprocessing steps including filtering to preserve frequency components in our region of interest, averaging EEG signals to the common reference signal to construct a spatial voltage distribution with zero-mean, removing baseline recording and Electrooculography (EOG) artifacts that are caused by eyeball movement. A technique is also employed to eliminate redundant information before inputting video data into the proposed model described in Section 6.1. This is because facial changes in videos are frequently subtle and do not change rapidly.
3. When preprocessed clean data is obtained, the rest of the emotion recognition network is comprised of two major parts, i.e., discriminative feature extraction and emotion classification. Feature extraction is the conversion of pre-processed input data into a set of features. This is a common task in most emotion recognition models as it is expected to obtain more descriptive non-redundant learning which also simplifies the processing of the deep network. There are plenty of feature extraction techniques in the literature. A simple feature extraction technique is capturing time-dependent or frequency-dependent components of input data or a mix of both. Researchers have also deployed deep networks such as CNNs, RNNs, GNNs, or a combination of multiple deep networks to extract high-quality features from raw data. Deep auto-encoders and attention-based auto-encoders have also been deployed by researchers to extract high-quality discriminative features through encoding and decoding input data. Therefore, the pre-processed data

undergoes a feature extraction stage to extract high-quality discriminative non-redundant features to further be used for classification purposes.

4. The next component includes the choice of a network. In a unimodal setting, this network is used as a classifier to perform emotion classification. However, if we deal with multiple modalities, we must select a fusion technique that aims to integrate data from different sources and types into a global space. A typical classifier used for deep learning models is comprised of several fully connected dense layers. Modality fusion can also be achieved typically from one of the three main ways: feature-level, decision-level, and model-level fusion.
5. Lastly, the ultimate goal is predicting emotion which can be done using a categorical or n-dimensional model. However, the categorical model has some disadvantages that will be discussed in Section 2.5. Therefore, for our work, we pursue one of the most widely adopted n-dimensional models which specify two dimensions: valence and arousal, and disregard dominance, liking, and predictability labels from further processing given that they are seldom considered in the literature due to their limited importance for interpreting human emotions.

The overall architecture of an emotion recognition architecture is presented in Figure 1.1 which shows all the components involved in a such model. Proposed models typically encompass these components which complement each other to achieve the main goal.

1.4 Contributions

We summarize the contributions of this thesis as follows:

1. Proposing a hybrid model consisting of a cascade of GoogLeNet hierarchal CNN and a block-based residual LSTM for EEG-based emotion recognition. The method

was compared with previous benchmark models on EEG-based emotion detection. Moreover, it was compared with a cascade of the same CNN with standard LSTM-RNN under an identical setting. GoogLeNet was chosen due to the input dimensions' limitation among other pre-trained CNNs. However, comparison among different hierarchal CNNs is performed in an identical setting with an input feature set that is compatible with the rest of the pre-trained CNNs. The results were published in [30]:

S. S. Gilakjani, H. Al Osman, "Emotion Classification from Electroencephalogram Signals Using a Cascade of Convolutional and Block-Based Residual Recurrent Neural Networks," 2022 IEEE Sensors Applications Symposium (SAS), 2022, pp. 1-6, doi: 10.1109/SAS54819.2022.9881254.

2. Proposing a novel framework for EEG-based emotion recognition using a combination of Generative Adversarial Networks (GAN), Contrastive Learning (CL), and Graph Neural Networks (GNN). The framework leverages self-supervised learning with supervised learning to capture high-quality EEG representations and to overcome inter-subject and intra-subject emotion variability. An ablation study is also performed to systematically investigate the effect of employing GAN data augmentation, Contrastive Learning, and Graph Neural Network on emotion recognition performance by isolating each component and analyzing its effect through comparison with several benchmark competing models. The proposed model was tested on two popular public datasets. This contribution resulted in a journal paper [31]:

S. S. Gilakjani, H. Al Osman, "Contrastive Learning Generative Adversarial Network Based Graph Neural Network for Emotion Recognition," Submitted to IEEE Transaction on Affective Computing, March 2023.

3. Proposing a novel framework for deep bimodal emotion recognition deploying Transformers over different frequency regions of EEG and facial expression data. This model uses a Transformer-based fusion to fuse data from both modalities. We systematically investigate the effectiveness of the proposed model by comparing it with existing benchmark bimodal and unimodal emotion recognition models. We replaced various fusion techniques in our proposed model to better understand the effect of the Transformer-based fusion method on the overall performance. The proposed bimodal Transformer-based emotion recognition model outperforms the existing models by approximately 4% in the valence and arousal emotion dimensions. This contribution resulted in a journal paper [32]:

   S. S. Gilakjani, H. Al Osman, " Transformer-based Bimodal Emotion Recognition Model with Fusion Transformer," Submitted to IEEE Transaction on Affective Computing, October 2023.

1.5 Thesis Outline

The outline for the rest of this thesis is as follows:

Chapter 2– Background offers foundational knowledge on machine learning and deep learning methodologies, as well as the features employed for emotion classification, evaluation metrics, and emotional models used in the literature.

Chapter 3– Related work presents a thorough review of pertinent studies, explaining state-of-the-art studies on data-driven emotion recognition models.

Chapter 4– Proposed EEG-based Emotion Classification Model Based on Hierarchal CNN and Block-Based Residual LSTM presents the proposed EEG-based emotion classification model constructed using a hierarchal CNN and block-based residual LSTM(s). We investigate the efficacy of four input feature sets for our proposed model and evaluate how the model performs when combining the pre-trained CNN with standard LSTMs. Additionally, this chapter provides emotion classification outcomes, predominantly focusing on the performance assessment of our model for the valence and arousal emotional dimensions. We compare our proposed solution to recent state-of-the-art methods and detail our findings.

Chapter 5– A Graph Neural Network for EEG-Based Emotion Recognition with Contrastive Learning and Generative Adversarial Neural Network Data Augmentation presents an EEG-based emotion recognition model comprised of three components. We perform a comprehensive analysis of the effect of each component of the proposed model in improving emotion recognition accuracy. The proposed model is also compared with existing competing emotion recognition models and performance evaluation results are provided.

Chapter 6– Transformer-based Bimodal Emotion Recognition Model with Fusion Transformer presents a Transformer-based bimodal emotion recognition model that uses EEG and facial data as input modalities. We assess the proposed model in both unimodal and bimodal contexts and contrast its performance with current state-of-the-art approaches.

Moreover, we explore various fusion techniques integrated into our model to gauge the effectiveness of our chosen fusion method.

Chapter 7– Conclusion and Further Research Plans provides conclusions to summarize the results and presents our future research plan. We discuss the proposed emotion classification models and then list suggestions for future work.

Chapter 2. Background

In this chapter, we provide background information on the use of machine learning and deep learning models in emotion recognition. We begin by introducing the emergence of machines in performing objective recognition in Section 2.1, followed by presenting deep neural networks and several deep learning architectures in Section 2.2 and Section 2.3, respectively. Moreover, in Section 2.4 we discuss evaluation metrics used on data-driven emotion recognition models. Finally, in Section 2.5 we represent emotional models used in the literature for emotion recognition.

2.1 Emergence of Machine Learning and Deep Learning

In general, machine learning is the application of algorithms to perform data learning and prediction automatically without help from humans through the use of data. Machine learning is a component of artificial intelligence [33]. Machine learning (ML) algorithms create a model to make predictions based on training data without being explicitly programmed to do so. Deep learning (DL) is considered an evolution of machine learning which uses neural networks to perform data prediction without human involvement. Traditional ML algorithms use hand-crafted data features to feed to their ML applications, unlike DL algorithms. This is the key difference between traditional ML and DL algorithms. The hand-crafted features are typically captured by applying several feature extraction techniques such as Empirical Mode Decomposition (EMD), Power Spectral Density (PSD), Local Binary Pattern (LBP), and many more. Then, traditional ML algorithms including Support Vector Machine (SVM), Linear Decrement Analysis (LDA), and Random Forest (RF), among others, are deployed to perform data prediction or

classification. In contrast, in the case of DL, features can be extracted automatically and represented hierarchically in several neural layers. This is the principal advantage of DL compared to traditional ML approaches. The development of deep learning has shown significant improvement in many classification/prediction tasks [34, 35].

The emergence of DL and artificial neural networks originates from our desire to create a computer system that mimics the human brain's neural network [36]. However, to create such a system, an understanding of the human cognitive system is required. Thus, tracing back to the early endeavors to understand the workings of the human brain, the origins of DL can be attributed to psychologist Frank Rosenblatt, who first introduced concepts related to human brain functionality. Rosenblatt collaborated with a team to construct a machine designed to recognize the letters of the alphabet. The machine was called the "perceptron", which became the prototype for modern Deep Neural Networks (DNN) [37]. Since then, various DNN architectures were proposed but until recently, they were very difficult to train. Consequently, they fell behind other recognition/classification methods.

In 2012, the deep learning revolution in machine learning began when Alex Krizhevsky et al. proposed a DNN-based model called AlexNet [38] that won several international competitions. It was the first time it was shown that deep learning models are typically superior to other machine learning models for large datasets where complex processing is required. Therefore, researchers have recently leveraged deep learning techniques to present more reliable outcomes for prediction/classification on larger datasets.

2.2 Deep Neural Networks (DNN)

Deep neural network (DNN) approaches can be categorized as supervised, semi-supervised, or partially supervised and unsupervised learning [39]. Supervised learning is

a learning category that uses labeled data. In supervised learning, algorithms are trained using labeled data to predict outcomes. Supervised learning is typically separated into two types of problems: classification and regression [39].

Unsupervised learning is a learning that uses unlabeled data where algorithms are trained to cluster unlabeled data. In unsupervised learning, the internal relationship among inputs is learned to discover unknown structures within the input data.

As collecting large amounts of labeled data is challenging and expensive, semi-supervised learning is typically used to address the issue of small labeled datasets. Semi-supervised learning is a type of learning that distills knowledge from both labeled and unlabeled data. Typically, in semi-supervised learning, the unsupervised component is used for representation learning which is also called self-supervised learning [40] where high-level features are extracted from low-level ones to obtain more useful information from data. On the other hand, the supervised component performs the recognition task.

In General, simple DNN is a network comprised of several hidden layers in between the input and output layers as depicted in Figure 2.1. Hidden layers consist of non-linear information processing units called nodes/neurons which provide the capability of high-level feature learning and pattern classification [41].

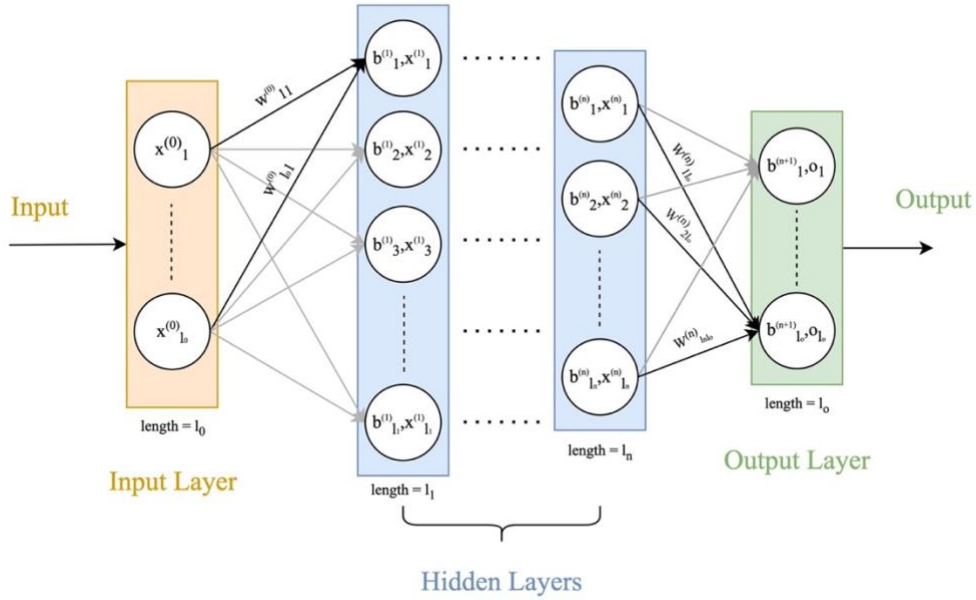


Figure 2.1. A typical Deep Neural Network (DNN) architecture.

As can be seen from Figure 2.1, neurons in DNN are capable of having synapses that provide connections to more than one neuron in the preceding layer. Each synapse has an associated weight that specifies the importance of the preceding neuron in the overall neural network. Once a neuron receives its input from the neurons of the preceding layer, it adds them up corresponding to their associated weights and finally adds a bias to the sum then passes them to an activation function as follows:

$$x_j^{(k)} = \mathbb{a}^{(k)}\left(\sum_{i=1}^{l_{k-1}} w_{ij}^{(k-1)} x_i^{(k-1)} + b_j^{(k-1)}\right) \qquad 1 \le j \le l_k \ , 2 < k \le n \tag{2.1}$$

Where (k) refers to layer number, x_i^{k-1} corresponds to the i-th neuron at layer $k-1, l_{k-1}$ represents the number of neurons in layer $k-1$, $w_{ij}^{(k-1)}$ represent weight from neuron i at layer $k-1$ to neuron j at layer k, $b_j^{(k-1)}$ refers to the bias value of neuron j at layer $k-1$. $\mathbb{a}^{(k)}$ refers to an activation function corresponding to layer k.

Training deep learning models is primarily related to adjusting the model's weights. To do so, a cost function is required to measure the error between predicted outputs and actual outputs. The lower the value of the cost function, the better the model's prediction capability. During training, typically the weights are initially assigned a random value. Then, the cost function is calculated based on the difference between predicted and actual outputs which shows the error. Then, the weights are optimized using an optimization algorithm over several steps such that eventually, the cost function reaches a minimum value. The learning rate is a parameter that controls the pace of optimizing/updating the weights. Optimizing the weights necessitates computing the derivative of the model's cost function. Backpropagation, a method developed for this purpose [42], involves propagating the observed error between predicted and actual outputs backward through the network by adjusting the weights to improve prediction accuracy. A widely employed optimization algorithm to update DNN weights is the Adam Deep learning optimizer which is an extension of Stochastic Gradient Decent (SGD) [43]. It utilizes a decaying learning rate that decreases with an increase in epochs, unlike SGD which uses a single learning rate throughout the optimization process [43].

In deep learning, we deal with two types of parameters: machine-learnable parameters and hyperparameters. Machine-learnable parameters are typically the weights or other parameters that algorithms estimate on their own during the training of a DNN model. Hyperparameters are the ones that are assigned specific values by machine learning engineers to control the way that algorithms learn or tune the performance of the model. The learning rate is a hyperparameter that must be initialized before the learning process. The choice of learning rate is crucial to the optimization process. If it is too large, the

network may diverge instead of converging. However, if it is too small, it will take a longer time for the network to converge. Hyperparameters can be tuned using hyperparameter optimization techniques [44] to reach a higher performance.

2.3 Several DNN Architectures

Since Krizhevsky et al.'s ground-breaking work on DNN in 2012 [38], numerous other DNN architectures have been introduced, demonstrating enhanced performance. We will delve into some of these architectures in the following sections.

2.3.1 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) are a class of deep neural networks (DNN) that were initially developed for image recognition and are now used in various fields including computer vision and natural language processing. The history of CNN can be traced back to 1988 when the first CNN structure was proposed by Fukushima [45] but it could not be used due to computer hardware limitations for training the model until 2012 when Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton from the University of Toronto achieved remarkable accuracy in image recognition using CNNs [38].

CNNs are designed to work with a grid-like topology, such as an image, which can be represented as a matrix of pixels. CNNs are highly advantageous over simple DNNs as they have been highly effective in extracting discriminative features from input data at different levels of abstraction. The CNN network consists of multiple layers, including convolutional layers, pooling layers, and fully connected layers as represented in Figure 2.2:

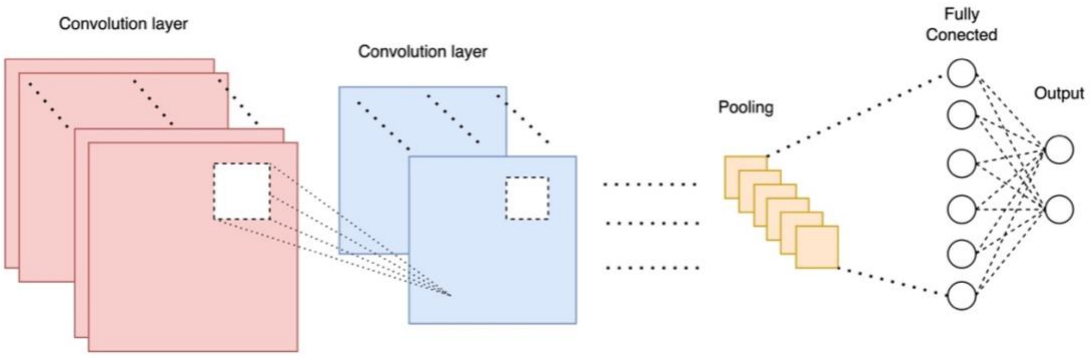


Figure 2.2. A typical CNN architecture

- The convolutional layer serves as the foundational layer of the CNN, performing operations on the input image to extract various features. As the building block of a CNN, this layer applies a set of learnable filters (often referred to as kernels) to the input image, resulting in a feature map.
- The pooling layer is used to reduce the spatial size of the feature map produced by the convolutional layer. In other words, it down-samples the feature map, reducing its spatial size while keeping the most important information. There are two commonly used pooling operations: max pooling and average pooling. Max pooling takes the maximum value of a set of neighboring pixels, while average pooling takes the average value of a set of neighboring pixels.
- The fully connected layers are usually placed at the end of the CNN network after building the desired number of convolutional and pooling layers. In CNN, the fully connected layers are used to produce the final output. The input to the fully connected layer is a flattened feature map, which is a 1D vector, produced by the

previous layers. The fully connected layers use a set of learnable weights and biases to compute the final output of the network.

Pre-trained Convolutional Neural Networks (CNNs) are CNN models that have been trained on large datasets for a specific task and can be fine-tuned for new tasks on smaller datasets using transfer learning. To do so, the early CNN layers are frozen and the last few layers which deal with prediction/classification are trained. Several pre-trained CNN models in the literature will be discussed in Chapter 4.

In our work, we use CNN as a feature extraction technique. The high-level features obtained from CNN will then be forwarded to another network for an emotion recognition task.

2.3.2 Recurrent Neural Network (RNN)

When humans process information, they relate it to other information they have previously processed. For instance, as the reader is reviewing this thesis, they understand new sentences or words based on previous ones. The CNN and simple DNN structures are not ideally structured to process sequential information as they ignore the relationships between data in a sequence. Hence, they present a weakness in terms of modeling sequential data such as text, speech, video, or EEG. Therefore, Recurrent Neural Networks (RNN) were developed in the 1980s to handle this type of problem [46]. RNNs are another class of DNNs that are designed to capture temporal dependencies in sequential data. RNN can be viewed as multiple copies of the same network (depending on the number of sequences) where each network passes information to the next network. RNNs use hidden states to store and pass information from the previous sequence to the next sequence. At each time step/sequence, the current input and the previous hidden state or previous output

(depending on the RNN model) are combined to update the current hidden state. The updated hidden state or output is then used as input to the next time step along with the input of the next sequence, allowing information to be propagated from one sequence to the next (Elman and Jordan RNN versions) [47, 48]. This enables the RNN to model temporal dependencies in sequential data.

Early RNN models faced the vanishing gradient problem, stemming from the repeated multiplication of gradients as they are propagated through many time steps or sequences. This caused the gradients or partial derivatives to become very small which makes it difficult to train the network. To mitigate the vanishing gradient problem, many solutions have been proposed in the past few decades. The Long-Short-Term-Memory (LSTM)-RNN [49] and the Gated Recurrent Unit (GRU)-RNN [50] were introduced and considered as the two possible effective solutions to address the aforementioned problem. GRU is a variant of LSTM that has fewer parameters, consequently, it is trained faster with lower complexity. On the other hand, LSTM has been shown to provide better performance while requiring higher computational power [51].

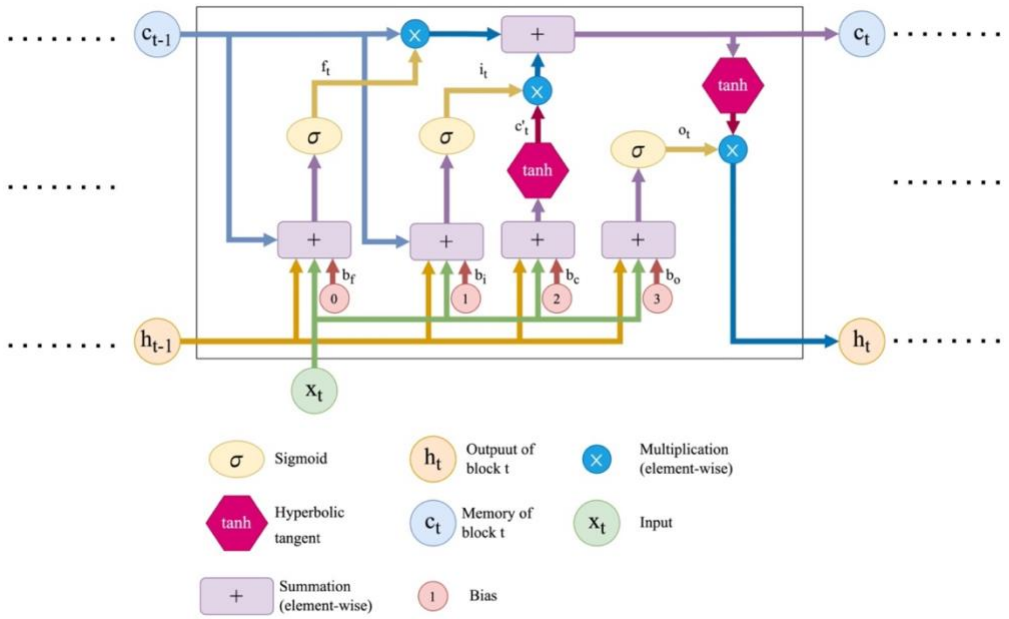


Figure 2.3. Standard LSTM-RNN network

Each LSTM layer consists of multiple parallel LSTM units depending on the number of sequences. Figure 2.3 shows the overall architecture of an LSTM unit. An LSTM unit is composed of three gates- input gate (i_t), forget gate (f_t) and output gate (o_t) which are used to remove or add information to the cell states (c_t) called remember gates which is the horizontal line at the top of Figure 2.3. the values of the gates are calculated using equations provided in (2.2). The three gates and cell state are controlled by the input and previous hidden state, allowing the LSTM layer to learn when to retain information and when to forget it. It is the cell state derivative that prevents the LSTM gradients from vanishing:

$$
\begin{aligned}
f_t &= \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \\
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
\mathrm{c}_\mathrm{t}' &= tanh(W_c x_t + U_c h_{t-1} + b_c) \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \mathrm{c}_\mathrm{t}' \\
h_t &= o_t \odot tanh(c_t)
\end{aligned}
\tag{2.2}
$$

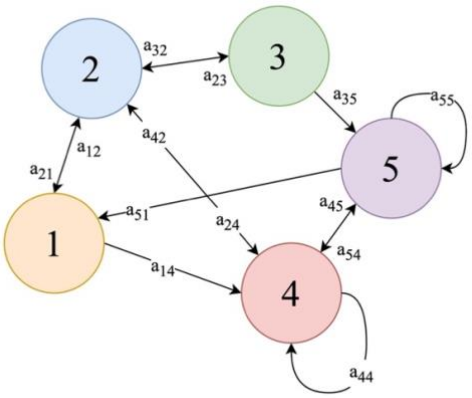
$\odot$ is a Hadamard product (element-wise product), σ is a sigmoid activation function W and U are weights that correspond to the input and previous hidden state, respectively.

We use LSTM-RNN in our work to model temporal dependencies in EEG data to perform emotion recognition.

2.3.3 Graph Neural Network (GNN)

To learn more discriminative features from the input data, the Graph Neural Network (GNN) was introduced in the early 2000s. GNNs are another class of DNNs that are used for graph-structured data. GNNs are also deployed in the field of EEG-based affect recognition to exploit the topological structures of EEG channels to learn more discriminative emotional features.

A graph is defined as $\mathcal{G} = \mathcal{V}, \mathcal{E}, A$ where $\mathcal{V}$ denotes a set of nodes with a number $|\mathcal{V}| = N$, $\mathcal{E}$ represents a set of edges connecting the nodes, and $A \in \mathbb{R}^{N \times N}$ represents an adjacency matrix defining the connection between any two nodes. The element a_{ij} of the adjacency matrix with i and j representing the row and column numbers, shows the weight which corresponds to the importance of the connection between nodes i and j. Data on $\mathcal{V}$ can be represented by $X \in \mathbb{R}^{N \times d}$ where d represents the dimension of input features.



$$A_{5\times5} = \begin{bmatrix} 0 & a_{12} & 0 & a_{14} & 0 \\ a_{21} & 0 & a_{23} & a_{24} & 0 \\ 0 & a_{32} & 0 & 0 & a_{35} \\ 0 & a_{42} & 0 & a_{44} & a_{45} \\ a_{51} & 0 & 0 & a_{54} & a_{55} \end{bmatrix}$$

(a) An example of graph representation (b) Adjacency matrix of the graph with 5 nodes

Figure 2.4. Example of a graph and the associated adjacency matrix

Figure 2.4 demonstrates an example of a graph containing 5 nodes and the edges connecting the nodes of the graph. When dealing with EEG data, the number of nodes refers to the number of EEG channels. The figure also illustrates the adjacency matrix corresponding to the graph. The arrows on the graph representation denote the edges connecting two nodes. Each node represents a channel of input data with its corresponding dimension. For 1D channels like EEG signals, each node contains a vector with a specific length.

In a basic GNN model, similar to (2.1), each neuron is comprised of a weighted sum and a bias. However, in graph representation, we have N nodes where each node is comprised of several hidden neurons. So, neurons on each node are computed as follows:

$$x_{m,j}^{(k)} = \mathbb{a}^{(k)}\left(\sum_{z=1}^{N}\left(\sum_{i=1}^{l_{k-1}} w_{ij}^{(k-1)} a_{mz} x_{zi}^{(k-1)} + b_j^{(k-1)}\right)\right) \quad (2.3)$$

$$, \quad 1 \leq \mathrm{m} \leq N$$

Where m denotes the node number, a_{mz} represents connection weight between node m and node z (i.e., from adjacency matrix $A \in \mathbb{R}^{N\times N}$). $x_{zi}^{(k-1)}$ represents *i*th neuron from node z on layer $k-1$. j refers to the neuron number in each node which depends on the size of a layer. All other parameters are identical to (2.1).

Since EEG data can be represented as a graph, we will use GNN to capture more discriminative emotional features in Chapter 5. In our proposed model, each EEG channel corresponds to a node, the relationship between every two channels corresponds to the edges of the graph, and the elements of the adjacency matrix describe the importance of the channels' relationship. A greater value of an element on the adjacency matrix indicates a closer relationship between the two channels.

2.3.4 Contrastive Learning (CL)

Contrastive Learning (CLs) is another class of DNNs that is used for general data-representation learning. CL captures a deep representation of data by maximizing the similarity between two similar data samples (called a positive pair) and minimizing the similarity between two different data samples (negative pair) using contrastive loss. The idea of maximizing the agreement between representations of two similar data samples was first introduced by Becker and Hinton in 1992 [52]. Then, it was extended by applying data augmentation, network architecture, and contrastive loss to evolve into the Contrastive Learning model. The overall structure of CL is depicted in Figure 2.5.

As can be seen from Figure 2.5, the general contrastive learning architecture contains four sub-components: pair loader, channel encoder, channel projector, and contrastive loss function. The pair loader creates a batch of several positive pairs. The channel encoder extracts representations from samples in pairs. Then, the channel projector maps the

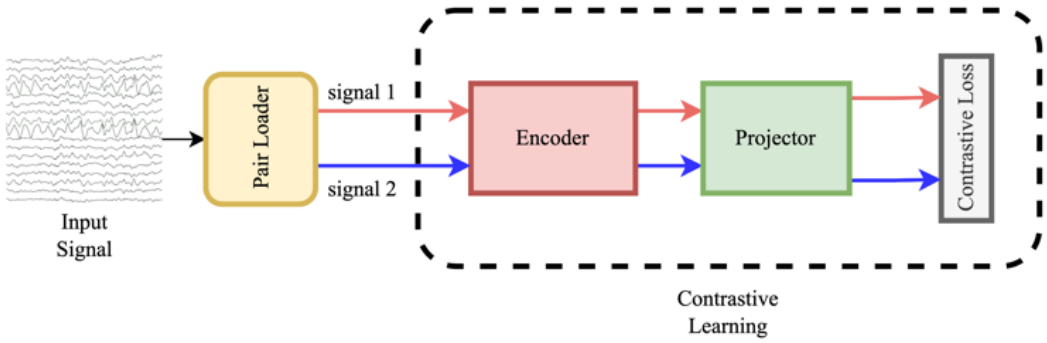


Figure 2.5. General architecture of Contrastive Learning.

extracted representations to another latent space to maximize their similarities using the contrastive loss function. The parameters of the channel encoder and channel projector are optimized such that contrastive loss is minimized.

In general, the most crucial step in CL is the choice of positive pairs which highly impacts the quality of CL's output signal [53]. We deploy CL with a novel pairing mechanism in the second proposed emotion recognition solution described in Chapter 5.

2.3.5 Generative Adversarial Networks (GAN)

Since training data are limited in size, it is difficult to train deep networks with satisfactory accuracy using deep learning architectures due to the high model parameters which require a large amount of data for training. To overcome this problem, Data Augmentation (DA) techniques were introduced to synthetically enlarge the size of datasets [54]. DA refers to the process of generating new data samples by transforming existing samples in a dataset. This technique can increase the accuracy and stability of the recognition tasks. Generative Adversarial Networks (GANs) are a widely used DA technique which is also a class of DNNs. GAN was first introduced by Goodfellow et al. in 2014 [55].

As can be seen in Figure 2.6, in general, GAN is a network comprised of two competing parts: generator and discriminator which are both parameterized as deep neural networks. The generator learns how to generate synthetic data that resembles real data. The discriminator evaluates the probability that a sample originates from the real data distribution. In the training process of a GAN, the generator (G) attempts to deceive the discriminator (D) by generating synthetic data while, the discriminator attempts to improve its discrimination to avoid being deceived by the synthetically generated data. The two parts are optimized simultaneously to eventually reach Nash equilibrium. The adversarial training procedure is formulated as a minimax problem, expressed as:

$$\min_{\theta_g} \max_{\theta_d} L(X, Z) = E_{x_i} \sim X\left[log\left(D(x_i)\right)\right] + E_{z_i} \sim Z[log(1 - D(G(z_i))] \tag{2.4}$$

where θ_g and θ_d denote the parameters of the generator and discriminator, respectively. X represents the real data distribution and Z represents a noise distribution where it can be a uniform or Gaussian noise distribution. The training of GAN is performed in two steps-maximization of the discriminator's loss and minimization of the generator's loss. For the first step, the optimum D is found by maximizing this function with a fixed G and Z, then, in the second step, this function is minimized to find optimal G by using previously computed optimal D.

GAN is also adopted in our proposed emotion recognition model presented in Chapter 5 to enlarge the size of the EEG training dataset. We will expand this minimax problem in Section 5.2.2 to describe the two steps in more detail.

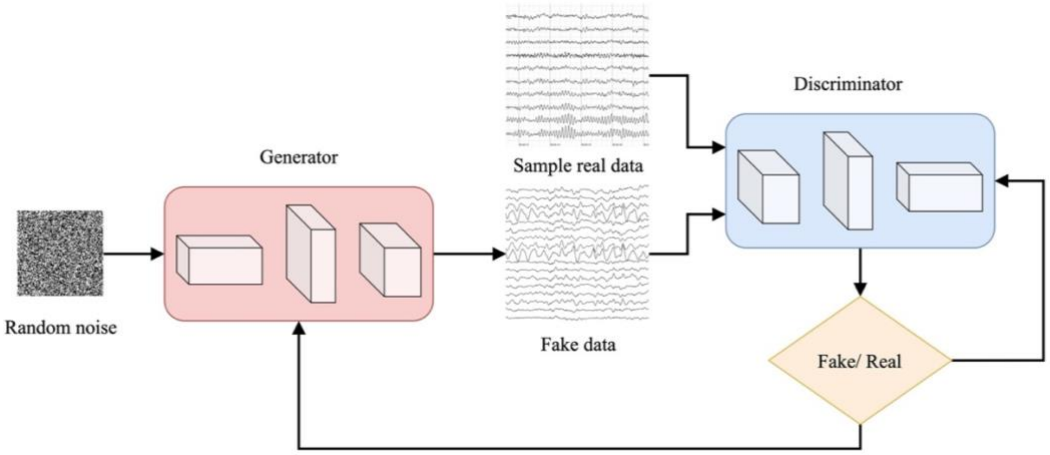


Figure 2.6. A typical Generative Adversarial Network architecture.

2.3.6 Autoencoders (AE)

Autoencoders (AEs) are a class of DNNs that are used for representation learning from input data. As shown in Figure 2.7, a typical AE is composed of two main components: an encoder, which maps/encodes the input data (x) to a latent space typically for data dimensionality reduction, compression, and many more, and a decoder, which attempts to map/decode the encoded representations back to the original input space (reconstructed input). During the training of the AE, the autoencoder is optimized using an objective function that measures the reconstruction error between the input and the output of the decoder which must be minimized.

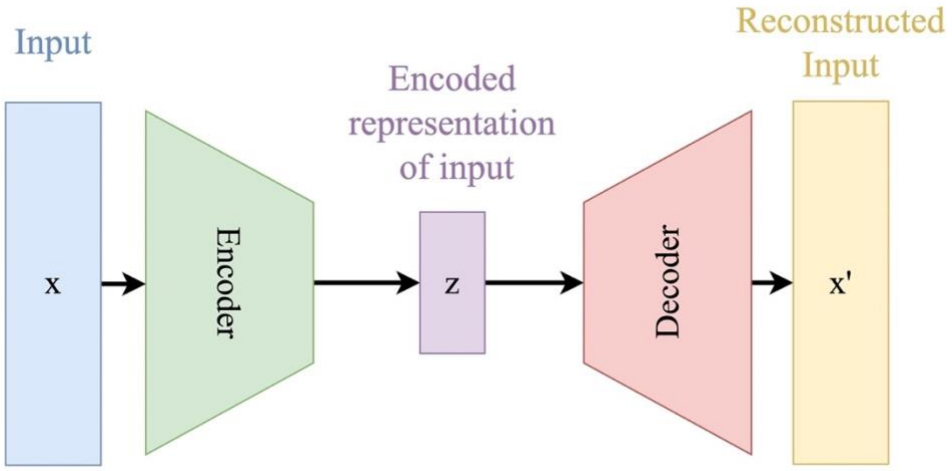


Figure 2.7. The overall architecture of an Autoencoder (AE).

The Transformer is a deep neural network based on the AE architecture that was introduced in a ground-breaking paper by Vaswani et al. [27]. It has since become the foundation for many state-of-the-art natural language processing (NLP) models, including BERT, GPT, and more. The Transformer architecture is known for its ability to handle sequential data efficiently, making it particularly well-suited for tasks like machine translation, text generation, and emotion analysis. Originally, the Transformer architecture was composed of two main components: the Transformer encoder and the Transformer decoder. These components work together to process input sequences and generate output sequences. However, in various applications, only one of these components is employed depending on the specific requirements. For instance, in the emotion recognition field, the Transformer encoder is typically solely used to process and understand the contextual information from input sequences, without the need for a decoder to generate a subsequent sequence. The architecture of the Transformer encoder is described in Section 6.3.

2.4 Evaluation Metrics

There are various evaluation metrics to assess the performance of a deep model. For emotion classification tasks, the following metrics are commonly used.

Confusion matrix

A confusion matrix or an error matrix is used in the field of machine learning and deep learning to visualize the performance of a model using a tabular layout. Each row of the matrix shows actual class occurrences whereas, each column represents predicted class occurrences. The following terms are typically used in a 2D confusion matrix for two-class classification problems with negative (class 0) and positive (class 1) classes:

True Positive (TP): both predicted and actual class values are 1.

True Negative (TN): both predicted and actual class values are 0.

False Positive (FP): the predicted class value is 1 but the actual class value is 0.

False Negative (FN): the predicted class value is 0 but the actual class value is 1.

Accuracy

Accuracy is the ratio of the true predicted values (i.e., the sum of TP and TN) to the total predicted values.

Precision

Precision is he ratio of TP to the total positive predictions (i.e., the sum of TP and FP) as follows:

$$precision = \frac{TP}{TP + FP} \tag{2.5}$$

Recall

Recall is the ratio of TP to the total that should have been predicted as positive (i.e., the sum of TP and FN) as follows:

$$Recall = \frac{TP}{TP + FN} \tag{2.6}$$

F1-Score

F1-score is the harmonic mean of precision and recall which is computed as:

$$\begin{aligned} F1 - score &= 2 \times \frac{precision \times recall}{precision + recall} \\ &= \frac{2TP}{2TP + FP + FN} \end{aligned} \tag{2.7}$$

This metric describes the prediction capability of the assessed model.

2.5 Emotional Model

Emotion can be represented in two ways [56]. The most straightforward way is the categorical model, which uses discrete emotion labels such as fear and happiness. However, the categorical model has some disadvantages. For example, emotions do not have exact translations in different languages, e.g., "disgust" does not have an exact translation in the Polish language [57]. Also, the categorical model may not encompass all emotions, as it is restricted to a limited set of categories. Alternatively, emotions can be represented using an n-dimensional model [58]. One of the most widely adopted dimensional models specifies two dimensions, valence, and arousal [59]. Therefore, we only consider these two emotional dimensions for our work, disregarding the dominance, liking, and predictability labels that are sometimes reported in some datasets. In particular, the liking and predictability labels are application-specific and are seldom

considered in the literature. In the datasets we employed, these dimensions are represented using numbers between 1 and 9. Valence measures the event's pleasantness on a scale from negative (1 - very unpleasant) to positive (9 - very pleasant). Arousal gauges the emotion's intensity, ranging from 1 (very calm) to 9 (very excited). In our datasets, these dimensions are numerically represented between 1 and 9.

Discrete emotions can lie in the four quadrants of the valence-arousal (VA) dimensional model as depicted in Figure 2.8. Subsequently, discrete emotions can be estimated through this model. For example, if valence and arousal scales are both greater than 5, the emotional state falls in the first quadrant, which can correspond to Excited or Happy emotion.

In our work, we classify the valence and arousal emotional dimensions into two classes: high and low with a threshold set to 5 to distinguish between the high or low intensity of each emotional dimension, i.e., ratings from 6 to 9 refer to the high class and ratings from 1 to 5 refer to the low class. However, this approach also comes with a limitation that needs to be considered. This limitation is the inherent loss of information due to binarizing continuous emotional ratings. By dividing valence and arousal scores into high and low classes, slight differences in emotional states are inevitably overlooked. For instance, assigning a valence rating of 4 to the low class and a valence rating of 5 to the high class may fail to capture subtle differences in emotional intensity between these two adjacent ratings. As a result, the model may struggle to distinguish between slight differences of emotional states that fall close to the threshold value.

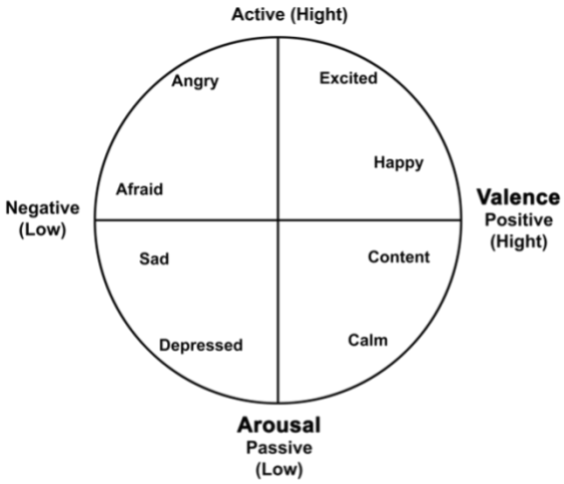


Figure 2.8. Valence-arousal dimensional emotion model.

Chapter 3. Related Work

This chapter discusses state-of-the-art data-driven emotion recognition models. Moreover, we review different manually engineered and non-manually engineered feature extraction techniques used in the context of emotion recognition.

3.1 Previous EEG-based Studies with DNNs

The most widely used modalities for emotion recognition are brain activity signals measured by EEG, facial expression, text, and speech. As the quality judgment process occurs inside the brain and depends on the user's emotional features, EEG signals are expected to carry relatively more significant information about one's emotional state [7]. Therefore, in recent years, the use of EEG in the field of emotion prediction has received remarkable attention from many researchers.

Wang et al. compared the outcomes of using various features derived from EEG data for emotion assessment [60]. They showed that power spectrum features outperform others for emotion classification. Lokannavar et al. employed an SVM to classify discrete emotions such as fear, sadness, and happiness [61]. They used Auto Regression (AR) and Fast Fourier Transform (FFT) input features extracted from EEG signals.

With the success of AlexNet in 2012, it was shown that deep learning functions are typically superior to other machine learning methods for large datasets [38]. Therefore, researchers started to adopt deep learning techniques to present more reliable outcomes for larger EEG datasets. He et al. used a hierarchical Convolutional Neural Network (CNN) to classify positive, negative, and neutral emotions using EEG data [62]. They proved that their method outperformed K-Nearest Neighbors (KNN), SVM, and stacked auto-encoder

algorithms. Tripathi et al. compared CNN with four fully connected Deep Neural Networks (DNN) [63]. They showed CNN improves emotion classification accuracy on the DEAP [64] dataset. Islam et al. [65] used CNN to recognize emotion from Pearson Correlation Coefficient (PCC) featured images obtained from EEG signals. They compared their results with earlier emotion recognition methods such as Deep Forest, dynamical graph CNN, SVM, and Random Forest. They indicated that their method outperforms existing ones by achieving lower computational complexity and requiring less memory to execute in addition to attaining improved emotion recognition accuracy. Ding et al. proposed the TSception model which consists of temporal-spatial convolutional layers (TCN) for EEG-based emotion recognition [66]. They collected EEG data from 18 healthy participants in an immersive virtual reality environment to evaluate their proposed model. They compared their model with SVM, EEGNet, and LSTM in terms of emotion classification accuracy and concluded that TSception achieves a relatively higher classification accuracy. CNN is an enriched deep-learning technique that extracts local and spatial features through convolutional and non-linearity operations. However, it only captures spatial information disregarding temporal information for emotion prediction [60]. Therefore, combining CNNs with Recurrent Neural Networks (RNN) can better handle spatial as well as temporal dependencies of EEGs. Alhagry et al. tested the capability of Long-Short Term Memory (LSTM) Recurrent Neural Network in emotion estimation with raw EEG signals as input and compared the results with previous prominent methods [67]. They indicated that their method outperforms several existing methods for emotion recognition. Nath et al. also evaluated the performance of their proposed LSTM as well as other benchmark classifiers such as KNN, SVM Decision Tree, and Random Forest for emotion recognition [68]. They

compared subject-dependent and subject-independent strategies for EEG-based emotion recognition using their proposed LSTM-based model. The subject-dependent strategy involves training separate LSTM models for each participant, while the subject-independent strategy involves training a single LSTM model for all participants. They found that the subject-dependent strategy outperforms the subject-independent strategy in terms of accuracy and F1-score. Their results also showed that their proposed LSTM-RNN achieves a higher classification accuracy for emotion recognition. Multilayer perceptron (MLP), a feedforward deep neural network (DNN) with multiple hidden layers is also proven as a powerful deep learning technique in various recognition tasks [69, 70]. Pandey et al. examined the capability of wavelet coefficients of a single EEG electrode in estimating the user's emotional state by MLP [71]. They compared their outcomes with two previous outstanding findings in the area of EEG-based emotion detection [72, 73]. They achieved the highest accuracy when they used the wavelet coefficients of the F4 electrode in the theta band. Shen et al. proposed a four-dimensional (4D) convolutional recurrent neural network structure for emotion recognition using differential entropy (DE) features of EEG signals on a segment-level basis over four different frequency bands [74]. They compared their result with 1D, 2D, and 3D CNN structures and indicated that the emotion recognition accuracy of their 4D convolutional recurrent neural network outperforms. the others. Li et al. proposed a hybrid deep model comprised of CNN and LSTM [75]. Their results demonstrated significant improvements over several existing baseline methods such as SVM, CNN, and Random Decision Forest. Yang et al. proposed a Bidirectional LSTM (BiLSTM) network for EEG-based emotion classification [76] using DE features where the network models sequential information in the backward and forward

directions. They compared their proposed model with two previous models using the same dataset and found that BiLSTM performs better than both of the previous models.

3.1.1 Feature Extraction

In general, a typical EEG emotion recognition network is comprised of two major parts, i.e., discriminative EEG feature extraction and emotion classification. As mentioned previously in Section 1.2.2, EEG features used for emotion recognition can be captured through their time-dependent or frequency-dependent components or a mix of both. One of the most commonly used EEG frequency-dependent feature extraction methods maps the EEG signal frequency range into several bands, namely δ (0 to 4Hz), θ (4 to 7Hz), α (8 to 12Hz), β (13 to 30Hz) and γ (31-50Hz) [77], and extract features from each band. There are extensive studies investigating various sets of EEG features for emotion recognition [78, 79]. Differential Entropy (DE) [80] features have been widely used in state-of-the-art emotion recognition models [81, 82]. These features have outperformed other EEG feature sets such as differential asymmetry (DASM), rational asymmetry (RASM), differential causality (DCAU), and PSD [83-85]. DE is generally equivalent to the logarithmic spectral energy for a fixed-length EEG signal in a specific frequency band [76]. Cheng et al. constructed a 2D frame for each sample, derived from the spatial distribution of EEG channels. This frame was utilized as the feature input to their classification model [85]. They used Cascade Forest which is constructed by Deep Forest (DF) to perform emotion classification. They compared the performance of their model with several existing emotion recognition models on two benchmark datasets and achieved a higher emotion recognition accuracy for their proposed model.

3.1.2 Representation Learning using Deep Networks

Beyond DE features, researchers have also leveraged deep networks to learn EEG representations and model the relationship between different EEG channels. For example, as there is long-term dependency in EEG data [20], a Recurrent Neural Network (RNN) [21] was used to capture more robust features in sequential data. BiLSTM network deployed in [76], models sequential information in the backward and forward directions to capture more discriminative features for emotion recognition.

Convolutional Neural Networks (CNNs) and attention mechanisms have been deployed to extract emotion-related EEG representations [86-89]. Cui et al. proposed regional asymmetric CNN to extract more discriminative features from EEG signals followed by a fully connected layer to perform emotion classification [86]. Their CNN structure is composed of three different networks to extract temporal, asymmetric, and regional features. Then, the output of the three networks is combined to further be fed to the classifier. They compared their model with several existing benchmark emotion recognition models on two public datasets and demonstrated that their proposed model improves emotion recognition accuracy. Maheshwari et al. proposed a multi-channel deep CNN to extract high-level EEG features [87]. They mapped EEG signals into different frequency bands to use as the input to their CNN model. Their results show that their proposed model achieves a higher accuracy in emotion classification, outperforming other state-of-the-art methods. Tao et al. proposed an attention-based convolutional recurrent neural network to extract more discriminative EEG representations [88]. Their proposed model deploys a channel-wise attention mechanism first to update the weights of different EEG channels adaptively. Then, CNN is applied to the encoded EEG signals to extract

spatial information followed by LSTM-RNN integrated with a self-attention mechanism to capture temporal information. Their results demonstrate that their proposed model achieves a higher emotion recognition classification in comparison with SVM, CNN, CNN-RNN, GCNN, and channel-wise attention mechanism combined with CNN-RNN and CNN-RNN with extended self-attention mechanism. Simonyan et al. proposed a deep hierarchical CNN called VGG [89]. They tested their model on a large public image dataset and found that increasing the depth of their network leads to better performance. They compared their proposed model with several state-of-the-art models and concluded that their proposed model achieves a better classification accuracy. Graph Neural Networks (GNNs) have been employed to learn the spatial relation between EEG features among different EEG electrodes [90-92]. The dynamic graph convolutional neural network (DGCNN) was proposed in [90] to capture intrinsic relationships of EEG channels by a trainable adjacency matrix. They compared their method with a GCNN with a predetermined adjacency matrix, SVM, and Deep Belief Network (DBN) and achieved higher emotion classification accuracy by their proposed method. Zhong et al. proposed two regularizers as node-wise domain adversarial training and emotion-aware distribution learning which they incorporated with DGCNN to deal with inter-subject EEG variations [91]. They tested their model on two publicly available datasets and demonstrated that their proposed model consistently outperforms the previous state-of-the-art models. Zhang et al. proposed a DGCNN by applying a sparseness constraint on the graph [92]. They showed that their proposed model achieves better emotion classification accuracy compared to SVM, DBN, and DGCNN. In addition to comparing their proposed model with existing models, they

also compared different features and spectral bands such as DE, PSD, DASM, and RASM on different bands of several publicly available datasets.

Some researchers combined various deep networks to extract deep EEG representations [93-95]. Yin et al. combined Graph Convolutional Neural Network (GCNN) with LSTM-RNN to capture both spatial and temporal relationships among EEG channels [93]. They compared their model with several state-of-the-art models and demonstrated that their model achieved a better emotion classification result. Li et al. took advantage of the attention mechanism combined with bidirectional LSTM-RNN for multimodal emotion recognition [94]. They indicated that the deep features extracted from their proposed network could reach a higher emotion recognition accuracy. Du et al. proposed a model constructed of an attention-based auto-encoder, an LSTM-based feature extractor, and a domain discriminator [95]. Their proposed model is designed to be efficient, using a smaller number of LSTM cells and fewer parameters than other state-of-the-art models while achieving comparable performance. They compared their proposed model with SVM, DBN, GCNN, and DGCNN and demonstrated that their model outperforms. Yang et al. utilized the combination of CNN and LSTM to learn deep representations of EEG signals [96]. Their proposed model consists of parallel convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which are used to extract spatial and temporal features from the EEG signals, respectively. The extracted features are then combined and fed into a fully connected layer for emotion recognition. They compared their model with several previous emotion recognition models and showed that their model achieved a higher emotion classification performance.

Deep auto-encoders and attention-based auto-encoders have also been deployed by researchers to extract high-quality EEG features through encoding and decoding input data [97-99]. One popular approach is to use recurrent neural networks (RNNs) for encoding and decoding EEG signals. RNNs are well-suited for processing sequential data, and EEG signals can be considered as a sequence of data points over time. In this approach, the RNNs are trained to learn the temporal dependencies between the EEG signals and use these dependencies to extract high-quality features to further perform emotion recognition. Zhang et al. deployed a deep recurrent autoencoder to recognize emotion from EEG signals. The AE is trained separately to extract EEG representations and then the network is followed by two fully connected dense layer classifier which performs emotion classification task [97]. They compared their method with several popular existing state-of-the-art emotion classification methods and demonstrated that their method outperforms the previous methods. Deep neural networks (DNNs) are also often used as building blocks for both the encoder and decoder in deep autoencoder models. Authors in [98] have combined CNN with a deep sparse auto-encoder made of three-layer DNN to extract EEG latent representations then, it is followed by a deep neural network (DNN) which is comprised of three fully connected dense layers to perform emotion classification. They evaluated their proposed model using two popular public datasets and compared it with several emotion classification models where their model showed superior performance than the existing emotion classification models. Rajpoot et al. proposed a model consisting of an LSTM with a channel attention autoencoder to extract high-level EEG representations [99]. They also deployed a CNN with an attention mechanism to perform emotion classification. Another approach is to use convolutional neural networks (CNNs) for

encoding and decoding EEG signals. CNN-based autoencoders are particularly more effective for image data, where they can learn feature hierarchies that capture both local and global spatial patterns in the image [100]. Chen et al. proposed convolutional autoencoder neural networks (CAE) for medical image analysis which showed improvement in the accuracy and efficiency of medical diagnoses [100]. The authors demonstrated the effectiveness of their proposed approach by applying it to three different medical image analysis tasks: brain tumor segmentation, lung nodule classification, and diabetic retinopathy diagnosis. In each case, the CAE model is trained on a large dataset of medical images, and the learned features are used for image classification and segmentation tasks. The authors compared their proposed models with several state-of-the-art models and reported that their approach outperforms existing methods on each of the tasks. Chai et al. propose a novel approach for EEG-based emotion recognition using unsupervised domain adaptation techniques based on auto-encoder models [101]. The authors note that EEG signals recorded from different subjects may have different characteristics, making it difficult to generalize emotion recognition models trained on one subject to another subject. To address this problem, the authors propose an unsupervised domain adaptation technique based on auto-encoders, which can learn a shared feature representation between the source and target domains. The proposed model consists of an encoder that maps the EEG signal to a lower-dimensional space, a decoder that reconstructs the signal from the lower-dimensional representation, and a domain classifier that distinguishes between source and target domains. The model is trained using both labeled data from the source domain and unlabeled data from the target domain and is optimized to minimize the reconstruction error and the domain classification loss. The authors evaluated their proposed approach on a

dataset of EEG signals collected from multiple subjects and compared it to several baseline methods. They showed that their proposed method outperforms the baseline methods and achieves a better performance on the dataset.

Contrastive learning has also been widely and successfully used for general data-representation learning [102, 103]. Le-Khac et al. provide an extensive review of recent applications of contrastive learning [102], including image and video recognition, natural language processing, speech recognition, and generative modeling. The authors highlighted the advantages of contrastive learning over other unsupervised learning methods, such as autoencoders, and discussed how contrastive learning can be used in conjunction with supervised learning to improve model performance. CL has achieved superior performance in various fields such as bioinformatics [104], natural language processing (NLP) [105], and computer vision [106]. CL intends to project the data into a space where different views of the same input sample have highly similar representation. Although, it was originally used for image classification applications [106, 107] and generally, it is widely adopted in the computer vision domain [108] but has also been applied to physiology-based emotion recognition [109, 53, 110] recently. Li et al. proposed a method based on a graph convolutional neural network (GCNN) that is trained using self-supervised contrastive learning [104]. They used GCNN to encode molecular structures into global representations that capture important features relevant to drug discovery. The self-supervised contrastive learning component ensures that the learned representations are invariant to various molecular transformations and can capture subtle differences between molecules. They evaluated their proposed model on a dataset of molecular structures and compared it with other state-of-the-art methods for molecular representation learning. The

results show that the proposed method outperforms other methods for several downstream tasks. Authors in [106] proposed a simple framework for contrastive learning of visual representations, which is a technique used to train neural networks for visual tasks such as image recognition. The proposed framework, called SimCLR, involves training a neural network to learn representations of image patches that are similar to each other while being dissimilar to patches from other images. The authors showed that SimCLR outperforms previous state-of-the-art methods on several benchmark datasets for image classification. They also showed that their proposed model is more efficient computationally than the existing methods. They also provided an ablation study to show the importance of various components of the SimCLR framework. Khosla et al. presented a new contrastive learning method for learning discriminative and invariant feature representations using labeled data [107]. They evaluated their proposed method on several benchmark datasets for image classification and found that their proposed method achieves a higher performance in comparison with the state-of-the-art methods. Mohsenvand et al. [109] used a similar method to SimCLR presented in [106] which is called SeqCLR to learn similarities between differently augmented transforms of the same EEG data sample disregarding the emotional state of the data sample. Augmented transforms were generated using temporal masking, linear scaling, time shifting, DC shifting, band-stop filtering, and Gaussian noise adding. Their proposed model is comprised of CL and BiLSTM to perform representation learning and classification, respectively. They also compared their model with several state-of-the-art models on EEG-based emotion recognition and concluded that their model achieves a higher classification accuracy.

Pinitas et al. [53] proposed a model to learn general affect-infused multimodal representations which was built upon the Contrastive Learning framework introduced in [107]. Their model is a combination of CL and one dense layer to perform representation learning and classification, respectively. The authors evaluated their method on several benchmark datasets for affect modeling and showed that their proposed model outperforms several baseline methods. They also presented an ablation study to show the importance of different components of the proposed method. Their results show that using CL improves multimodal affect modeling tasks. However, in their proposed CL, they only considered a single emotional dimension for their pairing mechanism.

Shen et al. employed CL to address the problem of inter-subject variability which they define as the disparity in the EEG signals of any two subjects exposed to the same stimulus [110]. Hence, they proposed a model using CL which maximizes the similarity between the representations of the EEG signals that are collected in response to identical stimulus followed by three dense layers to perform emotion classification. They compared their model with SeqCLR, CorrCA, SA, and simple multilayer perceptron using DE as input features and proved that their model outperforms. The drawback of their CL method is that the effect of intra-subject variability was neglected in their work. They also did not consider any augmentation technique for their limited dataset.

In general, inter-subject variability refers to the difference in brain functionality across different subjects whereas, intra-subject variability refers to the difference in brain functionality within one subject. In Chapter 5, we propose a model that addresses the problem of both inter-subject variability where different subjects are exposed to the same stimulus, and, intra-subject variability where one single subject is exposed to different

stimuli with the same emotional state. To the best of our knowledge, this is the first instance of using CL for EEG-based affect modeling to address both inter-subject and intra-subject variability.

3.1.3 Data Augmentation

As processing is more complex in deep learning architectures-i.e., more model parameters are involved-, a relatively large amount of data is required for training a deep model. However, high-grade EEG data collection is expensive and requires a relatively high amount of time from both investigators and participants. Therefore, researchers have deployed DA techniques to synthetically enlarge training datasets which has been shown to increase the DNN model's performance considerably [54]. There are several approaches to data augmentation, including traditional signal processing techniques, generative models such as GANs, and rule-based methods [54]. The choice of data augmentation technique depends on the type of data being used, the specific recognition task, and the available computational resources.

Wang et al. explored the use of data augmentation techniques for improving the performance of CNNs in EEG-based emotion recognition [111]. They added Gaussian noise to training EEG data to produce new samples for the emotion recognition task. They used DE features of the five frequency bands for their proposed deep network. They evaluated the performance of the augmented datasets against the original dataset using accuracy and F1-Score. Their experimental results showed that by augmenting the training dataset 30 times, the accuracy of their model improved notably. Salama et al. proposed a deep method for EEG-based emotion recognition using 3D CNNs. They augmented their training EEG data by adding noisy signals constructed from Gaussian noise with zero mean

and unit variance [112]. They compared the performance of their proposed method against several baseline methods, including a 2D CNN and SVM, and demonstrated that the DA technique improved the performance of their proposed 3D CNN for emotion recognition tasks. Shankar et al. presented a comparative study of data augmentation techniques for deep learning-based emotion recognition from EEG signals [113]. The authors evaluated the performance of five data augmentation techniques, including random cropping, random rotation, random flipping, random rescaling, and random translation on three different deep learning models, namely 1D CNN, 2D CNN, and LSTM on two popular datasets. The results show that all data augmentation techniques improve the performance of the deep learning models compared to the original datasets. The best performance is achieved by using a combination of random cropping, random rotation, random flipping, and random translation. The 2D-CNN model outperforms the other models, achieving an accuracy of 73.58% using the augmented data.

Generative Adversarial Networks which were first introduced by Goodfellow et al. have been widely applied as a DA technique in many fields including EEG-based recognition [114-116]. Authors in [114] proposed a novel method called EEG-GAN, which is a GAN-based approach for generating realistic EEG brain signals. The authors use the Wasserstein GAN architecture with a gradient penalty to ensure stable training and to produce high-quality EEG signals. The generator network in the EEG-GAN model generates synthetic EEG signals, while the discriminator network distinguishes between real and synthetic signals. The authors evaluated the performance of the EEG-GAN model by comparing the generated EEG signals with real EEG signals in terms of spectral power, coherence, and event-related potential (ERP) components. The results show that the EEG-GAN model can

generate EEG signals that are visually similar to real EEG signals and have similar spectral power and coherence properties. The EEG-GAN model also generates ERP components that are consistent with those observed in real EEG signals. Luo et al. proposed a data augmentation method for enhancing EEG-based emotion recognition using deep generative models [115]. The authors used two types of deep generative models, namely, variational autoencoder (VAE) and GAN to generate synthetic EEG signals. The synthetic EEG signals are then combined with the original dataset to form an augmented dataset, which is used to train a deep learning model for emotion recognition. The authors evaluated the performance of the deep learning model trained on the augmented dataset using the leave-one-subject-out cross-validation method. The results showed that the VAE-based data augmentation method slightly outperformed the GAN-based method and the baseline model trained on the original dataset in terms of accuracy, F1-score, and Cohen's kappa coefficient. Lue et al. proposed a Conditional Wasserstein GAN (CWGAN) to augment EEG data to further perform emotion recognition classification with an SVM as a classifier [117]. They adopted a set of indicators to judge the quality of the generated data. Their results indicated that their proposed CWGAN significantly improves the emotion classification accuracy when appending generated EEG data to the actual dataset. Luo et al. adopted a Conditional Boundary Equilibrium GAN (CBEGAN) [118] to generate artificial training data for multimodal emotion recognition [116]. They further used an SVM with a linear kernel as their model's classifier. They tested their model on two different datasets and their results indicated that the GAN-based data augmentation method improves emotion recognition accuracy compared to the baseline model trained on the original dataset.

3.2 Facial Expression-Based Emotion Recognition

Facial expression changes during an interaction can be used to extract the emotional state. Mehrabian showed that when expressing feelings or attitudes, 55% of communication is conveyed through visual expressions, 38% through vocal tone, and 7% through the actual words used [119]. Hence, researchers place significant emphasis on facial expression recognition for tasks related to emotion detection.

Facial expressions were traditionally extracted using manually engineered feature extraction techniques, such as Local Binary Patterns (LBP) and Local Directional Patterns (LDA) [120]. However, with the advancements in deep learning, deep neural networks have proven to be more efficient and successful in automatically extracting facial features without the need for manual techniques [120].

Tarnowski et al. [121] used a K-NN (nearest neighbor) classifier and an MLP deep neural network to recognize emotion from facial expressions. Their dataset contained pictures of individuals mimicking discrete emotions such as joy, surprise, sadness, fear, etc. according to certain instructions on a computer screen. Yu et al. proposed a model termed spatiotemporal convolutional with nested LSTM to perform emotion recognition using facial expressions [122]. They tested their model on four benchmark datasets and showed that their proposed model achieves higher performance than the state-of-the-art methods. Jain et al. proposed a model to classify images into six facial emotion classes [123]. Their proposed model is based on a deep CNN which contains convolutional layers and deep residual blocks. Their experimental results demonstrated that their proposed model outperforms recent state-of-the-art facial expression-based emotion recognition models.

Even though, facial expressions provide valuable cues about one's emotion, facial expression-based emotion recognition is not widely used using facial expression as the sole modality, especially in certain research areas, datasets, and real-world applications. This is due to the inherent challenges and limitations such as ambiguous facial expressions and the impact of cultural differences [120]. To overcome these limitations, researchers recently used multimodal emotion recognition approaches.

3.3 EEG Facial Expression Bimodal Emotion Recognition and Different Fusion Techniques

Lopez-Gil et al. [124] showed that fusing multiple modalities captured from multiple data sources in a synchronized fashion results in a better performance in emotion recognition. Gupta [7] used a hybrid BCI consisting of both EEG and functional near-infrared spectroscopy (fNIRS) to perform emotion recognition. He showed that using a hybrid BCI improves the results significantly compared to using a single modality-based BCI. His emotion recognition model consists of three modalities: EEG, fNIRS, and heart rate (HR). However, the third modality did not improve the performance significantly [124]. Hence, researchers are typically fusing multiple modalities of data typically collected through various sensors to perform emotion recognition. We selected EEG and facial expression modalities due to their recognized significance in emotion recognition [125, 119].

The fusion of modalities can be performed in typically three ways: feature-level [126], decision-level [127], and model-level fusion [128]. In feature-level fusion, features of various modalities are combined into a single feature vector, which is then processed by a

classification system. It is assumed that all modalities are perfectly synchronized temporally. In decision-level fusion, each modality is processed by a particular classification system, and then the outputs of all classifiers are combined to obtain the final result. In decision-level fusion, there is no interaction among multimodal features. This may be a disadvantage as the features from different modalities might be correlated [129]. However, it has some advantages compared to feature-level fusion. For example, we can adjust the contribution of each modality to the final result through a weighting scheme; moreover, the temporal synchronization assumption does not need to be necessarily valid as asynchronous characteristics of different modalities can exist with no interruption [130]. In model-level fusion, just like decision-level fusion, each modality may be associated with its model. However, instead of fusing the outputs of these models, we typically combine intermediate representations, such as the outputs of different hidden layers of the models. The combined representation is then fed to another classification system, thus mimicking at this stage feature-level fusion [131].

Huang et al. proposed a bimodal emotion recognition model using EEG and facial expressions [132]. They detected face position from video frames using the AdaBoost algorithm. Then, they extracted and fed the facial features to a feedforward neural network for classification [132]. For the EEG data, they first mapped the data to eight frequency regions. Then, they extracted PSD features using Short Time Fourier Transform (STFT) with a 1-s window and non-overlapping Hamming window. They also employed two decision-level fusion methods, the sum strategy and the decision-making strategy on the outputs of both classifiers to fuse the modalities. The sum strategy decision-level fusion method presented in equation (3.1) was used to combine the scores of the two classifiers

as represented in (3.1). Where, predicted outputs S_{face_j} and S_{EEG_j} (two outputs from each classification system as *j*=1, 2) from each classification system are combined. Afterwards, the index of maximum output specifies the class number as represented by r_{out} (here class 0 or 1).

$$S_{out_j} = S_{EEG_j} + S_{face_j}$$
$$r_{out} = \text{argmax}_j(S_{out}) \tag{3.1}$$

They tested their model under two experimental settings and their results show that their proposed model outperforms the state-of-the-art methods.

Canonical Correlation Analysis (CCA) was used in [133] to fuse the EEG and facial expression modalities through the feature-level fusion method. They employed manual face feature descriptors to construct a facial feature vector, while the EEG feature vector was constructed using Spectral Power (SP) complemented by Spectral Power Differences (SPD). Their results demonstrate the effectiveness of their approach in accurately analyzing and recognizing emotions in comparison with several benchmark emotion recognition models. Guo et al. investigated four different modality combinations across EEG, eye movement, and eye image data [134]. From the EEG data, they extracted DE features across five frequency bands. For eye movement features, they used a thirty-three-dimension feature set, which included pupil diameter as referenced in [135]. To extract features from eye images, they utilized a deep network composed of CNN and LSTM. To integrate the modalities, they employed a feature-level method and a Bimodal Deep Autoencoder (BDAE). In the first fusion method, they concatenated the features from various modalities (MFC) and fed them to an SVM with a linear kernel for classification.

In their second approach, each modality's features were independently inputted into a deep autoencoder to extract high-level representations. These advanced features were then classified using an SVM with a linear kernel. Their findings suggest that by integrating all three modalities, both fusion techniques effectively enhance emotion recognition accuracy.

Enumerate and AdaBoost decision-level methods have also repeatedly been used in multimodal emotion recognition research [136-139]. The AdaBoost algorithm is presented in equation (3.2).

$$S_{boost} = 1/\left(1 + exp\left(-\sum_{j=1}^{n} h_j s_j\right)\right) \tag{3.2}$$

Where, S_{boost} is the final predicted output which is used to classify emotional dimension. s_j refers to each modality's score where there is only s_1 and s_2 denoting the facial expression score and EEG score, relatively. h_j represents modalities' coefficients which are updated during the training process.

Equation (3.3) refers to the bimodal enumerate fusion method.

$$S_{enum} = \alpha S_{EEG} + (1 - \alpha) S_{face} \tag{3.3}$$

Where, S_{EEG} and S_{face} are the predicted output scores of EEG and facial expression classification systems. α Indicates the importance level of each score in the final classification based on S_{enum}. This parameter (α) varies from 0 to 1 during training to find the optimum value. The final predicted output S_{enum} is classified into high or low classes based on a threshold such that above 0.5 classifies as high and below 0.5 classifies as low.

The model presented in [136] utilizes multiple modalities, including EEG peripheral physiological signals, and facial expressions. It then combines the classification scores of all these modalities using an enumerate fusion method. Their findings underscore that the accuracy of multimodal emotion recognition surpasses that of unimodal strategies.

Since the introduction of the Transformer architecture [27], originally designed for Natural Language Processing (NLP) tasks, remarkable advancements have been achieved in diverse areas including computer vision and reinforcement learning. The Vision Transformer (ViT) [28], a modified version of the architecture introduced in [27], presents a novel approach to computer vision tasks, specifically for image classification. Transformers have been successfully used for emotion recognition due to their ability to capture contextual relationships and dependencies within sequential data. The incorporation of the attention mechanism and parallel input processing enhances the Transformer's capacity to comprehend dependencies in sequential data with heightened proficiency. Originally, the Transformer architecture consisted of two main components: Transformer encoder and Transformer decoder. The Transformer encoder is the main component employed for emotion recognition tasks.

Recently, the literature [140-144] has reported the deployment of Transformer-based emotion recognition models in both unimodal and multimodal configurations. Wang et al. proposed a Transformer-based model to perform emotion recognition using EEG signals [140]. They classified EEG signals into different brain regions and then used a Transformer to integrate information from different brain regions. They tested their proposed model on two benchmark datasets and achieved better performance compared to state-of-the-art emotion recognition models. Huang et al. proposed a Transformer-based model to fuse

audio-visual modalities using model-level fusion to perform emotion recognition [144]. They tested their model on a benchmark dataset and showed the superiority of their model under the model-level fusion configuration. These recent research studies show that the Transformer is an effective model for extracting discriminative features from modalities of information. Hence, it has gained notable attention lately for various tasks such as NLP, language translation, text generation, image generation, and sentiment analysis.

Chapter 4. Proposed EEG-based Emotion Classification Model Based on Hierarchal CNN and Block-Based Residual LSTM

As mentioned in Section 1.1, the need for automated emotion recognition has recently grown due to the increasing integration of human-computer interactions, advancements in personalized content recommendations, and the broader application of artificial intelligence in mental health and well-being platforms. However, existing emotion recognition models still do not address the challenges described in Section 1.2. Therefore, we are motivated to propose novel solutions for automated emotion detection. In this chapter, we present our first automated emotion recognition solution. This chapter is organized as follows. In Section 4.1, we describe the dataset used in this work; in Section 4.2, we present our proposed model; in Section 4.3, we evaluate the performance of the proposed model and illustrate the experimental results and analysis.

4.1 Dataset and Feature Extraction

We employ the DEAP dataset [64] to train and test our proposed model. The DEAP dataset is especially expected to be a valuable benchmark as it contains the highest number of participants compared to other publicly available EEG datasets for emotion recognition. DEAP presents data collected from 32 healthy adults aged between 19 and 37 years. During dataset collection, each subject was fitted with 32 EEG and 8 peripheral physiological sensors to produce 40 channels of signals. Each subject underwent 40 trials. For each trial, a subject was asked to watch a one-minute-long excerpt of a music video. EEG and

peripheral physiological signals as well as videos-which capture the user's facial expressions-were recorded. Since emotions can be represented using dimensional models [56], the participants self-assessed their level of arousal, valence, dominance, and liking, thus creating four labels per trial. As mentioned in Section 2.5, we only consider these two emotional dimensions for our work. The self-assessment manikins were used for the self-assessment, which resulted in a rating between 1 to 9 for each emotional dimension (1 being the least intense). EEG signals were collected by a Biosemi ActiveTwo[1] device with 32 active AgCl electrodes according to the international 10-20 system. All unprocessed data was stored in BioSemi.bdf format at a 512Hz sampling rate. However, for our work, we used the pre-processed down-sampled EEG signals at 128Hz and band-pass filtered to preserve frequency components in our region of interest (from 4 to 45Hz) to eliminate noise. Moreover, Electrooculography (EOG) artifacts that are caused by eyeball movement are removed and the EEG signals are averaged to the common reference signal to construct a spatial voltage distribution with zero-mean.

Table 4.1 refers to the number of participants, trials, EEG and peripheral channels, and samples in the pre-processed data (32 × 40 × 40 × 8064). This table also indicates the number of labels in the dataset where labels correspond to the participant ratings for valence, arousal, dominance, and liking (32 × 40 × 4) however, we only use valence and arousal for our work as these two are the key dimensions commonly used to describe emotion [29]. Many emotion recognition models and studies focus primarily on these two dimensions because they provide a meaningful and interpretable representation of

[1] http://www.biosemi.com

emotions. By using valence and arousal, we can capture the core emotional aspects without

Table 4.1. Available DEAP dataset information and features for each participant

Data Description	Data shape	Data contents
Data per Trial	40×8064	Channels × Samples
All Data	32×40×40×8064	Participants × Trials × Channels × Samples
Labels	32×40×4	Participants × Trials × Labels

dealing with additional complexity from dominance and liking, which might be less relevant to the emotion recognition task.

As can be seen in Table 4.1, we have 258,048 EEG samples per trial (32 channels × 8064 data samples) which is relatively a large number for one single trial. As training a machine learning model on raw data can be sometimes difficult due to the high input dimensionality, researchers tend to use feature extraction techniques to reduce the dimensionality of the input data while maintaining relatively the most informative data.

Therefore, we investigate four feature sets of inputs based on time domain and frequency domain characteristics to evaluate our proposed emotion recognition model as shown in Table 4.2. These input sets were previously considered for emotion recognition tasks and showed notable results in comparison with several other input sets described in the literature [63, 80, 145-147].

Tripathi et al. demonstrated the effectiveness of employing time domain features to classify emotions using EEG records [63]. Hence, for the first input set, we divide each one-minute trial into ten non-overlapping segments (each segment is 6 seconds long). As

in [63], we calculate the mean, median, maximum, minimum, standard deviation, variance, range, skewness, and kurtosis features from the one-minute preprocessed signal and its non-overlapping segments. However, as opposed to [63], we do not consider subject or stimulus-specific features, as we aim to develop a general model that does not depend on this information. We aim to assess the capacity of our model to infer the emotional status from EEG signals for an unknown subject. As shown in Table 4.2, the first input set contains 99 features per channel.

Song et al. showed that DE [80] and PSD [145] frequency-domain features result in a higher recognition accuracy compared to three other commonly used EEG features: differential asymmetry, rational asymmetry, and differential causality, especially, when DE or PSD features of the delta, theta, alpha, beta and gamma frequency bands are combined [90]. Hence, for the second input set, we collect a feature vector consisting of DE and peak value of PSD combined over the five frequency bands of interest for each channel which produces 10 features per channel.

Bashivan et al. demonstrated the advantage of using spectral images of EEG time-series signals in the context of representation learning and classification from EEG data [146]. Ozdemir et al. leveraged the same technique to construct spectral images using azimuthal equidistant projection and then used it as input to their deep learning network [147]. Therefore, for the third input set, we collect gray scale EEG spectrum images obtained from EEG signals using a 512-point fast Fourier transform [146].

Shao et al. showed that using the EEG signals directly as the input to a deep learning network can improve recognition accuracy by comparing their results with several state-

of-the-art feature-based methods [148]. Therefore, for the last input set, we employ the preprocessed EEG signals without further feature extraction.

Table 4.2. Description of the four different sets of inputs extracted from the DEAP dataset

Input set description	Data shape	Data contents
1) Time domain features	32×99	Channels × Features
2) Peak PSD and DE features	32×10	Channels × Features
3) Spectral Images	32×50×50	Channels × Pixels × Pixels
4) EEG Signals	32×8064	Channels × Samples

As mentioned in Section 2.5, we associate each emotional dimension with two classes: high and low with a 5 rating as the boundary. This allows us to distinguish between a low or high intensity of each emotional dimension (e.g., low arousal vs high arousal).

4.2 Proposed Model

We propose a deep-learning solution to classify emotions using the DEAP dataset. The proposed neural network consists of two main components: a pre-trained CNN and a residual LSTM as shown in Figure 4.1. The cascading of CNN and LSTM has already been presented in the literature [17]. Nonetheless, the proposed method presents a novel structure for emotion detection that leverages a pre-trained CNN, namely GoogLeNet [149], for feature extraction and a block-based residual LSTM network for classification. We selected a pre-trained CNN for the proposed architecture given the size of our dataset. GoogLeNet has shown notable performance in extracting features over image sequences for serval applications [150]. As described in Section 1.2, a conventional LSTM can

effectively model long sequences without being affected by the vanishing and exploding gradient issues that can afflict RNNs. However, a residual LSTM includes additional spatial skip connections from lower layers, which is assumed to improve the network's training by creating shortcut paths [22].

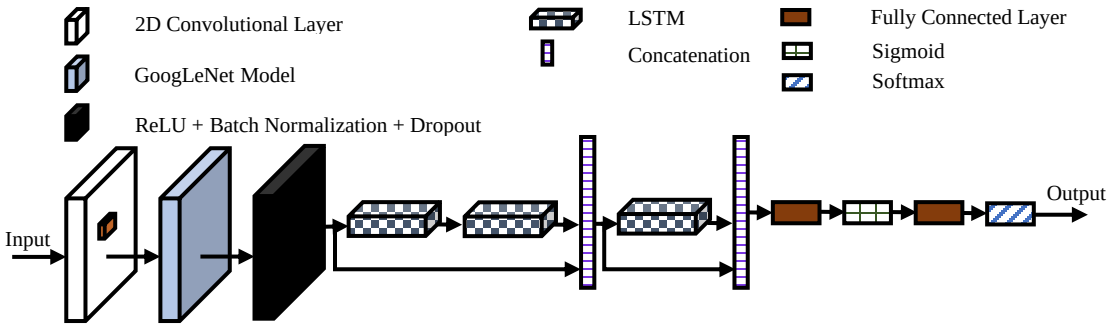


Figure 4.1. Architecture of the proposed model.

The idea of residual learning was first introduced by [151] to train deep CNNs for image recognition. These networks use shortcut connections such that the input of a layer is the summation of the outputs of the previous layer and a layer 2 or 3 levels below the previous layer. LSTM residual connection is presented in [22] and applies the same concept to the LSTM layers. In the proposed work, we present a block-based residual LSTM where the output of a block is a concatenation of the input vector to the block and the residual function learned by the corresponding block.

The architecture of the proposed model is depicted in Figure 4.1. To feed the inputs represented in Table 4.2 into the pre-trained CNN, we increase the number of channels to 3 as in [149]. Hence, we use a 2D convolutional layer with a kernel size of 3 x 3 and a stride of 1 over the input to satisfy this criterion. We utilize a ReLU (Rectified Linear Unit) activation function and batch normalization on the output of the convolutional layer. Then,

we feed the output of the batch normalization layer to the pre-trained GoogLeNet CNN network. We remove the top fully connected layer of the CNN as we employ the CNN as a feature extractor. Hence, the output of the pre-trained CNN corresponds to the model's intermediate features, a vector with a size of 1024. We divide the LSTM layers in the proposed networks into blocks. To realize the residual connections, the input and output of each block are concatenated and fed to higher layers of the network. Each block can observe inputs from previous blocks except the first block, which receives inputs from the pre-trained CNN component. The idea behind blocking the LSTM layers is inspired by the TCN proposed in [152].

We feed 16 sequences of 64 intermediate feature vectors produced by the pre-trained CNN to the first LSTM block. Each LSTM block's output propagates to the next LSTM block's input as a concatenation vector of the last LSTM block's input and output. If we have an odd number of optimum LSTM layers (according to the hyperparameter optimization process), then the last block only contains one LSTM layer.

There are various search/optimization techniques in the literature [44] to find the optimal hyperparameters for a model. The Gaussian search method [153] is used to yield the best hyperparameter combination for the proposed model. The hyperparameter optimization often helps improve the performance of machine learning models [152]. The hyperparameters used in this work for optimization include learning rate, number of LSTM layers (which consequently decides the number of blocks), number of stacked LSTMs on each LSTM layer, size of the first LSTM layer, and step size to decrease the output size of each LSTM layer accordingly. The hyperparameter optimization process rendered the

following values: learning rate of 0.0001, 2 stacked LSTMs, hidden size of 100 for the first LSTM, 3 LSTM layers, and step size of 15.

We use linear layers for the classification of the related model's output. We feed the output of the last LSTM layer to two fully connected dense layers for the classification of the related model's output. As activation functions, we use a sigmoid activation function for the first dense layer and a softmax activation function for the second dense layer to output class probabilities.

4.3 Results and Performance Analysis

To assess the performance of the proposed model, the data of one subject is excluded as a testing set. Thus, we train and validate the model on the data of the remaining 31 subjects. We use a batch size of 50 and for the choice of epoch, we used Grid Search [44] over 100 epochs. We chose the best epoch number based on when the training and testing accuracy reached the maximum. To train the model, we employ a 4-fold cross-validation over our train-validation set. We use a Cross-Entropy (CE) loss function and Adam optimizer [43] with a 0.9 momentum, 10^{-6} weight decay, and 0.0001 learning rate.

Equation 4.1 shows how we compute the CE loss [154], where y_i is true class number corresponding to input and h is an array of output probabilities for different classes. So, h_{y_i} is the true label probability for the corresponding input and j is the total number of classes. To decrease the risk of over-fitting, we use dropout with a probability of 0.2 on the output of pre-trained CNN and 2D convolutional layer.

$$\text{Loss}_i = -h_{y_i} + \log \sum_j e^{h_j} \tag{4.1}$$

Equation 4.1 shows how we compute the CE loss [154], where y_i is true class number corresponding to input and h is an array of output probabilities for different classes. So, h_{y_i} is the true label probability for the corresponding input and j is the total number of classes. To decrease the risk of over-fitting, we use dropout with a probability of 0.2 on the output of pre-trained CNN and 2D convolutional layer.

We evaluate the proposed model using two testing phases. In the first testing phase, we compare the performance of the proposed model to existing state-of-the-art models. We test four instances of the proposed model which are identical except for the use of a different input set (see Table 4.2). In the second testing phase, we compare the performance of the proposed model to a CNN+LSTM network identical to the proposed one except for the use of a conventional LSTM structure instead of the block-based residual LSTM. This allows us to isolate the effects of the block-based residual on the performance.

4.3.1 First Phase

We perform a four-fold cross-validation on the training/validation set. We calculate the average accuracy for the four folds. Once the model is fit, we compare the cross-validation performance results to those of the test set. In addition to accuracy, we calculate the F1-score for each class to ensure the result's reliability as described in [155]. This metric describes the prediction capability of the assessed model for each class (low=0 and high=1).

Table 4.3 compares the results of the proposed model with four different input sets to seven popular recent state-of-the-art benchmarks, namely DNN and CNN by Tripathi et al. [63], RNN by Alhagry et al. [67], wavelet coefficient based multilayer perceptron (MLP) in two different frequency bands – alpha and theta by Pandey et al. [71], TSception

composed of temporal and spatial CNN by Ding et al. [66], CNN+LSTM architecture followed by DNN layers from Hizlisoy et al. [17], and BiLSTM by Yang et al. [76]. We implemented all compared models using the PyTorch library [156].

The proposed model can be deployed in a variety of configurations. For instance, since all models developed by PyTorch are compatible with edge computing technology, the proposed model may be trained on a high-end server but deployed on an edge computer, if such a system does not have a real-time processing requirement. Note that the model has a minimal memory footprint and requires approximately 26.58 MB of storage space. Alternatively, the model may be deployed in a cloud computing setup where an edge computer communicates collected inputs to cloud servers for processing. Such deployment has the potential to reduce latencies associated with the recognition task significantly. The size of a DEAP dataset trial is 7.88 MB. Hence, using the latest communication technologies (e.g., fiber cable, 4G, and 5G), the data can be relayed quickly to cloud servers where it would be processed swiftly for an output to be returned to the edge computer.

As we can see in Table 4.3, the proposed model with EEG signals as input achieves an average validation accuracy of 0.61 and 0.63 for valence and arousal, respectively. The same model achieves a testing accuracy of 0.65 and 0.68 for valence and arousal, respectively. The proposed model with EEG signals as input provides the smallest discrepancy between the validation and testing performance and relatively higher F1 scores for both classes compared to the proposed model with the other three considered input sets. This can happen because the raw EEG signals contain all the available information about the data, and feature engineering may not always capture the most important aspects of the data for the task at hand. Moreover, deep learning models such as CNNs are specifically

designed to extract informative features directly from raw input data, a process that is often expected to yield better performance than using manually engineered features.

The discrepancy between the validation and testing results for DNN [63], CNN [63], CNN+LSTM [17], and BiLSTM [76] may be due to over-fitting on the training/validation set. RNN [67], MLP with wavelet coefficients in theta band [71] and TCN [66] do not seem to overfit, however, they achieve inferior accuracy and F1-score results to the proposed model. RNN [67] achieves a 0 F1-score for class 0; this means that the model is unable to predict low valence and low arousal and TCN [66] achieves a 0 F1-score for class 0 of the arousal dimension. Therefore, despite achieving relatively reliable classification accuracy, these three models [66, 67, 71] do not perform as well as the proposed model. Conversely, DNN [63] performs poorly for the prediction of class 1 for both emotional dimensions according to the F1-score results.

As mentioned before, we used GoogLeNet as a pre-trained CNN for the proposed model as it was the only pre-trained CNN compatible with the size of all four input sets. However, as can be seen from Table 4.3, if we disregard the proposed model with raw EEG signals as input, the best performance is achieved by the proposed model with time-domain features as input. If we consider the time-domain feature set as the input for the proposed model, we can test several other pre-trained CNNs for the proposed model to investigate its performance in the case of other pre-trained CNNs. Therefore, before moving forward to the second testing phase, we briefly compare the effect of several pre-trained CNNs on the proposed structure in the following.

Table 4.3. Performance comparison for valence and arousal classification for 2 classes on the DEAP dataset

Models	Valence			Arousal		
	F1-Score: Class 0 Class 1	Train Acc.	Test Acc.	F1-Score: Class 0 Class 1	Train Acc.	Test Acc.
DNN [63]	0.64 0.00	0.61	0.47	0.34 0.00	0.63	0.20
CNN [63]	0.57 0.66	0.97	0.63	0.34 0.00	0.56	0.20
RNN [67]	0.00 0.69	0.54	0.53	0.00 0.88	0.55	0.70
MLP (theta band) [71]	0.36 0.42	0.56	0.55	0.48 0.63	0.57	0.62
MLP (alpha band) [71]	0.57 0.52	0.56	0.47	0.40 0.53	0.57	0.35
TCN [66]	0.67 0.26	0.51	0.54	0.00 0.75	0.57	0.59
CNN+LSTM [17]	0.42 0.47	0.99	0.51	0.14 0.56	0.99	0.44
BiLSTM [76]	0.44 0.54	0.79	0.50	0.68 0.00	0.79	0.52
Proposed model: time-domain features	0.80 0.80	0.56	0.80	0.27 0.83	0.62	0.73
Proposed model: DE+PSD features	0.52 0.65	0.56	0.62	0.40 0.86	0.58	0.78
Proposed model: EEG spectrums	0.16 0.67	0.56	0.52	0.02 0.82	0.58	0.70
Proposed model: EEG signals	**0.56** **0.71**	**0.61**	**0.65**	**0.52** **0.69**	**0.63**	**0.68**

Table 4.4. Comparison among benchmark pre-trained CNNs for the proposed model

Models	Valence			Arousal		
	F1-Score: Class 0 Class 1	Train Acc.	Test Acc.	F1-Score: Class 0 Class 1	Train Acc.	Test Acc.
VGG11	0.37 0.67	0.57	0.57	0.00 0.82	0.67	0.70
ResNet18	0.32 0.69	0.63	0.57	0.14 0.81	0.66	0.70
EfficientNet	0.18 0.68	0.56	0.55	0.40 0.86	0.60	0.77
GoogLeNet	0.60 0.60	0.60	0.60	0.15 0.83	0.62	0.72

Table 4.4 indicates the results of pre-trained CNN combined with block-based residual LSTM with time-domain features as its input in the case of four different architectures for the CNN component (VGG11 [89], ResNet18 [151], EfficientNet [157], and GoogLeNet [149]). We selected these architectures as they have shown notable performance in extracting features over image sequences for serval applications [158, 150]. There exist deeper versions of the same pre-trained CNNs such as VGG19 and ResNet50 which are comprised of a higher number of convolutional layers. Therefore, they include more parameters than the versions we selected, which makes the processing and training of the models computationally more expensive and longer.

Also, for small datasets, it is often recommended to use a simpler neural network architecture with fewer parameters, as complex models like ResNet50 or VGG19 may overfit the data and result in poor generalization to new, unseen data. In general, these architectures contain a series of convolutional stages and fully connected layers. There are

several numbers of pre-set hyperparameters in these configurations to hierarchically extract features.

Table 4.4 compares the results of the four pre-trained CNNs combined with block-based residual LSTM models with different pre-trained CNNs such as VGG11, ResNet18, EfficientNet, and GoogLeNet. The same hyperparameters as Section 4.2 were used to optimize these models. As we can see in Table 4.4, among the several pre-trained CNNs, GoogleNet and ResNet show a better performance compared to the other two pre-trained CNNs. VGG11 and EfficientNet achieve inferior accuracy on the valence dimension. Also, VGG11 achieves 0 F1-score for class 0 on arousal dimension which means the model is unable to predict low arousal. ResNet18 achieves slightly better training accuracy on both valence and arousal dimensions compared to GoogLeNet; however, GoogLeNet provides a slightly higher accuracy and F1-score on the test dataset on both valence and arousal dimensions expect for class 1 on valence dimension. This means that GoogLeNet is identifying positive cases in each class more accurately than ResNet on the test dataset. In general, the F1-score of class 0 for the arousal dimension is relatively low for all models which might be due to imbalanced instances of our test dataset for class 0.

4.3.2 Second Phase

In this phase, we compare the proposed model to a network composed of a pre-trained CNN (GoogLeNet [149]) and n number of conventional LSTM layers without the use of residual blocks. We refer to this model as conventional CNN+LSTM. To decide on the number n, we tested an interval of [2, 6] and achieved the best recognition accuracy when n is 3. We adopt the preprocessed EEG signals as the input set given that it performed best in the previous testing phase (as discussed in Section 4.3.1).

The experimental results in Table 4.5 demonstrate that the proposed model outperforms the conventional CNN+LSTM model by at least 5% on the training accuracy. The proposed model also achieves more reliable results on the F1-scores for both classes as the conventional CNN+LSTM model performs poorly on class 0 for both dimensions (0.30 and 0.40 for valence and arousal respectively). The proposed model shows relatively less discrepancy between the validation and test accuracy for the arousal dimension compared to the conventional CNN+LSTM model. However, it is important to acknowledge a limitation regarding the size of the test data, as only data from a single subject was included as test data. Consequently, placing significant reliance on the testing accuracies may not be advisable due to the restricted diversity and representation inherent in such a small dataset.

Table 4.5. Comparison of the proposed model with conventional CNN + LSTM

Models	Valence			Arousal		
	F1-Score: Class 0 Class 1	Train Acc.	Test Acc.	F1-Score: Class 0 Class 1	Train Acc.	Test Acc.
CNN+ conventional LSTM	0.30 0.64	0.56	0.52	0.40 0.86	0.57	0.77
Proposed model	0.56 0.71	0.61	0.65	0.52 0.69	0.63	0.68

Generally, supervised deep neural networks require a large training dataset to reach optimal classification performance. Our progress in improving classification accuracy was partly hampered by the small size of the dataset. This also limited our ability to create a larger test set since an increase in the size of the latter would reduce the size of the

training/validation set and hence hinder training. An approach to address this limitation is to expand the size of the dataset artificially which will be presented in the next chapter.

Chapter 5. A Graph Neural Network for EEG-Based Emotion Recognition with Contrastive Learning and Generative Adversarial Neural Network Data Augmentation

In this chapter, we present our second automated emotion classification solution which addresses several existing challenges in an integrated architecture consisting of CL, GAN, and GNN.

The chapter is organized as follows: in Section 5.1, we introduce the datasets we employ to train, validate, and test our solution. In Section 5.2, we describe the proposed solution. In Section 5.3, we perform experimental analysis and evaluation of the proposed model.

5.1 Datasets

To evaluate the performance of the proposed model, we conducted experiments on the publicly available DEAP [64] and MAHNOB-HCI [159] emotion recognition databases. The MAHNOB-HCI database was created under similar experimental conditions as the DEAP database. Although these datasets contain multiple modalities, we only utilized the EEG modality from the DEAP and MAHNOB-HCI databases to evaluate our proposed model.

MAHNOB-HCI dataset presents data collected from 27 healthy adults aged between 19 and 40 years old. Similar to DEAP, during the dataset collection, each subject was fitted with EEG and peripheral physiological sensors. Facial expressions and body movements of users were also captured using six cameras. Each subject watched 20 music videos

which resulted in 20 trials per subject. Although the length of the videos ranged from 94 to 176 seconds, only the recordings captured during the final 60 seconds of each stimulus were used for sub-sequent processing and analysis [160]. Similar to the DEAP dataset, at the end of each trial, the subjects were asked to self-rate their arousal, valence, dominance, and sense of predictability on a scale of 1 to 9. Furthermore, they were told to self-report the emotion they felt using emotional keywords. To ensure a more efficient emotion recognition, we applied the same pre-processing steps that were used on the EEG signals in the DEAP dataset (described in Section 4.1).

There are 3s pre-trial baseline recordings in the DEAP dataset that were also removed. Similar to the DEAP dataset, the EEG data for MAHNOB-HCI was also captured using a 32-channel Biosemi ActiveTwo device.

A comparison between DEAP and MAHNOB-HCI datasets information is presented in Table 5.1.

For classification purposes, the same emotional model as presented in Sections 2.5 and 4.1 is used to analyze the performance of the proposed model.

The best input feature set for a deep network depends on the specific task and the type of data being used. We investigated the effect of four different input sets for an identical task and identical data (EEG) in Chapter 4 and demonstrated that using EEG signals directly without any hand-crafted features performed better than feature-engineered input sets. Therefore, we use preprocessed EEG data for our work.

Table 5.1. Information on the DEAP and MAHNOB-HCI databases

Feature	DEAP / MAHNOB-HCI Description
Number of subjects	32 / 27
Recorded signals	EEG, respiration signal, PSG, EOG, EMG, GSR, skin temperature / EEG, respiration signal, ECG, GSR, skin temperature
Recorded video	Face / Face and body (6 cameras)
Number of experiments	40 /20
Number of EEG channels	32 / 32
Experiment length	60s (128hz) / 94s-176s (256hz)
Rating scales	Valence, arousal, dominance and liking / Emotional keywords, Valence, arousal, dominance, and predictability
Rating values	1-9 / 1-9

5.2 Proposed Model

To ensure a fair performance assessment, we trained and evaluated our model using two strategies:

- We shuffled all the trials and partitioned the dataset, allocating 80% for training and 20% for testing. This approach ensures that the model is never exposed to the test set during the training process.
- We implemented a leave-one-subject-out cross-validation (LOSOCV) strategy. For each cross-validation fold, data from one subject is reserved for validation, while data from the remaining subjects is used for training.

This method has its limitations as it results in a small testing set that corresponds to the data of a single subject. However, importantly, it provides a subject-independent evaluation strategy.

In both strategies, we augment the training data with synthetic data using a GAN component. The GAN component exclusively uses the training data to generate synthetic data (Section 5.2.2). Our model is comprised of three main components: CL, GAN, and GNN as depicted in Figure 5.1. We will explain the different components of the proposed model in the subsequent sections.

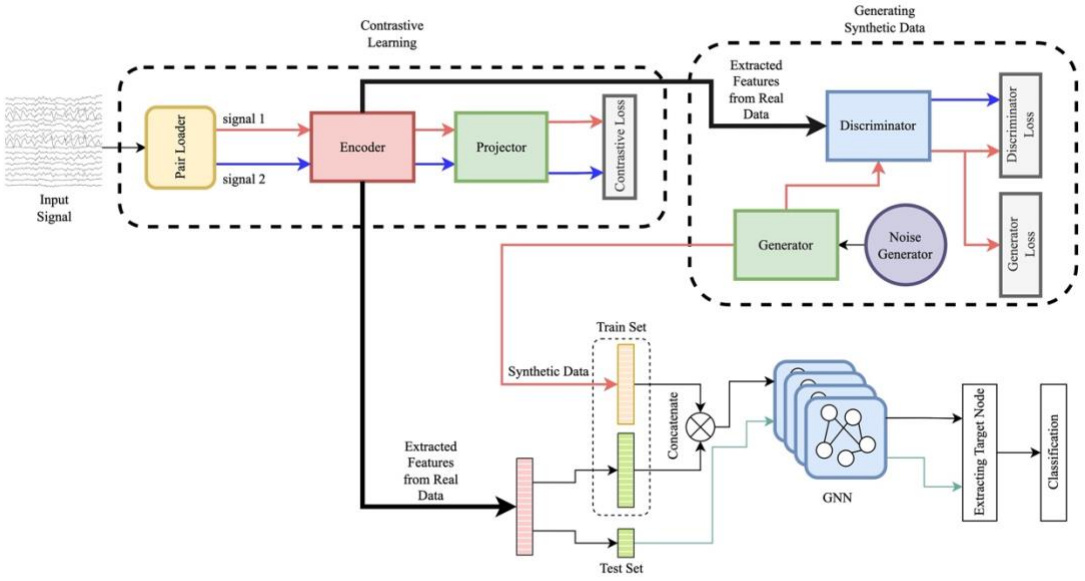


Figure 5.1. Architecture of the proposed model.

5.2.1 CL Component

The overall architecture of the proposed model is presented in Figure 5.1. The input signal is fed to the encoder of the CL component which in turn produces a latent representation. However, prior to deploying this encoder, the CL component must be

trained, and CL loss must be minimized. As described in Section 2.3.4, the encoder is trained using a CL approach. The description of training the CL is provided in the sequel.

CL has already been adopted lately in the context of physiology-based emotion recognition [109, 53, 110]. However, in [109], the focus was mainly on creating augmented instances of each EEG signal to construct positive pairs, without addressing inter-subject and intra-subject emotion variabilities. In contrast, the CL method presented in [53, 110] was applied to limited data, as they did not leverage augmented transforms of the EEG signals, which may cause overfitting issues. Additionally, in [110] the pairing was based on the EEG signals of the same trials over different subjects, neglecting intra-subject emotion variabilities for trials pertaining to a subject and evoking the same emotional state.

The CL component has a dual role. It acts as an encoder to extract high-quality features from the raw EEG signals and minimizes inter-subject and intra-subject emotion variabilities. Figure 2.5 (which is also presented in Figure 5.1 by a dashed line encapsulating the inner components involved in CL) shows that the first step for training the CL component is to load signal pairs. To train the CL component, we propose a pairing mechanism that leverages trial categorization based on the valence-arousal emotional model. By doing so, we aim to maximize the representation similarity across EEG signals corresponding to the same emotional state, regardless of the subject or trial number. Therefore, in our CL strategy, the model learns to recognize whether two EEG signals correspond to the same emotional state.

To achieve our goal, we categorized the EEG data into four emotional categories: High Valence/High Arousal (HVHA), High Valence/Low Arousal (HVLA), Low Valence/High Arousal (LVHA), and Low Valence/Low Arousal (LVLA). Then, we employ a pair loader

to create the mini-batches we use for training. We describe the pairing mechanism for the DEAP dataset below. The same approach was used for the MAHNOB-HCI dataset.

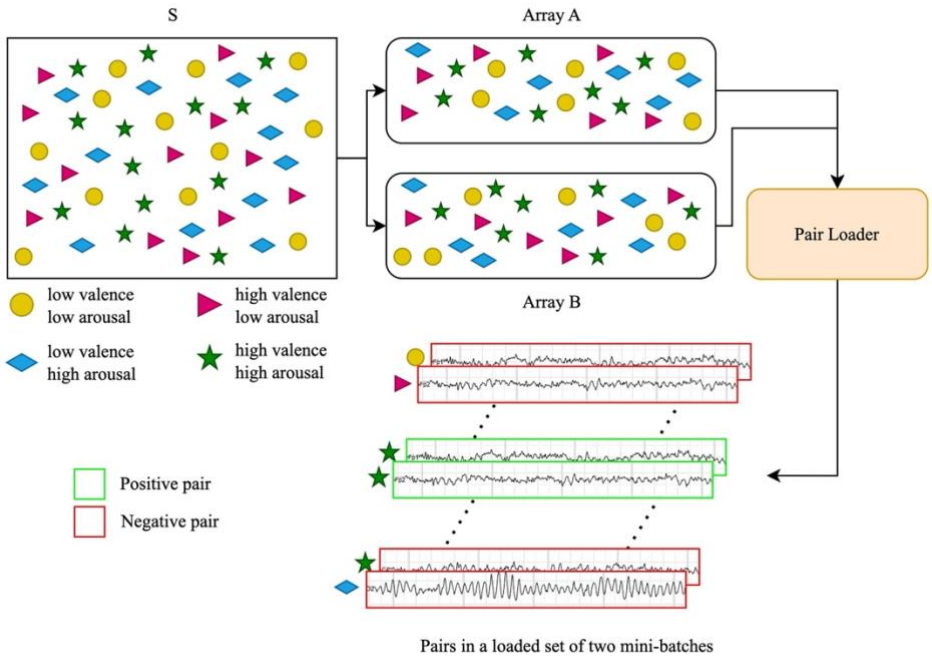


Figure 5.2. Demonstration of the proposed pairing model based on emotional categories.

We flattened all EEG signals as:

$$S = \{s_{1,1,1}, s_{1,1,2}, \ldots, s_{i,j,k}\} \tag{5.1}$$

Where $s \in \mathbb{R}^{7680}$ represents each DEAP EEG signal with 7680 samples. i represents the subject number ($1 \leq i \leq 32$), j represents the trial number ($1 \leq j \leq 40$) and k represents EEG channel number ($1 \leq k \leq 32$). We created 20s segments from the pre-processed EEG signals as S becomes S' where $s' \in \mathbb{R}^{2560\times 3}$ to prepare the data to be fed to the channel encoder.

Then, we evenly divided the EEG signals across two arrays, A and B (Figure 5.2) as follows:

$$A = \{a_1, a_2, \ldots, a_n\}$$
$$B = \{b_1, b_2, \ldots, b_n\}$$
$$Count(A) = Count(B) \tag{5.2}$$
$$Count(A_{xy}) = Count(B_{xy})$$

Where a_n and b_n represent signals in arrays *A* and *B*, respectively ($1 \leq n \leq 16380$). *Count* is a function that returns the array length. Also, x and y represent valence and arousal labels, respectively. Where $x, y \in \{l, h\}$ and *l* and *h* represent the low and high class for each label.

We used a pair loader to create mini-batches for training. The mini-batches contained pair samples, with each signal in the pair originating from a different array. The signals in the mini-batches were pulled randomly from the arrays as follows:

$$pair\_loader(l) \rightarrow A_{minibatch}, B_{minibatch} \tag{5.3}$$

Where $A_{minibatch}$ and $B_{minibatch}$ are mini-batches of size l that are loaded from the *A* and *B* arrays, respectively. Each signal from $A_{minibatch}$ is paired with all the signals in $B_{minibatch}$. Therefore, the total number of pairs formed using both mini-batches is $l \times l$. We considered a pair as positive if both of its signals had the same label, and we considered it as negative otherwise. Positive and negative pairs are labeled as 1 and 0, respectively.

As described in Section 2.3.4, the CL component we use consists of four sub-components. Therefore, when pairs are loaded, the EEG signals of any typical positive pair of a_i and b_j are passed to the channel encoder of the CL component to generate high-quality inter-subject/intra-subject aligned representations for the EEG signals over trials with the same emotional state. The output of the encoder is forwarded to a simple

multilayer perceptron channel projector. As it was found useful in [100], we apply the contrastive loss on the output of the channel projector. As in the contrastive loss function used in [97, 103], we deploy the normalized temperature-scaled cross-entropy loss which has been modified according to our proposed pairing mechanism. The loss attempts to increase the similarity between the two EEG signals of a positive pair. The contrastive loss function for our proposed pairing mechanism is defined as follows:

$$loss(A_{minibatch}, i) = -\sum_{j} log \frac{exp\left(sim\left(a'_i, b'_j\right)/\tau\right)}{1 + \sum_{k=1}^{l} \mathbb{I}(a_i, b_k)\, exp(sim(a'_i, b'_k)/\tau)} \tag{5.4}$$

Where l is the total number of pairs that can be constructed with an EEG signal a_i of $A_{minibatch}$, j is the indices of signals in $B_{minibatch}$ which constructs a positive pair with a_i, τ is the temperature parameter to adjust the scaling of the similarity scores, and a'_i and b'_j are the output of the channel projector in response to signals a_i and b_j for a positive pair derived from:

$$a'_i = projector(encoder(a_i)) \tag{5.5}$$

$sim(a, b)$ is the cosine similarity of a and b which is calculated as follows:

$$sim(a, b) = \frac{a.b}{\| a \| \| b \|} \tag{5.6}$$

Where, $\| a \|$ and $\| b \|$ are the Euclidean norms of a and b, respectively. $\mathbb{I}(a_i, b_k) \in \{0,1\}$ which is set as 1 if a_i and b_k makes a negative pair otherwise, it is 0.

The final loss of the set of two mini-batches is computed as follows:

$$L = \sum_{i=1}^{l} \frac{Loss(A_{minibatch}, i) + Loss(B_{minibatch}, i)}{2} \tag{5.7}$$

Our channel encoder and channel projector network architectures are inspired by [103] and presented in Table 5.2 and Table 5.3, respectively. The hyperparameters of these networks were further manipulated and the networks with the presented parameters achieved the top performance. k, f, and s refer to kernel size, filter size, and stride number, respectively.

For the CL training, we used a mini-batch size of 30 and trained the model for 8 epochs on the training set. Algorithm 1 presents the pseudo-code for the proposed CL component.

After completing the CL training, we integrated the trained encoder of the CL component into the overall solution presented in Figure 5.1.

5.2.2 GAN Component

As mentioned earlier in Section 2.3.5 and 3.1.3, due to the small size of the dataset which causes overfitting issues in the training process, we increase the number of data samples using GAN to generate synthetic realistic-like data for efficient emotion classification.

Therefore, as depicted in Figure 5.1, the features extracted from the CL's channel encoder are considered real data and fed to a GAN to generate the synthetic data.

Table 5.2. Architecture of the channel encoder

Layer	Parameter	Activation
Conv1D	k=20, f=100, s=2	ReLU
Batch Norm	—	—
Conv1D	k=10, f=90, s=2	ReLU
Batch Norm	—	—
Conv1D	k=5, f=50, s=2	ReLU
Batch Norm	—	—
Conv1D	k=3, f=10, s=2	ReLU
Batch Norm	—	—
Dense	70	Sigmoid
Batch Norm	—	—

Table 5.3. The architecture of the channel projector

Layers	Parameter	Activation
Dense	100	ReLU
Dense	10	—

Algorithm 1: Contrastive Learning Algorithm

Inputs: Training data $\{A, B\}$, the learning rate α, the minibatch size l, the training epochs T

1: initialize parameters of the base encoder θ_e and the projector θ_p
2: for epoch = 1 to T do
3: repeat
4: sample l signals from A and B
5: obtain $\{sim_{i,j} | i, j = 1,2, \dots, l\}$ by (5.6)
6: calculate loss by (5.7)
7: update θ_e and θ_p by loss with α
8: until all possible pairs enumerated

Outputs: Features of data using parameters θ_e

The synthetic data and real data are merged to form the expanded training dataset to further train the GNN model for emotion classification.

Equations (5.8) and (5.9) show the discriminator and generator's loss functions, respectively.

$$loss_D = \nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[logD(x^i) + log\left(1 - D\left(G(z^i)\right)\right)\right] \tag{5.8}$$

$$loss_G = \nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} log\left(1 - D\left(G(z^i)\right)\right) \tag{5.9}$$

Where G is a generator, D is a discriminator, $z^i \in R^{1\times 256}$ is ith sample noise vector with i representing the trial number, $x^i \in R^{1\times 32\times 70}$ is our ith input data which is aimed to be reconstructed. θ_d is the parameter set for the discriminator, and θ_g is the parameter set for the generator.

We utilized the GAN component to produce four distinct sets of labeled synthetic data representing high valence, low valence, high arousal, and low arousal.

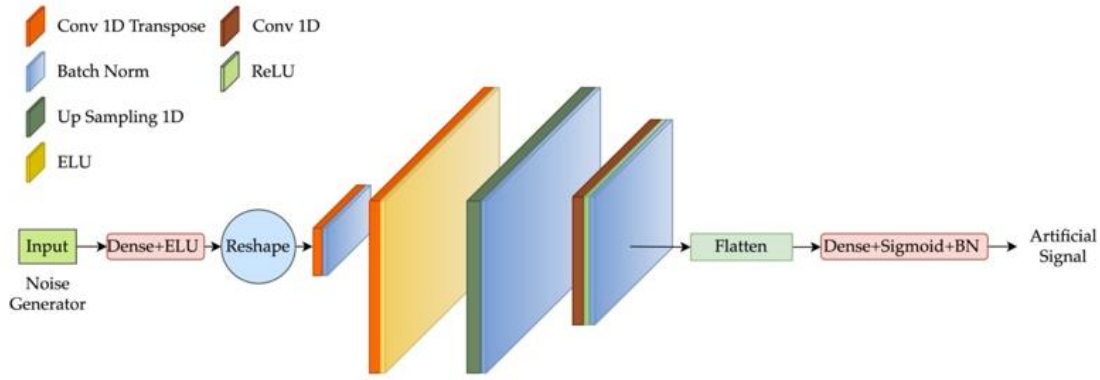
To train the GAN component, we used 100, 100, and 256 as batch size, number of epochs, and code size, respectively. Algorithm 2 illustrates the pseudo-code of the proposed GAN network.

Algorithm 2: GAN Algorithm

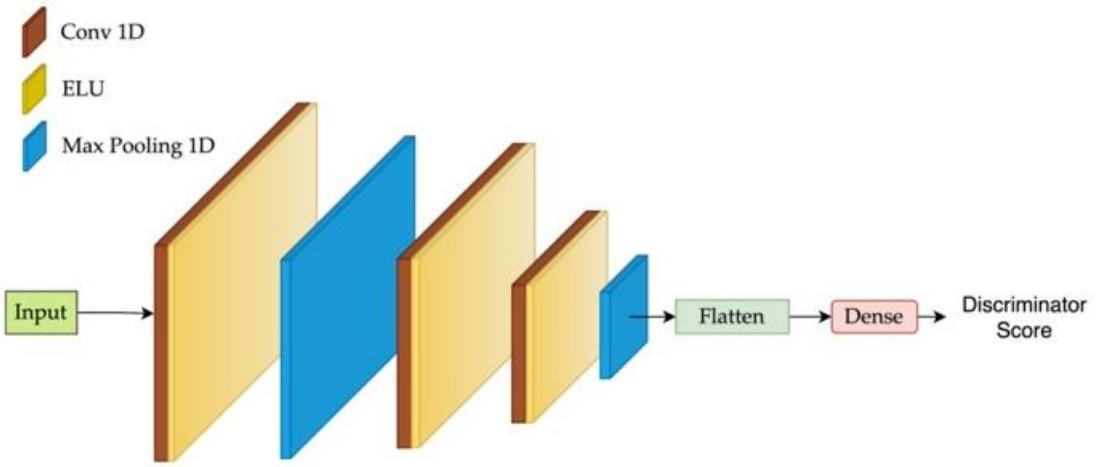
Inputs: Features of training trial's data {X}, the learning rate α_g and α_d, the Batch size N, the training epochs T

1: initialize parameters of the base generator θ_g and the discriminator θ_d
2: for epoch = 1 to T do
3: for iterate = 1 to 5 do
4: sample from trials $\{x_i | i = 1, \dots, N\}$
5: sample from noise generator $\{z_i | i = 1, \dots, N\}$
6: calculate loss by (5.8)
7: update θ_d by loss with α_d learning rate
6: sample from noise generator $\{z_i | i = 1, \dots, N\}$
7: calculate loss by (5.9)
8: update θ_g by loss with α_g learning rate
9: generate synthetic data using θ_g

Outputs: synthetic realistic-like data



a) Architecture of the generator which generates realistic-like data.



b) Architecture of the discriminator which distinguishes the real and artificial data.

Figure 5.3. Architecture of the GAN network.

After generating synthetic data using GAN, the synthetic data is appended to the real data obtained from the channel encoder of CL. Then, we conducted two rounds of training and classification of the GNN, the first time for the valence and the second time for the arousal dimension. We used the DEAP and MAHNOB-HCI databases for this study with 1280 and 530 total number of trials, respectively.

As in [118], the discriminator parameters were updated 5 times per epoch, while the generator parameters were updated once per epoch.

The generator and discriminator networks used in this study were optimized through a trial-and-error process. Table 5.4 and Table 5.5 show the deployed architectures for the generator and discriminator networks, respectively. The architecture of the GAN component is illustrated in Figure 5.3.

Table 5.4. The architecture of the generator network

Layer	Parameters	Activation Function
Dense	768	ELU
Reshape	(3, 256)	—
Conv1D Transpose	k=5, f=40, s=1	—
Batch Norm (BN)	—	—
Conv1D Transpose	k=5, f=40, s=1	ELU
UpSampling1D	k=2	—
Conv1D Transpose	k=5, f=35, s=1	—
Batch Norm	—	—
Conv1D	k=5, f=32, s=1	ReLU
Batch Norm	—	—
Flatten	—	—
Dense	2240	Sigmoid
Reshape	(32, 70)	—
Batch Norm	—	—

Table 5.5. Architecture of the discriminator network

Layer	Parameters	Activation Function
Conv1D	k=5, f=50, s=1	ELU
MaxPool1D	k=2	—
Conv1D	k=5, f=50, s=1	ELU
Conv1D	K=5, f=40, s=1	ELU
MaxPool1D	k=3	—
Flatten	—	—
Dense	1	—

5.2.3 GNN Component

The appended data is fed to the GNN component for classification. We propose a GNN with multiple linear layers. Each layer is composed of a dense layer, sigmoid activation function. We also deploy the dropout layer on the first and before the last dense layer. We also applied regularization over the network.

First, we setup a dynamic square adjacency matrix with a dimension corresponding to the number of EEG channels (32 for both DEAP and MAHNOB-HCI databases). The following equation describes each layer's forward propagation:

$$output = WX_{i,j} + b, \qquad W \in R^{u \times l}$$
$$X_{i,j} = S_{i,j}A, \qquad S_{i,j} \in R^{l \times c}, A \in R^{c \times c} \tag{5.10}$$

Where $S_{i,j}$ is the extracted features corresponding to the ith subject and jth experiment. W and b are the parameters of the layer. A is the adjacency matrix where the ith row and jth column define the relation between the ith node and jth one. c is the number of

Table 5.6. Architecture of the proposed GNN

Layer	Number of hidden units	Activation Function	Dropout rate
Dense	45	Sigmoid	
Dropout	—	—	22%
Dense	26	Sigmoid	
Dense	7	Sigmoid	
Dropout	—	—	5%
Dense	2	Softmax	

channels, l is the length of extracted features, and u is the number of hidden units of the corresponding linear layer.

The architecture of our proposed GNN is described in Table 5.6. The hyperparameters of the proposed GNN model are optimized using Gaussian search [153]. We adopted the learning rate, dropout, layer's output size (hidden size), and a number of dense layers for hyperparameter optimization. The hyperparameter optimization process rendered the following values: a learning rate of 0.00005, 4 dense layers with hidden sizes of [45, 26, 7, 2], respectively, and a dropout rate of 22% and 5% as in Table 5.6.

Algorithm 3 presents the pseudo-code of the proposed GNN.

We introduced a new parameter to our network, which we call the "target node", to update the weights during backpropagation. The model is fed input features of a target node and its neighboring nodes and is tasked with predicting the target output value. Target node optimization in a graph neural network involves training a model to predict the properties or behaviors of specific target nodes in a graph. Node-level backpropagation attempts to categorize nodes into several classes, which can improve performance. In our work, we

Algorithm 3: GNN algorithm

Inputs: Features of training trial's data {X}, Features of artificial trial's data {F}, the learning rate α_c, the batch size N, the training epochs T, the number of channels C

1: initialize parameters of the GNN θ_{gnn} and the adjacency matrix $\{A_{i,j}|i,j=1,\dots,C\}$
2: {S} ← merge {X} and {F}
3: for epoch = 1 to T do
4: repeat
5: sample from trials $\{s_i|i=1,\dots,N\}$
6: calculate the output of all layers by (5.10)
7: get the p-value using (5.11)
8: calculate loss with p-value
9: update θ_{gnn} parameters and $A_{i,j}$ by loss with α_c
10: until all possible batches enumerated

Outputs: θ_{gnn} and $A_{i,j}$ parameters

considered the number of nodes to be equal to the number of EEG channels in each trial, which is 32. We used grid search to optimize this parameter before performing hyperparameter optimization. Hence, from eq. (5.10) we have:

$$out \in R^{b\times 32\times 2}$$
$$p = out_t \quad 1 \le t \le 32, out_t \in R^{b\times 2} \tag{5.11}$$

Where out is the last output from our GNN. b is the batch size of the processed data. t refers to the target node and p denotes the output processed data of the corresponding node which is used further for the calculation of loss.

For the loss function and optimizer, we used Categorical Cross Entropy loss [154] and Adam optimizer [43]. The batch size and number of epochs are 100 and 400, respectively. However, we used early stopping to avoid overfitting.

5.3 Results and Analysis

This section describes the experimental setup and evaluation of the proposed model using two testing scenarios. Firstly, we performed an ablation study to analyze the impact of each component (CL, GAN, and GNN) of the proposed model on improving emotion classification accuracy. To do so, each component is replaced with a competing similar component from the literature or omitted entirely. Secondly, we compared the performance of the proposed model with that of several recent competing emotion recognition models. We implemented the existing models and trained and tested them on the same datasets for a fair comparison. All the models were implemented using the Keras framework libraries with a Tensorflow backend in Python and trained on NVIDIA GeForce RTX 2080 Ti GPU.

5.3.3 Experimental Setup

We performed two sets of evaluations. The first evaluation involved the partitioning of the dataset into training and testing subsets. The training set is used to train both the proposed model and existing models (for comparison purposes), computing cross-entropy loss and updating model parameters using the Adam optimizer [43]. The testing set is used to evaluate the trained model's ability to identify the level of arousal and valence of the testing samples.

To split a dataset into training and testing subsets, we shuffled all trials for different subjects and separated 20% of the data as testing samples while using the remaining 80% for training. We used 1024 trials as training datasets and 256 trials as the testing dataset for the DEAP database. Similarly, for the MAHNOB-HCI database, we used 424 and 106 trials as training and testing datasets, respectively. Moreover, we performed four-fold cross-

validation on the training set and reported the average accuracy of the four folds as the training accuracy.

For the second evaluation, we assess the performance of the proposed model in a subject-independent manner where the data of one subject was excluded as a testing dataset and the data for the remaining subjects (31 and 26 for DEAP and MAHNOB-HCI, respectively) were used for training and validating of the proposed model using the LOSOCV evaluation strategy. We reported the average accuracy of the folds (31 folds and 26 folds for DEAP and MAHHNOB-HCI datasets, respectively) as the training accuracy. This test provides an unbiased estimate of the model performance for individual subjects since each subject serves as a test set in LOSOCV.

For both evaluations, when comparing to existing state-of-the-art methods, we detail the accuracy of the training and testing subsets in the case of valence and arousal. Additionally, we provide average accuracy across both emotional dimensions for the training and testing datasets.

5.3.4 First Evaluation Strategy: Splitting Dataset into Training and Testing Sets

In this section, we evaluate the proposed model by splitting the dataset into training and testing subsets. We start with an ablation study (Section 5.3.4.1) and then proceed to compare the proposed model to state-of-the-art models on the same datasets (Section 5.3.4.2).

5.3.4.1 Component-Based Analysis (Ablation Study)

Performance of the GAN Component

In this section, we evaluate the performance of the GAN component of the proposed model based on the amount of generated data added to the training set for both the DEAP and MAHNOB-HCI databases.

Table 5.7 and Table 5.8. present the performance of the proposed model for different amounts of synthetically generated data appended to the training sets for the two databases. The amount of appended synthetic data that results in the best performance is used for further classification. The training and testing accuracies of the valence and arousal emotional dimensions are used to determine the amount of synthetic data to be appended to the training sets. The amount of appended data is expressed as a multiple of the real training set. For instance, 0 denotes that we will use only the original real training dataset without any synthetic data added. On the other hand, 0.5 indicates that we will add synthetic data that corresponds to half the size of the original real set. Specifically, we would add 512 synthetic trials for the DEAP database and 212 synthetic trials for the MAHNOB database. The result for the best node is chosen to represent in the table. The results for the rest of the nodes are provided in Appendix A.

From the tables, we can see that adding a number of synthetic samples equal to that of the real ones (i.e., the amount of appended synthetic data is equal to 1) achieves the best results for the DEAP database while adding synthetic samples that correspond to half the number of real samples (i.e., amount of appended synthetic data is equal to 0.5) performs best for the MAHNOB database based on the accuracies obtained for the valence and arousal classification tasks. However, both tables show that employing synthetic data still

Table 5.7. Performance evaluation on the volume of appended synthetically generated EEG data in the DEAP dataset

Data Appended	Valence Acc. (%)		Arousal Acc. (%)	
	Train	Test	Train	Test
× 0 dataset (0)	66.01	59.37	69.33	62.10
× 0.5 dataset (512)	72.39	62.10	70.89	63.67
× 1 dataset (1024)	**74.65**	**64.84**	**74.21**	**66.40**
× 1.5 dataset (1536)	76.71	64.06	73.63	64.84
× 2 dataset (2048)	77.53	64.84	74.27	64.45

Table 5.8. Performance evaluation on the volume of appended synthetically generated EEG data in the MAHNOB-HCI dataset

Data Appended	Valence Acc. (%)		Arousal Acc. (%)	
	Train	Test	Train	Test
× 0 dataset (0)	72.68	66.03	73.15	61.32
× 0.5 dataset (212)	**74.32**	**66.98**	**78.12**	**71.69**
× 1 dataset (424)	72.92	66.03	71.25	62.26
× 1.5 dataset (636)	74.32	63.20	76.42	67.92
× 2 dataset (848)	77.88	64.15	74.50	64.15

improves the performance of the model. For further analysis, we selected 1 and 0.5 times as the amount of appended synthetic data for the DEAP and MAHNOB-HCI databases,

respectively, as they achieve the best accuracies on both emotional dimensions and the least discrepancy between the train and test datasets.

Performance of the CL Component

In the next step, we compared the performance of the encoder from the proposed CL component to other feature extractors, namely SeqCLR Encoder [109] and CLISA Encoder [110], which are encoders trained using CL for EEG-based classification, Label-Based CL Encoder [53], which is an encoder trained using CL for multimodal classification, Attention Encoder [99], which is an encoder with an attention mechanism for EEG-based emotion classification, and VGG16 [89], which is a widely used CNN for feature extraction employed in various applications. To evaluate the performance of a feature extractor, we replace the proposed encoder in our model (shown in Figure 5.1) with the feature extractor. Then, we train and test the resulting model on the DEAP dataset. Our comparison results are presented in Table 5.9.

Table 5.9. Performance evaluation on different CL pairing methods and feature extractors on the DEAP dataset

Data Appended	Valence Acc. (%)		Arousal Acc. (%)	
	Train	Test	Train	Test
SeqCLR Encoder [109]	87.84	64.06	69.33	64.06
Label-Based CL Encoder [53]	80.95	60.54	74.41	62.89
CLISA Encoder [110]	65.11	61.66	75.96	65.41
VGG16 [89]	90.86	60.15	92.52	62.10
Attention-Encoder [99]	74.65	59.76	76.26	61.71
Proposed CL component Encoder	**74.65**	**64.84**	**74.21**	**66.40**

As shown in Table 5.9, the proposed CL component encoder outperforms. We evaluated the use of VGG16 [89] for feature extraction from EEG signals. Although VGG16 is a deep convolutional neural network typically used for image feature extraction, it has been employed for other feature extraction applications as well. Based on our results, we observed that using the VGG16 feature extractor caused the model to overfit, leading to high performance on the training dataset but relatively low performance on the testing dataset. The Attention Encoder proposed in [99] performs relatively well, but our proposed CL component encoder shows approximately a 5% improvement in testing accuracies for the valence and arousal emotion classification. The CLISA Encoder [110] achieves more reliable results than other existing models, but our proposed CL component encoder achieves higher testing accuracies for both valence and arousal classification. The SeqCLR Encoder [109] and Label-Based CL Encoder [53] show a relatively high discrepancy between training and testing accuracies, which might be due to overfitting.

Performance of the GNN

In the last step, we investigated the effect of using GNN as a classifier by comparing it with some benchmark classification models presented in Table 5.10. All evaluated classifiers have been proposed for EEG-based emotion classification in recent work [76, 87, 98, 117]. To evaluate the performance of a classifier, we replace the proposed GNN classifier in our model (shown in Figure 5.1) with the classifier. Then, we train and test the resulting model on the DEAP dataset.

As shown in Table 5.10, the choice of classifier has a significant impact on the emotion classification accuracy. Our proposed classifier outperformed the existing popular classification models. The training and testing accuracies for the valence and arousal

Table 5.10. Performance evaluation on different classification models on the DEAP dataset

Model	Valence Acc. (%)		Arousal Acc. (%)	
	Train	Test	Train	Test
DNN [98]	99.99	52.73	79.79	48.44
CNN [87]	98.73	52.34	74.22	53.52
BiLSTM [76]	95.07	56.25	84.64	60.16
SVM [117]	84.76	50.78	66.16	59.76
The proposed GNN	**74.65**	**64.84**	**74.21**	**66.40**

dimensions have a high discrepancy for DNN [98], CNN [87], and BiLSTM [76] models which indicate that the model overfits the training set. SVM [117] performs more reliably on the arousal dimension but it performs poorly on the testing dataset on both arousal and valence dimensions. BiLSTM achieves better testing accuracies in comparison to DNN [98] and CNN [87]. Nonetheless our proposed classifier still showed a significant improvement over the other methods.

5.3.4.2 Performance Comparison

In this section, we compare our proposed model with several recent emotion classification models on both the DEAP and MAHNOB-HCI datasets. The DEAP dataset is particularly valuable as it has the highest number of participants compared to other publicly available EEG datasets for emotion recognition.

We compared our proposed model with existing benchmark EEG-based deep models, namely:

- Attention-based LSTM combined with domain discriminator denoted as ATDD-LSTM where DE features from different frequency bands are used as input [95]
- Attention-based convolutional recurrent neural network (ACRNN) using EEG features of time and frequency domains as input [88]
- Graph convolutional neural network combined with LSTM (ECLGCNN) with DE features as input [18]
- Conditional Wasserstein GAN (CWGAN) using DE features as input [117]
- CapsNet with attention mechanism (ACapsNet) using segmentations of raw EEG signals as input [145]
- NAS-optimized emotion recognition model with raw EEG signals as input [142]
- Multi-task learning using DF with a manual 2D frame for each sample according to EEG channels' distribution as input [85]
- Attention-based LSTM autoencoder (ALSTM autoencoder) + attention-based CNN (ACNN) using raw EEG signals with additive white Gaussian noise as input [99]
- Hierarchical Spatial Learning Transformer (HSLT) model with DSP features as input [140]
- EEG emotion Transformer model (EeT) using segmentations of raw EEG signals as input [143]
- Contrastive learning followed by three dense layers for emotion recognition (CLISA) using EEG signals as input [110].

Table 5.11 shows the emotion recognition accuracies on the training and testing sets for the proposed and existing competing models on the valence and arousal for the DEAP database. Our model outperforms competing EEG-based emotion recognition models with significant improvement. However, the models presented in [85], [88], [99], [110], [140], [142], [143] and [145] achieve a relatively high training accuracy but present a significant discrepancy between training and testing accuracies which may be due to overfitting on the training set. The models presented in [18], [95], and [117] achieve the least discrepancy between the training and testing accuracies but overall, these three models perform poorly for emotion classification compared to the proposed model. Specifically, the GAN-based model presented in [117] achieves the highest train and test accuracy for emotion classification outperforming the ten state-of-the-art models while maintaining a moderate level of discrepancy. This finding provides a valuable understanding of the impact of expanding datasets in case of depletion of deep learning models. It suggests that by incorporating a larger and more diverse dataset, there is a potential enhancement in the model's performance. The proposed model achieves a training accuracy of 74.65% and 74.21% for valence and arousal in the DEAP dataset, respectively. The same model achieves a testing accuracy of 64.84% and 66.40% for valence and arousal in the DEAP dataset, respectively. This improvement also shows that the EEG representation

Table 5.11. Comparison of emotion recognition models on DEAP database

Model	Valence Accuracy (%)		Arousal Accuracy (%)		Average Accuracy for Valence and Arousal (%)	
	Train	Test	Train	Test	Train	Test
CLISA [110]	90.72	50.59	85.11	58.79	87.91	54.69
ACRNN [88]	68.26	45.31	79.00	55.47	73.63	50.39
ECLGCNN [18]	56.64	56.25	58.50	60.55	57.57	58.45
ATDD LSTM [95]	54.93	51.12	59.37	55.76	57.15	53.44
CWGAN [117]	69.92	61.71	69.43	70.31	69.67	66.01
ACapsNet [145]	99.90	51.56	96.19	50.39	98.04	50.97
NAS [142]	99.31	52.73	98.24	48.04	98.77	50.38
DF [85]	96.87	53.51	95.01	55.46	95.94	54.48
ALSTM autoencoder + ACNN [99]	99.80	54.68	99.12	57.42	99.46	56.05
HSLT Transformer [140]	73.14	42.18	71.48	54.29	72.31	48.23
EeT Transformer [143]	88.96	48.43	87.10	53.51	88.03	50.97
Proposed model	**74.65**	**64.84**	**74.21**	**66.40**	**74.43**	**65.62**

Table 5.12. Comparison of emotion recognition models on MAHNOB-HCI database

Model	Valence Accuracy (%)		Arousal Accuracy (%)		Average Accuracy for Valence and Arousal (%)	
	Train	Test	Train	Test	Train	Test
CLISA [110]	54.15	50.00	61.26	60.37	57.70	55.18
ACRNN [88]	84.60	45.28	83.41	61.32	74.00	53.30
ECLGCNN [18]	99.76	52.83	99.29	41.50	99.52	47.16
ATDD LSTM [95]	99.53	37.73	99.82	50.00	99.67	43.86
CWGAN [117]	66.98	66.03	66.50	67.92	66.74	66.97
ACapsNet [145]	99.52	37.73	97.64	34.90	98.58	36.31
NAS [142]	99.76	45.28	99.29	45.28	99.52	45.28
DF [85]	91.74	56.60	90.80	53.77	91.27	55.18
ALSTM autoencoder + ACNN [99]	91.98	45.28	95.28	42.45	93.63	43.86
HSLT Transformer [140]	68.39	53.77	70.28	50.00	69.33	51.88
EeT Transformer [143]	94.10	47.16	92.45	49.05	93.27	48.10
Proposed model	**74.32**	**66.98**	**78.12**	**71.69**	76.22	69.33

learning by our contrastive learning component is more efficient than simple DE features for emotion recognition classification.

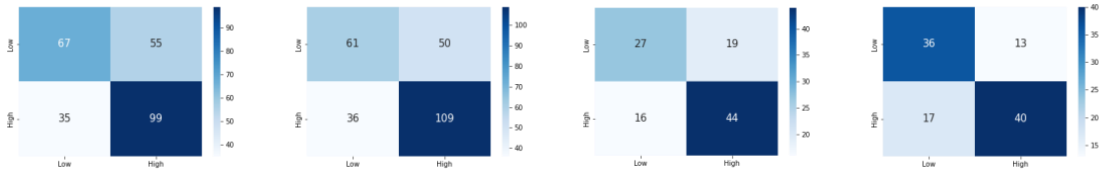
We have also calculated the average accuracy over valence and arousal emotional dimensions across all models to assess their performance relative to the proposed model. As can be seen from Table 5.11, the proposed model achieves a higher average emotion recognition accuracy on the test dataset with less discrepancy between the average

accuracies of the train and test datasets. The model presented in [117] demonstrates a 0.39% increase in average test accuracies on both emotion dimensions compared to the proposed model. However, the proposed model achieves 4.76% higher average train accuracies on both valence and arousal emotion dimensions compared to the model presented in [117]. The model presented in [18] achieves a moderate average accuracy for the test dataset compared to the other nine benchmark models which is 7.17% lower than the average test accuracy of the proposed model. However, this model results in a low average train accuracy of 57.57%.

We evaluated the effectiveness of our proposed model using the MAHNOB-HCI database. Table 5.12 presents the training and testing accuracies of the compared models for the MAHNOB-HCI database on the valence and arousal emotion dimensions. The results indicate that the proposed model outperforms the competing recent models with 74.32% and 78.12% as training accuracies and, 66.98%, and 71.69% as testing accuracies for valence and arousal, respectively. Despite achieving a moderate level of discrepancy between the training and testing accuracies, the model presented in [110] performs relatively poorly on emotion classification compared to our proposed model.

As can be seen from Table 5.12, the average train and test accuracies for the proposed model are relatively higher compared to state-of-the-art models, coupled with a moderate discrepancy between these accuracies. However, the GAN-based model presented in [117] shows the lowest discrepancy between average train and test accuracies compared to the proposed model. Despite this, the proposed model surpasses the performance of the GAN-based model in [117] for the training and testing sets for both emotional dimensions assessed.

The average test and train accuracies over valence and arousal for the proposed model are 69.33% and 76.22 % with a 6.89% drop which is significantly lower than the difference



(a) (b) (c) (d)

Figure 5.4. Confusion matrices of the proposed model.

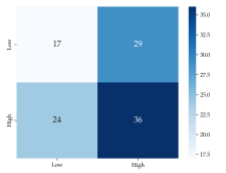
(a) Valence-DEAP dataset (b) Arousal-DEAP dataset (c) Valence-MAHNOB dataset (d) Arousal-MAHNOB dataset.

between the average train and test accuracies over the benchmark emotion recognition models. This indicates that the model has learned the relevant patterns from the training data without memorizing and consequently, it can generalize well to new unseen data.

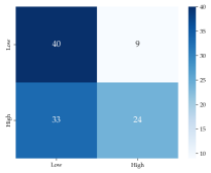
We present the confusion matrices for the test set for the proposed model for both databases in Figure 5.4. The proposed model achieved a better performance in the classification of high arousal and high valence for the DEAP dataset (73.88% and 80.74%, respectively) compared to the MAHNOB-HCI dataset (73.33% and 70.17%, respectively), which could be due to the test set in the DEAP dataset being slightly unbalanced in favor of high valence and high arousal.

We present the confusion matrices for the existing models for the MAHNOB-HCI database in Figure 5.5. We observe that the values on the diagonal of the proposed model's confusion matrices (Figure 5.4 (c) and (d)) are relatively greater than those on the diagonal of the other confusion matrices. This observation suggests that our proposed model has lower misclassification rates compared to the competing models. However, for the arousal

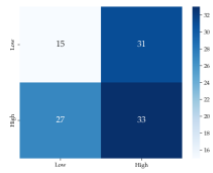
classification, CLISA [110], ACRNN [88] and ATDD-LSTM [95] outperformed our



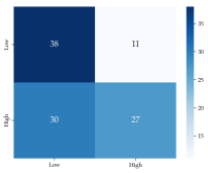
(a)



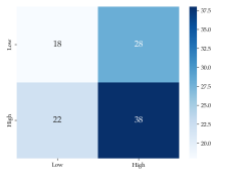
(b)



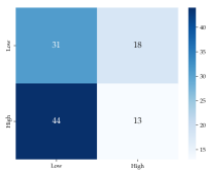
(c)



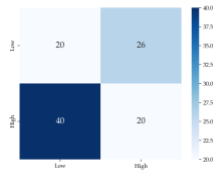
(d)



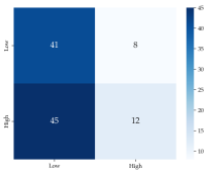
(e)



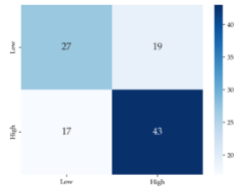
(f)



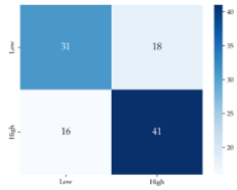
(g)



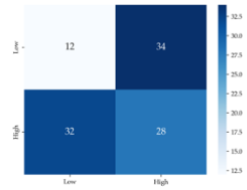
(h)



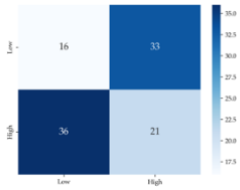
(i)



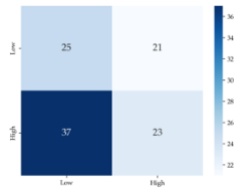
(j)



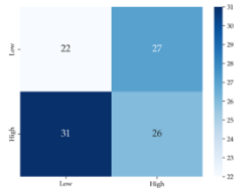
(k)



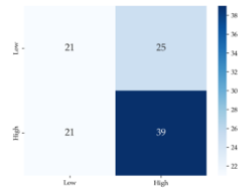
(l)



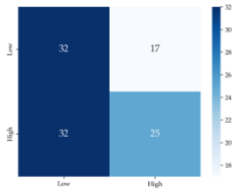
(m)



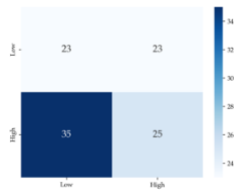
(n)



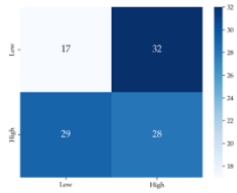
(o)



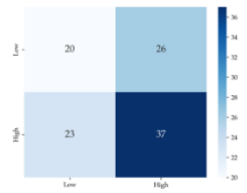
(p)



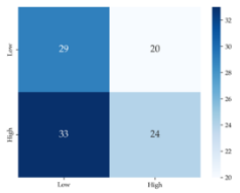
(q)



(r)



(s)



(t)

(u) (v)

Figure 5.5. Confusion matrices of test cases.

(a) Valence- CLISA (b) Arousal- CLISA (c) Valence-ACRNN (d) Arousal-ACRNN (e) Valence- ECLGCNN (f) Arousal- ECLGCNN (g) Valence- ATDD LSTM (h) Arousal-ATDD LSTM (i) Valence-CWGAN (j) Arousal-CWGAN (k) Valence-ACapsNet (l) Arousal-ACapsNet (m) Valence-NAS (n) Arousal-NAS (o) Valence-DF (p) Arousal-DF (q) Valence-ALSTM autoencoder + ACNN (r) Arousal-ALSTM autoencoder + ACNN (s) Valence-HSLT Transformer (t) Arousal-HSLT Transformer (u) Valence-EeT Transformer (v) Arousal-EeT Transformer.

proposed model in classifying low arousal with an 8%, 4%, and 10% higher accuracy, respectively, and CWGAN [117] outperformed our proposed model in classifying high arousal with a 1% higher accuracy. Nevertheless, our proposed model still outperforms the other models on both the arousal and valence emotion classification.

5.3.5 Second Evaluation: Leave-One-Subject-Out Cross-Validation

The evaluation results of the proposed model under the LOSOCV evaluation strategy are presented in Table 5.13 and Table 5.14 for the DEAP and MAHNOB-HCI datasets, respectively.

Table 5.13. LOSOCV of the proposed model on the DEAP database

Model	Valence Accuracy (%)		Arousal Accuracy (%)		Average Accuracy for Valence and Arousal (%)	
	Train	Test	Train	Test	Train	Test
CLISA [110]	50.36	50.00	50.00	50.00	50.18	50.00
ACRNN [88]	80.24	45.00	93.87	40.00	87.05	42.50
ECLGCNN [18]	56.77	50.00	58.55	70.00	57.66	60.00
ATDD LSTM [95]	56.21	67.50	59.19	50.00	57.70	58.75
CWGAN [117]	71.29	52.50	70.72	62.50	71.00	57.50
ACapsNet [145]	99.83	45.00	98.62	47.50	99.22	46.25
NAS [142]	99.35	37.50	98.14	55.00	98.74	46.25
DF [85]	96.12	52.50	95.96	52.50	96.04	52.50
ALSTM autoencoder + ACNN [99]	98.87	55.00	99.03	55.00	98.95	55.00
HSLT Transformer [140]	75.88	47.50	73.30	52.50	74.59	50.00
EeT Transformer [143]	87.82	52.50	85.48	50.00	86.65	51.25
Proposed model (LOSOCV)	**68.77**	**57.50**	**69.85**	**70.00**	**69.31**	**63.75**

Table 5.14. LOSOCV of the proposed model on the MAHNOB-HCI database

Model	Valence Accuracy (%)		Arousal Accuracy (%)		Average Accuracy for Valence and Arousal (%)	
	Train	Test	Train	Test	Train	Test
CLISA [110]	49.59	47.50	100.0	52.50	74.79	50.00
ACRNN [88]	88.71	55.00	75.56	45.00	82.13	50.00
ECLGCNN [18]	100.0	30.00	100.0	60.00	100.0	45.00
ATDD LSTM [95]	99.79	45.00	96.10	50.00	97.94	47.50
CWGAN [117]	67.25	60.00	67.05	45.00	67.15	52.50
ACapsNet [145]	99.79	50.00	97.84	40.00	98.81	45.00
NAS [142]	99.60	30.00	97.05	45.00	98.32	37.50
DF [85]	92.15	60.00	92.35	55.00	92.25	57.50
ALSTM autoencoder + ACNN [99]	89.80	40.00	93.33	45.00	91.56	42.50
HSLT Transformer [140]	69.41	60.00	72.15	55.00	70.78	57.50
EeT Transformer [143]	90.39	50.00	86.47	55.00	88.43	52.50
Proposed model (LOSOCV)	71.52	60.00	67.95	55.00	69.73	57.50

As can be seen from Table 5.13 and Table 5.14, the proposed model presents a relatively small discrepancy between the training and testing accuracies on the DEAP and MAHNOB-HCI datasets for both emotional dimensions. This small gap suggests that the proposed model can generalize well to unseen data, thereby avoiding overfitting or underfitting on the training set. However, for the proposed model, test accuracies for

valence on the DEAP dataset is as low as 57.50%, and 55% for arousal on the MAHNOB-HCI dataset. Nonetheless, the existing state-of-the-art models present similar performance limitations on the testing set. This could be attributed to limitations in the size of the test data, as we only included data from one subject as test data, which constitutes less than 4% of the entire dataset size for each dataset (40 and 20 experiments for the DEAP and MAHNOB-HCI datasets, respectively).

We calculated the average accuracies for the testing and training sets across the valence and arousal emotional dimensions to compare the overall performance of the proposed model with the existing models. As demonstrated by Table 5.13 and Table 5.14, the proposed model achieved the highest average test accuracy of 63.75% and 57.50% for the DEAP and MAHNOB-HCI datasets, respectively. In comparison to existing models, the proposed model demonstrates the smallest discrepancy between average test and train accuracies on both datasets. The existing models seem to overfit on the MAHNOB-HCI dataset, possibly due to its limited size. However, the GAN-based model presented in [117] exhibits the most consistent training and testing accuracies across both emotional dimensions compared to the existing benchmark models on both datasets. Despite this, the performance of the proposed model remains superior. The models outlined in references [85], [88], [99], [140], [142], [143] and [145] overfit on the DEAP dataset. Although the models discussed in references [110, 18, 95] do not overfit on the DEAP dataset, our proposed model nonetheless achieves superior accuracy.

Chapter 6. Transformer-based Bimodal Emotion Recognition Model with Fusion Transformer

Transformer-based models have made a significant impact on various fields within artificial intelligence and natural language processing (NLP) and revolutionized the way researchers approach sequence-to-sequence tasks. Hence, we propose our third automated emotion classification solution using Transformers.

The chapter is organized as follows: in Section 6.1, we introduce the datasets and the modalities we employ for this work in addition to the data preprocessing we require for the proposed model. In Section 6.2, we describe the proposed solution. In Section 6.3, we perform experimental analysis and evaluation of the proposed model.

6.1 Datasets and preprocessing

As in the previous chapters, we employ the publicly available datasets DEAP [64] and MAHNOB-HCI [159] for this work. This choice is motivated by the fact that these datasets contain the highest number of subjects compared to other publicly available EEG datasets for emotion recognition. We detect emotions using facial expressions and EEG signals, as these are the most commonly used modalities in the literature for automated emotion recognition [29].

Due to the limited computational resources available to us and the large size of the data, we employ feature extraction before feeding the data to the model. This approach condenses the input data into a more manageable and informative format. In our first solution [30], we evaluated various input feature sets and established the efficacy of

utilizing time domain features for the classification of emotions through EEG recordings. Therefore, we calculate time-domain features as in [30] creating 99 features for each EEG channel over five frequency regions: theta (4-8 Hz), slow alpha (8-10 Hz), alpha (8-12 Hz), beta (12-30 Hz), and gamma (30-45 Hz).

As for video recordings, we disregard traditional handcrafted feature extraction techniques since deep learning features have demonstrated better performance in recent years. However, we employ a strategy to reduce redundant video data as changes in the facial expressions do not often rapidly change for the subjects consuming video content in the datasets. Therefore, the facial videos are down-sampled to 4 Hz first, then from every 8 frames, we only keep one frame for further processing. This results in 30 frames in total for each experiment. Afterward, we use OpenCV [161] to crop the facial image in each frame. The extracted images were resized to 48 by 48 pixels and gray scaled to be then fed to the model as input facial data. This approach serves multiple purposes. Firstly, it significantly reduces the volume of data without sacrificing the essence of the facial expressions. By maintaining one frame from every couple of frames, we can still capture the essential visual changes that signify emotional responses while discarding redundant frames that might not contribute substantially to the emotion recognition process. Furthermore, this strategy is particularly useful for managing computational resources and improving the efficiency of subsequent analyses or models. Since processing and analyzing large amounts of data can be resource-intensive, our approach minimizes the computational burden while retaining the pertinent emotional cues.

Similar to Chapter 4 and Chapter 5, we employ the valence-arousal dimensional model to represent emotion for our work where we associate each emotional dimension with two classes: high and low with a 5 rating as the boundary.

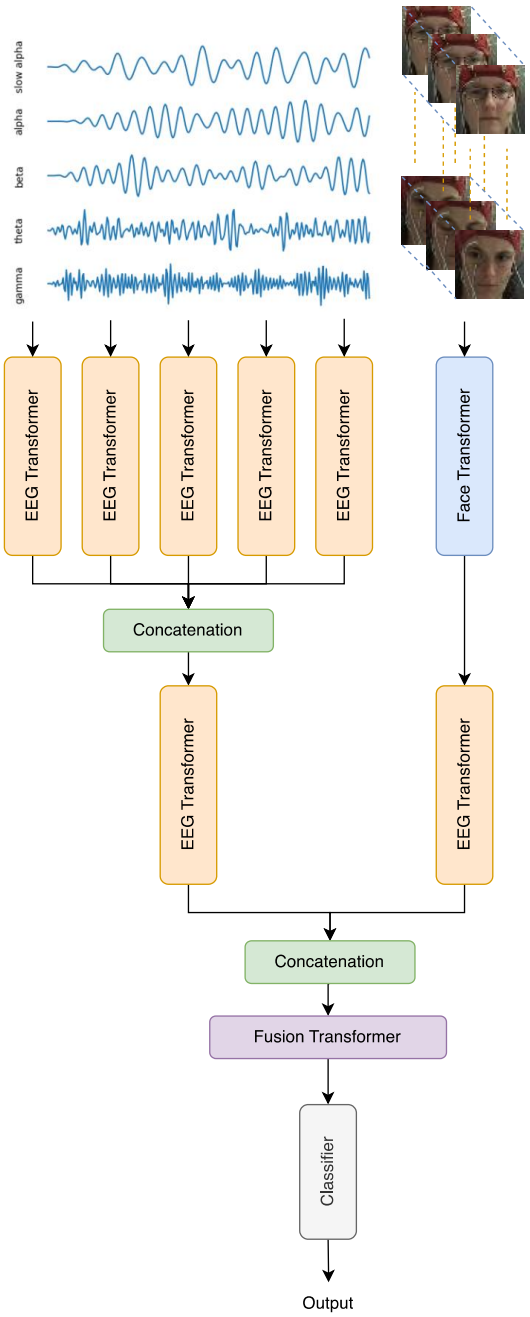


Figure 6.1. Architecture of the proposed model.

6.2 Proposed Model

To the best of our knowledge, this is the first time that a Transformer-based model has been proposed to recognize emotion using the EEG and facial modalities. Figure 6.1 illustrates the architecture of the proposed model. Our bimodal network leverages multiple Transformer stages. Initially, we employ five separate EEG transformers for the EEG modality, each responsible for handling sequences of temporal features of EEG signals extracted from distinct frequency ranges discussed in Section 6.1. These Transformers operate on the 32 EEG channels to capture independence in channel-level spatial-temporal learning. Subsequently, their outputs are combined and passed through another EEG Transformer to capture spatial-temporal dependencies across different frequency regions while maintaining the independence between EEG channels. As shown in Figure 6.2, the EEG Transformer is composed of input embedding, positional embedding, and N identical Transformer encoder blocks each consisting of a multi-head attention layer followed by a fully connected dense layer with Gaussian Error Linear Units as an activation function. The choice of activation function was based on its demonstrated performance enhancements in the context of natural language processing Transformer-based models [162]. As can be seen in Figure 6.2, in addition to the two layers, the Transformer encoder block also has residual skip connections around both layers along with Norm (Normalization) layers.

The multi-head attention module utilized in the Transformer encoder employs scaled-dot product attention across queries (Q), keys (K), and values (V) (all organized as matrices) within each attention head as in [27]. Practically, Q, K, and V first undergo

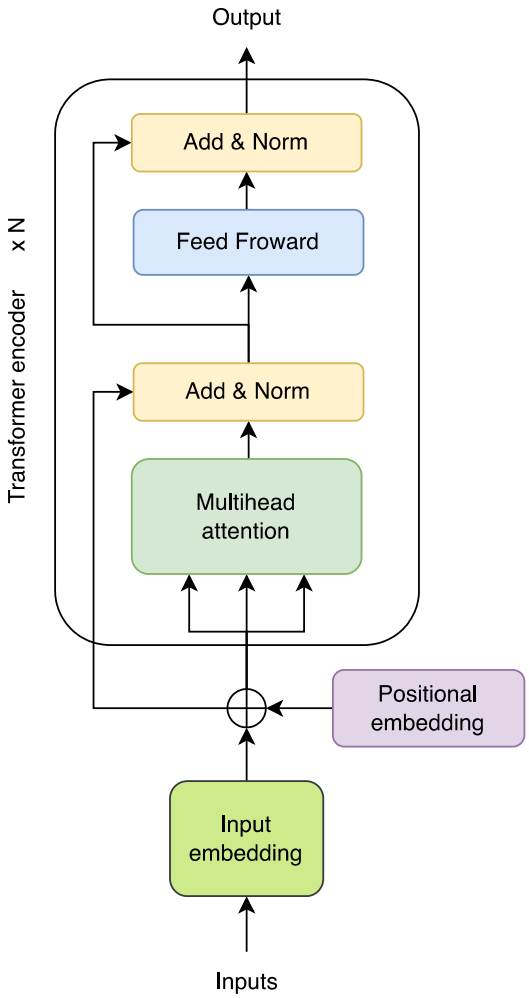


Figure 6.2. The architecture of the EEG encoder.

separate projections h times in different subspace $head_i$ ($i \in [1,\ h]$) with different learned linear projections W_i^Q, W_i^K and W_i^V with dimensions dk, dk, and dv as follows:

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{6.1}$$

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6.2}$$

$$MultiHead(Q,K,V) = Concat(head1,\dots,head_i)W^{out} \tag{6.3}$$

where i refers to the ith head of the total heads. $W_i^x (x \in [Q,K,V])$ is the projection parameter for ith head of multi-head Attention where, $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in$

$\mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, $W^{out} \in \mathbb{R}^{hd_v \times d_{model}}$ and $\frac{1}{\sqrt{d_k}}$ is a scaling factor, d_{model} is the output dimension of the model.

As depicted in Figure 6.1, we also incorporate facial information in addition to EEG data. The preprocessed facial data are fed to a face Transformer with an identical structure as the ViT model [28] albeit with distinct hyper-parameters fine-tuned for our work. Moreover, we have added a convolutional layer for image patching instead of pachify stem method used in ViT model [28]. The pachify stem method is implemented by a non-overlapping $p \times p$ image patches with stride of p, where p is chosen as 16 by default. However, adding a convolutional layer with *k* as kernel size and *s* as stride acts as a feature extractor layer that helps the model to converge faster and achieve a higher recognition accuracy [163]. The effectiveness of replacing the pachify stem that existed in the ViT model with convolutional layers composed of a small kernel size has been proven in [163]. Therefore, we adopted the convolutional patching technique to transform our image sequences into diverse patches. For our work, the number of patches is considered as a hyper-parameter which is further optimized. These patches were then flattened and fed to input embedding block and then to the positional embedding and L Transformer encoder blocks that are identical to the ones used for the EEG Transformer in terms of their structure but their weights will be optimized during training process. Subsequently, the output of the face Transformer is fed to another EEG Transformer as depicted in Figure 6.1.

The face Transformer structure is represented in Figure 6.3. As can be seen from Figure 6.3, the face Transformer is composed of a 3D convolution layer, input embedding, positional embedding, and L identical Transformer encoder blocks.

Finally, to fuse the modalities, the outputs of both modalities are concatenated followed by a fusion Transformer which has an identical structure to the EEG Transformer although its hyperparameters are optimized and fine-tuned. The fusion Transformer attends to interactions between the two modalities and captures the importance of high-level features extracted from each modality via the attention mechanism. As for the classification, we employ a single dense layer with a softmax activation function to perform emotion classification for the arousal and valence dimensions.

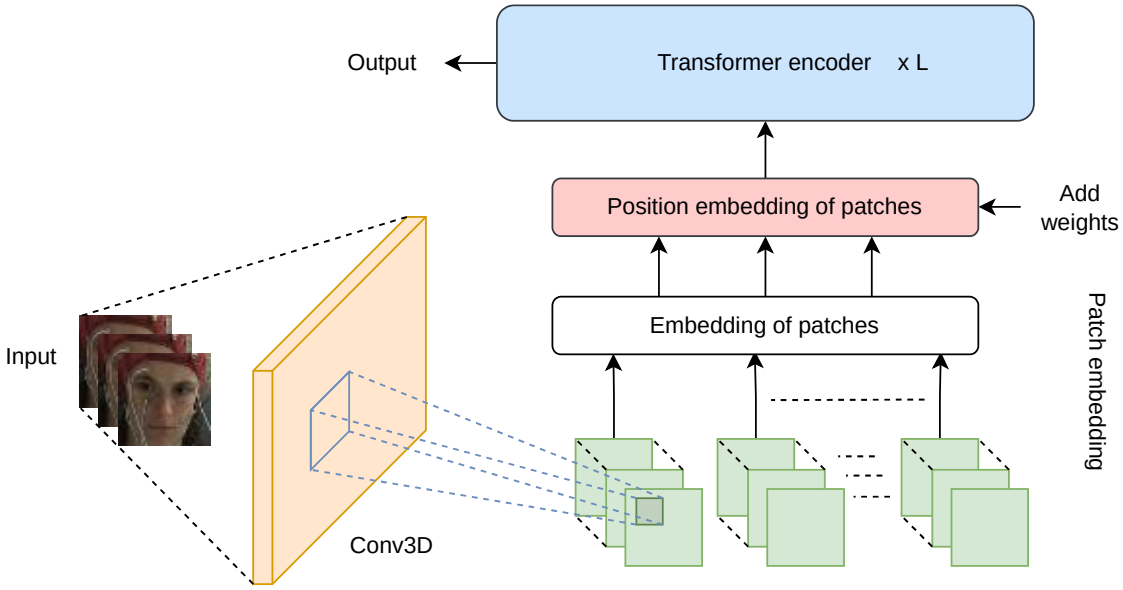


Figure 6.3. Architecture of the face encoder.

Our proposed model training process is composed of three distinct stages. In the initial stage, we focus on training the EEG transformer to encode EEG data. Moving to the second stage, our focus shifts to training the face Transformer and encoding the image data of faces extracted from the databases. In the third and final stage, we proceed to train the fusion Transformer, incorporating the concatenated encoded inputs obtained from both the EEG Transformer and face Transformer.

To optimize the network performance, we employed the Adam optimizer [43] with a learning rate set at 0.001. We selected Categorical Cross Entropy for the loss function [154]. The chosen batch size for training is 50.

For hyperparameter optimization across all Transformer models, we utilized Bayesian optimization with Gaussian Processes [153]. The specific hyperparameters we focused on optimizing included the number of encoder Transformers, the output dimension of encoder Transformers, the key dimension of multi-head attention, the number of multi-head attention heads, and the EEG sequence size.

Next, we will perform a comprehensive performance analysis to demonstrate the efficiency of the proposed model.

6.3 Performance Analysis and Experimental Results

In this section, recent notable deep learning models proposed for emotion estimation are implemented and a comparison is presented. The evaluation of the proposed model is performed using multiple phases. In the first phase, several recent EEG-based emotion recognition models are implemented and compared with the proposed model in a unimodal setting (using EEG as the sole modality) using the DEAP dataset. In the second phase, the proposed bimodal model (using EEG and facial expression) is compared with several recent multimodal emotion recognition models. We implemented all the existing models and trained and evaluated them on the same dataset for a fair comparison. In the last phase, we test the effectiveness of different fusion methods on the proposed model to demonstrate how the proposed fusion method contributes to improving the performance of the proposed model using DEAP and MAHNOB-HCI datasets.

We performed four-fold cross-validation on the training set and reported the average accuracy of the four folds as the training accuracy. Throughout the implementation of all models, 20% of the data is excluded as a testing dataset. Thus, we train and validate the model on the data of the remaining 80%. The training process involves 100 epoch iterations, where we select the best-performing model based on train and test accuracies. Once the best-performing model is selected, we compare the cross-validation performance results to those of the test set. In addition to accuracy, we calculate the F1-score for each class to ensure results reliability as described in [155]. This metric describes the prediction capability of the assessed models for each class (low=0 and high=1).

6.3.1 First Testing Phase

In this testing phase, we will evaluate a unimodal adaptation of the proposed model presented in Section 6.2. To construct the unimodal model, we discard the facial data and fusion Transformer. Hence, instead of fusing the EEG and video modalities, we attach the classifier directly to the output of the deepest EEG Transformer processing the EEG data. Figure 6.4 represents the unimodal adaption of the proposed model.

Table 6.1 compares the results of EEG-based emotion recognition of nine state-of-the-art models with the proposed unimodal model: CNN and DNN by Tripathi et al. [63], RNN by Alhagry et al. [67], wavelet coefficient based multilayer perceptron (MLP) in theta band proposed by Pandey et al. [71], TCN model proposed in [66], BiLSTM network proposed in [76] consisted bidirectional LSTM recurrent neural networks, a cascade of attention-based LSTM autoencoder and attention-based CNN [99], HSLT Transformer-based model [140], and EeT spatial-temporal Transformer-based model [143].

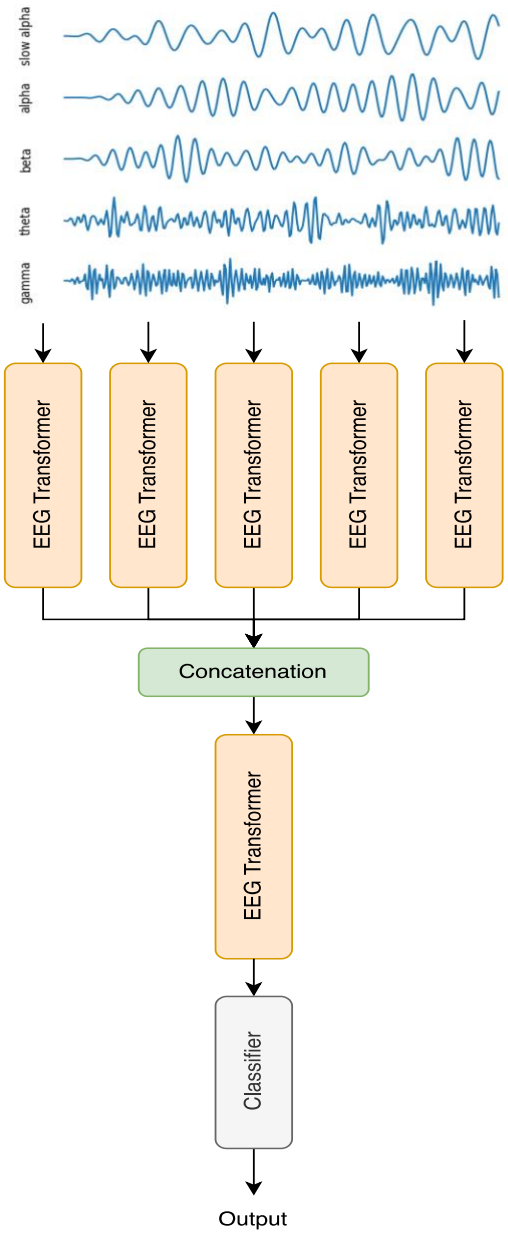


Figure 6.4. Architecture of the unimodal adaption of the proposed model.

Performance Analysis

As we can see in Table 6.1, the proposed unimodal Transformer-based model, reaches an average train accuracy of 0.71 and 0.67 for valence and arousal emotion dimensions, respectively. Whereas accuracy obtained for the test set with the same model is 0.65 and 0.62 for valence and arousal emotion dimensions, respectively.

Table 6.1. Comparison among different EEG-based models for valence and arousal classification for 2 classes

Models	Valence			Arousal		
	F1-Score: *Class 0* *Class 1*	Train Acc.	Test Acc.	F1-Score: *Class 0* *Class 1*	Train Acc.	Test Acc.
DNN [63]	0.64 0.00	0.61	0.47	0.34 0.00	0.63	0.20
CNN [63]	0.57 0.66	0.97	0.63	0.34 0.00	0.56	0.20
RNN [67]	0.00 0.69	0.54	0.53	0.00 0.88	0.55	0.70
MLP (theta band) [71]	0.36 0.42	0.56	0.55	0.48 0.63	0.57	0.62
TCN [66]	0.67 0.26	0.51	0.54	0.00 0.75	0.57	0.59
BiLSTM [76]	0.44 0.54	0.79	0.50	0.68 0.00	0.79	0.52
ALSTM Autoencoder + ACNN [99]	0.42 0.62	0.99	0.54	0.47 0.64	0.99	0.57
HSLT Transformer [140]	0.46 0.37	0.73	0.42	0.35 0.64	0.71	0.54
EeT Transformer [143]	0.56 0.37	0.88	0.48	0.27 0.65	0.87	0.53
Proposed EEG Model	**0.46** **0.75**	**0.71**	**0.65**	**0.39** **0.72**	**0.67**	**0.62**

Among the other nine models, the MLP model [71] seems to provide generally more reliable results (i.e. smaller discrepancy between the train and test results) however, the proposed model achieves a higher accuracy for both train and test datasets while maintaining a small discrepancy between train and test results and higher F1-scores on both

classes. The discrepancy between the train and test results for DNN [63], CNN [63], BiLSTM [76], ALSTM Autoencoder + ACNN [99], HSLT Transformer [140], and EeT Transformer [143] models may be due to over-fitting on the training set, thereby causing them to generalize poorly on the unseen test dataset. Model RNN [67] and TCN [67] do not seem to overfit, however, they all achieve inferior accuracy and F1-score results. In particular, RNN [67] achieves a 0 F1-score for class 0; this means that this model is unable to predict low valence and low arousal. Therefore, despite achieving reasonable classification accuracy, it does not perform as well as the proposed model.

In the next phase, the proposed bimodal Transformer-based model will be compared with recent multimodal state-of-the-art emotion recognition models.

6.3.2 Second Testing Phase

In this section, we tested recent deep bimodal frameworks for emotion recognition purposes. To the best of our knowledge, this is the first time different recent competing deep bimodal models are implemented under identical settings (e.g., same dataset, identical test and train samples, etc.), and a comparison with the proposed model is presented. The majority of previous studies only experimented with fusion methods on their model and did not thoroughly investigate its robustness in comparison to other models [132-134, 137].

The selected bimodal frameworks make use of facial expressions and EEG as these two are the most widely used modalities in the literature [8, 29].

Table 6.2 compares nine recent bimodal emotion recognition models with the proposed model. The nine models are listed as follows:

Table 6.2. Bimodal emotion recognition models

Model		Accuracy		F1-Score	
		Train	Test	Class 0	Class 1
Enumerate [137]*	Valence	0.58	0.60	0.52	0.65
	Arousal	0.60	0.62	0.21	0.75
AdaBoost [137]*	Valence	0.58	0.45	0.00	0.62
	Arousal	0.65	0.60	0.00	0.75
Sum Strategy [132]*	Valence	0.88	0.57	0.66	0.57
	Arousal	0.80	0.47	0.16	0.61
CCA [133]**	Valence	0.52	0.57	0.51	0.62
	Arousal	0.61	0.60	0.00	0.75
BDAE [134]**	Valence	0.64	0.61	0.27	0.72
	Arousal	0.70	0.51	0.00	0.82
MFC [134]**	Valence	0.47	0.52	0.67	0.09
	Arousal	0.61	0.60	0.00	0.75
Weight-based [136]*	Valence	0.66	0.52	0.50	0.55
	Arousal	0.71	0.58	0.33	0.70
Weight-based [164]*	Valence	0.56	0.51	0.00	0.67
	Arousal	0.62	0.62	0.00	0.76
Soft-Voting Strategy [165]**	Valence	0.99	0.52	0.50	0.54
	Arousal	0.98	0.61	0.48	0.68
Proposed Bimodal Model*	Valence	**0.66**	**0.65**	**0.50**	**0.73**
	Arousal	**0.72**	**0.66**	**0.51**	**0.66**

* Decision level fusion methods
** Feature level fusion methods

- The bimodal model implemented from [137] which uses CNN and SVM for face and EEG data, respectively. Enumerate and AdaBoost fusion methods are also tested for their proposed model to fuse the modalities.
- The bimodal model implemented from [132] which uses one feed-forward dense layer and SVM for face and EEG data, respectively where Sum strategy fusion method is also deployed.
- The bimodal model implemented from [133] which uses SVM and KNN for face and EEG data, respectively. The canonical Correlation Analysis fusion method is also deployed to fuse the modalities.

- The bimodal model implemented from [134] is also implemented. Bimodal Deep Auto Encoder (BDAE) and multiple feature concatenation fusion (MFC) methods are examined. In BDAE, high-level representative features from EEG and face data are extracted using BDAE such that, Restricted Boltzman Machine (RBM) is trained for each modality as an encoding part. Then, hidden layers of RBMs are concatenated and fed their classification system. In MFC, the hand-crafted features of EEG and high-level features of eye images outputted from a CNN+RNN model are concatenated directly and then fed to their classification system. SVM is used as classifier for their model to perform emotion classification.
- The bimodal model implemented from [136] which uses CNN+DNN and 3D CNN for EEG and face data, respectively. Weight-based fusion method (similar to enumerate fusion method) is also used to fuse the modalities.
- The bimodal model implemented from [164] which uses VGG-16 deep CNN followed by attention mechanism and VGG-16 followed by 5-layer LSTM for emotion classification from face and EEG data, respectively. Weight-based fusion method is also deployed for their proposed model.
- The bimodal model implemented from [165] which uses cascade of a 3D convolution, Gated Recurrent Unit (GRU) and a dense layer for facial data. On the other hand, a model comprised of 2D convolution, GRU and a dense layer is used for EEG data. Then, the output of final dense layers from both modalities are concatenated. Soft-voting Strategy fusion method is used for their proposed model where the concatenated features are fed to different classifiers as SVM,

Random Forest (RF), Logistic Regression (LR) and Extreme Gradient Boost to evaluate the performance of their proposed model. Soft-Voting strategy will assign a weight to each classifier based on their performance to perform emotion classification.

Performance analysis

Table 6.2 provides a comparison among previously presented bimodal frameworks in the literature. As can be seen in Table 6.2, the high discrepancy between train and test accuracies of bimodal models presented in [132] and [165] indicates a potential case of overfitting for these particular models. Bimodal models presented in [137] and [134] which are both tested under two fusion methods and [133] do not overfit however, their overall performance tends to be inferior compared to the proposed model as they demonstrate lower accuracies on train and test datasets with low F1-score across both classes. Specifically, in [134], F1-score of class 1 over arousal dimension under both fusion methods is relatively higher than the proposed model however the model represents 0 for F1-score of class 0 over arousal dimension which means the model is unable to predict low arousal under both fusion methods. The bimodal model presented in [137] shows a more reliable results under the enumerate fusion method however the proposed model achieves relatively higher train and test accuracy for both valence and arousal dimensions. The bimodal model presented in [136] demonstrates higher training accuracies on both emotion dimensions and higher F1-scores for both classes compared to the other eight state-of-the-art competing bimodal models. However, the proposed model achieves relatively higher test accuracies for both valence and arousal dimension. The bimodal model presented in

[164] performs poorly as it achieves 0 for F1-score of class 0 on both valence and arousal dimensions.

Next, we want to test different fusion techniques on the proposed model. To the best of our knowledge, this is the first work that tests various fusion techniques on a bimodal Transformer-based emotion recognition model using EEG and face modalities.

6.3.3 Third Testing phase

In this phase, we will examine the performance of all the applicable aforementioned fusion methods on the proposed model using DEAP and MAHNOB-HCI datasets. The results are presented by Table 6.3 and Table 6.4. Model hyper-parameters discussed in

Table 6.3. Performance comparison of multiple feature-level and decision-level fusion methods tested on Transformer-based model for DEAP dataset

Model		Accuracy		F1-Score	
		Train	Test	Class 0	Class 1
Enumerate [137]	Valence	0.60	0.60	0.18	0.73
	Arousal	0.90	0.62	0.52	0.69
AdaBoost [137]	Valence	0.56	0.58	0.00	0.73
	Arousal	0.90	0.62	0.52	0.69
Sum Strategy [132]	Valence	0.56	0.58	0.00	0.73
	Arousal	0.90	0.61	0.48	0.69
CCA [133]	Valence	0.68	0.54	0.43	0.62
	Arousal	0.92	0.62	0.57	0.67
MFC [134]	Valence	0.72	0.56	0.39	0.65
	Arousal	0.99	0.65	0.54	0.71
Soft-Voting Strategy [165]	Valence	0.99	0.61	0.50	0.69
	Arousal	0.99	0.60	0.31	0.72
Proposed Bimodal Model	Valence	**0.66**	**0.65**	**0.50**	**0.73**
	Arousal	**0.72**	**0.66**	**0.51**	**0.66**

Table 6.4. Performance comparison of multiple feature-level and decision-level fusion methods tested on Transformer-based model for MAHNOB-HCI dataset

Model		Accuracy		F1-Score	
		Train	Test	Class 0	Class 1
Enumerate [137]	Valence	0.68	0.61	0.47	0.70
	Arousal	0.71	0.67	0.67	0.69
AdaBoost [137]	Valence	0.67	0.58	0.45	0.67
	Arousal	0.71	0.67	0.67	0.69
Sum Strategy [132]	Valence	0.69	0.62	0.49	0.70
	Arousal	0.72	0.66	0.65	0.67
CCA [133]	Valence	0.55	0.57	0.18	0.71
	Arousal	0.99	0.55	0.56	0.54
MFC [134]	Valence	0.67	0.57	0.41	0.66
	Arousal	0.75	0.64	0.60	0.67
Soft-Voting Strategy [165]	Valence	0.99	0.52	0.46	0.56
	Arousal	0.99	0.59	0.53	0.64
Proposed Bimodal Model	Valence	**0.69**	**0.61**	**0.58**	**0.58**
	Arousal	**0.73**	**0.68**	**0.61**	**0.73**

Table 6.5. Proposed Transformer-based model hyper-parameters

Model	Number of encoder blocks	Output of Transformer dimension	Number of attention heads	Key dimension of Transformer	EGG sequence size
Face Transformer	7	18	4	18	33
EEG Transformer	5	50	5	59	
Fusion Transformer	2	21	3	39	

Section 6.2 are also optimized for each fusion method. The optimum hyper-parameters for the proposed model with Transformer fusion are presented in Table 6.5. Figure 6.5 and Figure 6.6 also show how decision-level and feature-level fusion methods are incorporated into the proposed model.

The objective here is to test the effectiveness of the Transformer fusion method on the proposed model. Therefore, we keep the overall structure of the proposed model and test different feature-level and decision-level fusion methods on the proposed model.

For feature-level fusion methods such as CCA and MFC, we train each modality separately and add a single dense layer with softmax activation function on the output of final EEG Transformer of each modality disregarding the fusion Transformer part represented in Figure 6.1.

Then, high-level features obtained from the last layer before each modality's classifier are used to further apply a feature-level fusion method. In MFC, these features are simply concatenated and fed to a single dense layer with softmax activation function classifier to perform emotion classification. in CCA feature-level fusion, CCA is applied on each modality's high-level features to extract the most informative features from each modality. Then, the extracted features are concatenated and fed to a single dense layer with softmax activation function classifier to perform emotion classification.

For decision-level fusion methods such as enumerate/weight-based, AdaBoost, sum strategy and soft-voting strategy, we train each modality separately as in previous part. Then, each decision-level fusion method is applied on the classifications results from the two modalities.

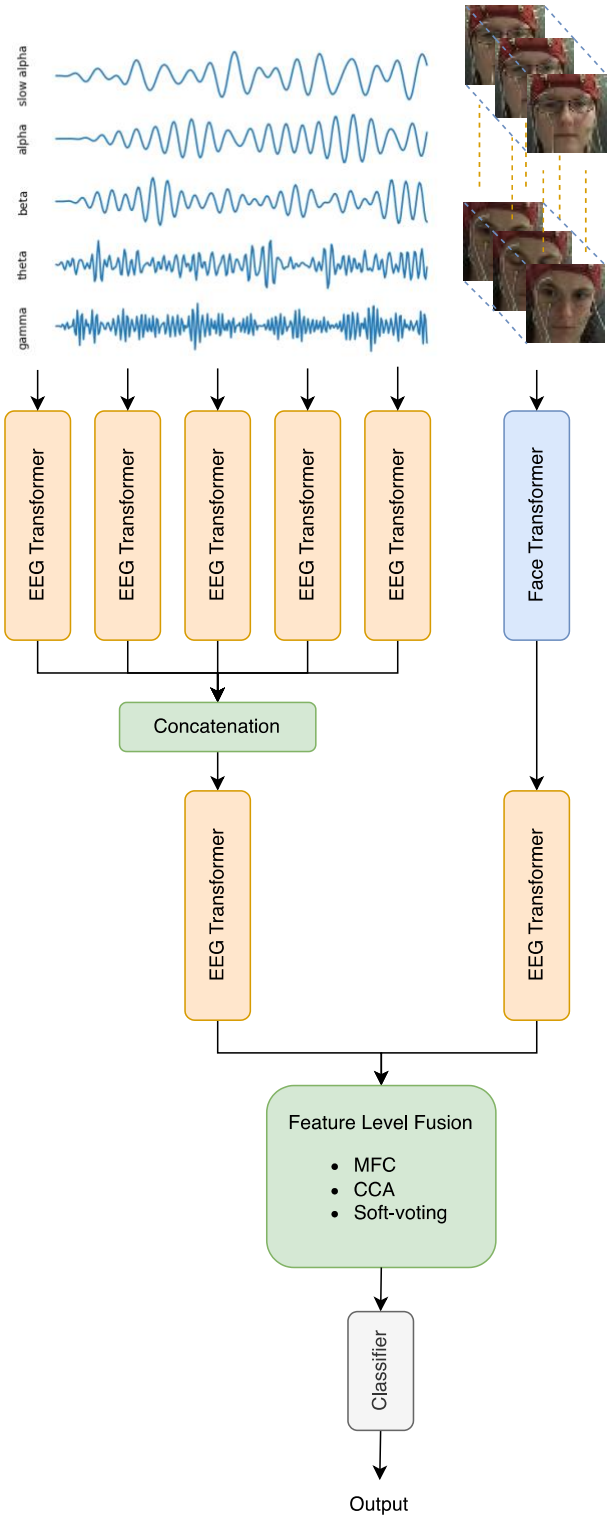


Figure 6.5. Architecture of the proposed model using state-of-the-art feature-level fusion methods.

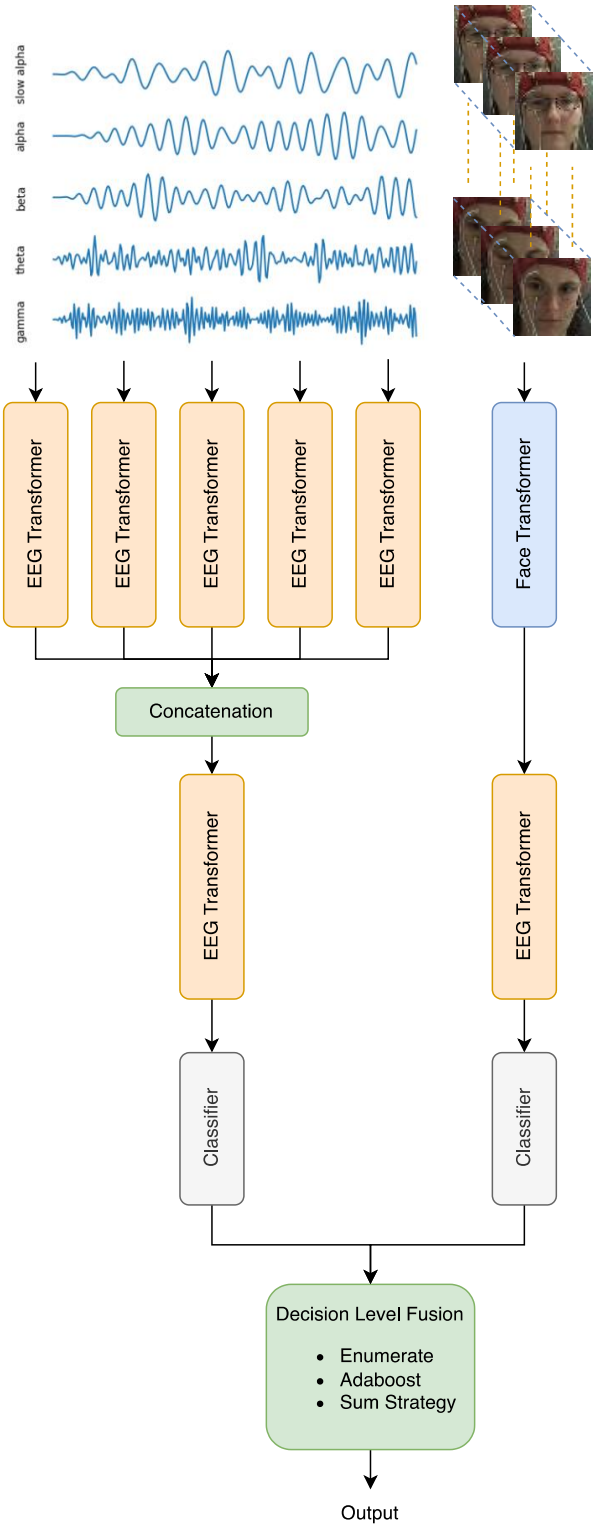


Figure 6.6. Architecture of the proposed model using state-of-the-art decision-level fusion methods.

Performance Analysis

As can be seen from Table 6.3 and

Table 6.4, the proposed model with Transformer fusion reaches a train and test accuracy of 0.66, 0.65, and 0.72, 0.66 for both valence and arousal on the DEAP dataset, respectively. Moreover, the proposed model with Transformer fusion reaches a train and test accuracy of 0.69, 0.61, and 0.73, 0.68 for both valence and arousal on the MAHNOB-HCI dataset, respectively. In comparison to other fusion methods, the proposed Transformer-based fusion method provides more accurate results with higher reliability (i.e., lower discrepancy between the validation and test results). Moreover, the results of the F1-scores across all fusion methods reveal that the proposed Transformer-based fusion method is more capable of identifying the correct low and high classes compared to the other six methods as it is the only method possessing over 50% F1-score of valence and arousal for both classes.

Chapter 7. Conclusion and Further Research Plans

Automated emotion recognition models have the potential to revolutionize many areas of human life, from improving mental health to enhancing customer experiences in retail and service industries. There are several challenges in developing accurate emotion recognition models including the need for large and diverse data sets. In this thesis, we put essential effort towards finding automated emotion recognition models to overcome some of the key challenges facing automated emotion recognition.

In Chapter 4, we proposed a model for emotional state prediction from EEG signals. The proposed model uses a cascade of a pre-trained CNN and residual block-based LSTM structure. To evaluate the proposed model, we implemented state-of-the-art methods and tested all models on the DEAP dataset. We found that the proposed model presents the least discrepancy between the validation and test results while achieving superior performance compared to state-of-the-art methods on the DEAP dataset. We also compared the proposed model to a conventional CNN+LSTM network. We showed that the proposed model achieves a lower discrepancy between the validation and testing results while achieving consistent performance for both emotional dimensions.

In Chapter 5, we proposed a novel Contrastive Learning GAN-based Graph Neural Network for the emotion recognition task. This model addresses several challenges in the area of affective computing at once. The CL component is a self-supervised framework that is deployed to learn high-quality EEG representations and to solve inter-subject intra-subject emotion variabilities. The GAN component is utilized to address the limitation of dataset size by adding artificial realistic-like data to the real data. The GNN component is deployed to take the topological structure of EEG channels into consideration. The

proposed model was implemented and compared with several competing state-of-the-art models in a subject-independent experimental setting in the area of EEG-based emotion recognition. The results demonstrated the superior performance of the proposed model in terms of achieving a higher recognition accuracy on both arousal and valence dimensions for both DEAP and MAHNOB-HCI databases.

Although a better classification accuracy was achieved by the proposed models, emotion classification models were only proposed in unimodal settings using EEG signals. Combining multiple modalities might also lead to achieving better classification accuracy. Therefore, in our third solution, we investigated the effect of bimodal emotion recognition settings.

Therefore, in Chapter 6, a Transformer-based bimodal emotion recognition model using facial expressions and EEG modalities is proposed. Recent bimodal emotion recognition models are also implemented and compared with the proposed model. Moreover, the model is tested in an unimodal setting and a comparison with recent competing state-of-the-art unimodal models is presented. Various fusion techniques are also tested on the proposed model to investigate the effectiveness of the proposed fusion method for our work. The findings demonstrated that the proposed model outperformed others by achieving higher accuracy in recognizing both arousal and valence dimensions across the DEAP and MAHNOB-HCI databases. Moreover, based on the findings of Chapter 6, the inclusion of the face modality alongside EEG data has the potential to bring about a modest enhancement in an emotion recognition model. This suggests that the combined utilization of facial and EEG information could lead to a slight improvement in the model's ability to accurately identify and classify emotions.

Although, we have tested bimodal EEG and facial expression-based emotion recognition models, combining multiple modalities might also lead to achieving better classification accuracy. On the other hand, the size of the dataset remains the primary constraint affecting the advancement of automated emotion recognition models. Although attempts are made to synthetically enlarge the dataset as proposed in Chapter 5, there is a fundamental recognition that having a larger original dataset is critical for achieving more accurate, reliable, and adaptable automated emotion recognition models. This is because, Emotions are multifaceted and context-dependent, making them inherently challenging to capture comprehensively. Therefore, a limited dataset cannot cover the entire spectrum of human emotions and their variations. A larger original dataset would provide a more accurate representation of the diverse emotional states people experience. However, when researchers attempt to artificially augment a small dataset, they often use techniques like data augmentation, which involves creating new examples by applying various transformations to the existing data. While this can help in some cases as noticed in Chapter 5, it might not replace the need for a genuinely extensive and diverse dataset. Models trained on larger, more representative datasets are more likely to generalize well to real-world scenarios, where emotional expressions can vary significantly. Moreover, the use of only a subset of EEG channels for emotion recognition might be promising as research avenue. While conventional approaches often utilize a large number of EEG channels to perform EEG-based emotion recognition, focusing on a select few channels may offer several potential advantages.

Therefore, this research can further be extended as follows:

- Delving deeper into the realm of emotion recognition by considering multiple modalities. To enhance the effectiveness of emotion recognition models, it is crucial to leverage various sources of information. These sources, or modalities, can include facial expressions, vocal intonations, physiological signals, neurophysiological signals, and more. For instance, combining facial expressions, EEG, and physiological signals like heart rate and skin conductance might provide a more accurate and robust assessment of an individual's emotional state compared to using just one or two modalities. To better understand the individual contributions of each modality, it is essential to investigate the effect of each modality separately in an unimodal setting. For example, analyzing how well a model performs when considering only heart rate modality, or when relying solely on facial expression features, can reveal the strengths and limitations of each modality. After thoroughly examining each modality's performance in isolation, it is substantial to investigate the potential advantages of combining these modalities in various bimodal or multimodal settings. This involves developing fusion strategies that integrate information from multiple sources to perform emotion recognition.
- Addressing the limitation of datasets in emotion recognition research is a critical challenge. Addressing this challenge requires substantial efforts in collecting more extensive data that reflects the diversity of human emotional experiences containing diverse cultural backgrounds and languages to ensure that research in this area remains relevant, equitable, and capable of producing models that are both accurate and adaptable to the complexities of human emotions in real-world scenarios.

- The exploration of different data augmentation techniques for emotion recognition represents a valuable avenue for future research although we tested one DA technique, GAN in Chapter 5. While DA techniques may not provide a comprehensive solution to all the challenges in understanding human emotions, they offer a pragmatic approach to improving the quality and diversity of training data, ultimately leading to more robust and effective emotion recognition models that can be applied in a wide array of contexts.
- The investigation of selecting few EEG channels for emotion recognition holds promise for future research. Selecting a subset of EEG channels may lead to improved computational efficiency and reduced processing requirements. By focusing on channels that are most informative for emotion recognition, researchers can potentially achieve comparable or even superior performance while minimizing computational resources and energy consumption.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," *MIT press,* 2016.

[2] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," *in International Conference on Machine Learning*, 2016, pp. 173–182.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, 2001, vol. 18, no. 1, pp.32-80.

[4] R. W. Picard, "Affective Computing," *MIT press,* 2000.

[5] M. Cheon, S. Kim, C. Chae, and J. Lee, "Quality assessment of mobile videos," pp. 99–127, *Springer International Publishing,* Cham, 2015.

[6] R. Gupta, "Using Affective Brain-Computer Interfaces to Characterize Human Influential Factors for Speech Quality-of-Experience Perception Modelling," *Human-centric computing and information sciences*, 6.1, (2016).

[7] R. Gupta, "Physiology-based Quality-of-Experience Assessment for Next Generation Multimedia Technologies," *Doctoral dissertation, Université du Québec, Institut national de la recherche scientifique*, 2016.

[8] S. Moon, and J. Lee, "Implicit Analysis of Perceptual Multimedia Experience Based on Physiological Response: A Review," *in IEEE Transactions on Multimedia,* vol. 19, no. 2, pp. 340-353, Feb. 2017.

[9] X. Hu, J. Chen, F. Wang, and D. Zhang, "Ten Challenges for EEG-Based Affective Computing," *Brain sci. adv.,* vol. 5, no. 1, pp. 1- 20, 2019.

[10] M. Sreeshakthy, and J. Preethi, "Classification of human emotion from DEAP EEG signal using hybrid improved neural networks with cuckoo search," *BRAIN. Broad Research in Artificial Intelligence and Neuroscience,* vol. 6, no. 3-4, pp. 60–73, 2016.

[11] K.S. Fleckenstein, "Defining affect in relation to cognition: a response to Susan McLeod," *Journal of Advanced Composition,* pp. 447-453, 1991.

[12] Y. Li, L. Wang, W. Zheng, Y. Zong, L. Qi, Z. Cui, T. Zhang, and T. Song, "A novel bi-hemispheric discrepancy model for EEG emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems,* vol. 13, no. 2, pp.354-367, 2020.

[13] Y. J. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, and Y. Shi, "Real-time movie-induced discrete emotion recognition from EEG signals," *IEEE Trans. Affect. Comput.*, 2017.

[14] J. Li, S. Qiu, Y.Y. Shen, C.L. Liu, and H. He, "Multisource transfer learning for cross-subject EEG emotion recognition," *IEEE transactions on cybernetics,* vol. 50, no. 7, pp.3281-3293, 2019.

[15] Z. Lan, O. Sourina, L. Wang, R. Scherer, and G.R. Müller-Putz, "Domain adaptation techniques for EEG-based emotion recognition: a comparative study on two public datasets," *IEEE Transactions on Cognitive and Developmental Systems,* vol. 11, no. 1, pp.85-94, 2018.

[16] P. Zhong, D. Wang, and C. Miao, "EEG-Based emotion recognition using regularized graph neural networks," *IEEE Trans. Affect. Comput.,* 2020,

[17] S. Hizlisoy, S. Yildirim, and Z. Tufekci, "Music emotion recognition using convolutional long short term memory deep neural networks," *Engineering Science and Technology, an International Journal,* vol. 24, p. 760–767, 2021.

[18] Y. Yin, X. Zheng, B. Hu, Y. Zhang, and X. Cui, "EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM," *Applied Soft Computing,* 100, p.106954, 2021.

[19] J. Liao, Q. Zhong, Y. Zhu and D. Cai, "Multimodal physiological signal emotion recognition based on convolutional recurrent neural network," *In IOP Conference Series: Materials Science and Engineering,* vol. 782, p. 032005, IOP Publishing, 2020.

[20] T. Zhang, W. Zheng, Z. Cui, Y. Zong and Y. Li, "Spatial–Temporal recurrent neural network for emotion recognition," *in IEEE Transactions on Cybernetics,* vol. 49, no. 3, pp. 839-847, March 2019.

[21] A. Graves, A. Mohamed and G. Hinton, "Speech recognition with deep recurrent neural networks," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing,* Vancouver, BC, pp. 6645-6649, 2023.

[22] J. Kim, M. El-Khamy, and J. Lee, "Residual LSTM: design of a deep recurrent architecture for distant speech recognition," *arXiv preprint arXiv:1701.03360,* 2017.

[23] S. Chaib, Y. Hongxun, G. Yanfeng, and A. Moussa. "Deep feature extraction and combination for remote sensing image classification based on pre-trained CNN models," *In Ninth International Conference on Digital Image Processing (ICDIP)*, vol. 10420, p. 104203D, 2017.

[24] J. Liu, X. Shen, S. Song, and D. Zhang, "Domain Adaptation for Cross-Subject Emotion Recognition by Subject Clustering," *in Int. IEEE/EMBS Conf. Neural Eng. (NER),* pp. 904-908, 2021.

[25] J. Teo, C. L.Hou, and J. Mountstephens, "Deep learning for EEG-based preference classification," *In AIP Conference Proceedings*, vol. 1891, no. 1, p. 020141, 2017.

[26] N.S. Suhaimi, J. Mountstephens, and J. Teo, "EEG-based emotion recognition: A state-of-the-art review of current trends and opportunities," *Computational intelligence and neuroscience,* 2020.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *In NIPS,* 2017.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and J. Uszkoreit, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929,* 2020.

[29] F. Muhammad, M. Hussain, and H. Aboalsamh, "A bimodal emotion recognition approach through the fusion of Electroencephalography and facial sequences," *Diagnostics,* vol. 13, no. 5, p.977, 2023.

[30] S. S. Gilakjani, H. Al Osman, "Emotion classification from Electroencephalogram signals using a cascade of convolutional and block-based residual recurrent neural networks," *2022 IEEE Sensors Applications Symposium (SAS),* pp. 1-6, 2022.

[31] S. S. Gilakjani, H. Al Osman, " Contrastive learning generative adversarial network based graph neural network for emotion recognition," *Submitted to IEEE Transaction on Affective Computing,* 2023.

[32] S. S. Gilakjani, H. Al Osman, " Transformer-based bimodal emotion recognition model with fusion transformer," *Submitted to IEEE Transaction on Affective Computing*, 2023.
[33] T.M. Mitchell, "Machine learning," *New York: McGraw-hill*, vol. 1, no. 9, 1997.
[34] A. Craik, Y. He, and J. L Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: a review," *Journal of neural engineering*, vol. 16, no. 3, 2019.
[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 2017.
[36] H. Wang, and B. Raj, "On the origin of deep learning," *arXiv preprint arXiv:1702.07800*, 2017.
[37] A.L. Fradkov, "Early history of machine learning," *IFAC-PapersOnLine*, vol. 53, no. 2, pp.1385-1390, 2020.
[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *in Proc. IEEE Conference on International Conference on Neural Information Processing Systems*, pp. 1097–1105, 2012.
[39] M.Z. Alom, T.M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M.S. Nasrin, B.C. Van Esesn, A.A.S. Awwal, and V.K. Asari, "The history began from alexnet: A comprehensive survey on deep learning approaches," *arXiv preprint arXiv:1803.01164*, 2018.
[40] Y. Bengio, A. Courville, P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
[41] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, 61, pp.85-117, 2015.
[42] D.E. Rumelhart, G.E. Hinton and R.J Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp.533-536, 1986.
[43] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[44] R. Cabada, H. Rangel, M. L. Estrada, and H. M. Lopez, "Hyperparameter optimization in CNN for learning-centered emotion recognition for intelligent tutoring systems," *Soft Computing*, vol. 24, pp. 7593–7602, 2020.
[45] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition," *Neural networks*, vol. 1, no. 2, pp.119-130, 1988.
[46] S. Sathasivam, and W.A.T.W. Abdullah, "Logic learning in Hopfield networks," *arXiv preprint arXiv:0804.4075*, 2008.
[47] J.L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp.179-211, 1990.
[48] M.I. Jordan, "Serial order: A parallel distributed processing approach," *In Advances in psychology*, vol. 121, pp. 471-495, North-Holland, 1997.
[49] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory”, *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.
[50] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[51] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," *In International conference on machine learning*, pp. 2342-2350, June 2015.
[52] S. Becker, and G. E. Hinton, "Self-organizing neural network that discovers surfaces in random-dot stereograms," *Nature,* vol. 355, no. 6356, pp. 161–163, 1992.
[53] K. Pinitas, K. Makantasis, A. Liapis, and G.N. Yannakakis, "Supervised contrastive learning for affect modelling," *In International Conference on Multimodal Iinteraction,* pp. 531-539, November 2022.
[54] E. Lashgari, D. Liang, and U. Maoz, "Data augmentation for deep-learning-based electroencephalography," *Journal of Neuroscience Methods*, 346, p.108885, 2020.
[55] I. Goodfellow et al., "Generative adversarial nets," *in Proc. Adv. Neural Inf. Process. Syst.*, pp. 2672–2680, 2014.
[56] I. B. Mauss, and M. D. Robinson, "Measures of emotion: a review," *Cognition and emotion,* vol. 23, no. 2, p. 209-237, 2009.
[57] J.A. Russell, "Culture and the categorization of emotions," *Psychological Bull.*, vol. 110, no. 3, pp. 426-450, 1991.
[58] J.A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Research in Personality*, vol. 11, no. 3, pp. 273-294, Sept. 1977.
[59] D. Garg, and G.K. Verma, "Emotion recognition in valence-arousal space from multi-channel EEG data and wavelet based deep learning framework," *Procedia Computer Science,* pp.857-867, 2020.
[60] XW. Wang, D. Nie, BL. Bu, "Emotional state classification from EEG data using machine learning approach," *Neuro-Computing,* pp. 94-106, 2014.
[61] S. Lokannavar, P. Lahane, A. Gangurde, P. Chidre, "Emotion recognition using EEG signals," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 4, pp. 54-56, 2015.
[62] J. Li, Z. Zhang, and H. He, "Hierarchical convolutional neural networks for EEG-based emotion recognition," *Cognit. Comput.*, pp. 1–13, 2017.
[63] S. Tripathi, S. Acharya, R.D. Sharma, , S. Mittal, and S. Bhattacharya, "Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset," *in Proc. IAAI,* pp. 4746–4752, 2017.
[64] S. Koelstra, C. Muehl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, "DEAP: A database for emotion analysis using physiological signals," *IEEE Transaction on Affective Computing, Special Issue on Naturalistic Affect Resources for System Building and Evaluation,* vol. 3, pp.18-31, 2012.
[65] M.R. Islam, M.M. Islam, M.M. Rahman, C. Mondal, S.K. Singha, M. Ahmad, A. Awal, M.S. Islam, and M.A. Moni, "EEG channel correlation based model for emotion recognition," *Computers in Biology and Medicine,* 136, p.104757, 2021.
[66] Y. Ding, N. Robinson, Q. Zeng, D. Chen, A. Wai, T. S. Lee and C. Guan, "Tsception: a deep learning framework for emotion detection using EEG," *In 2020 International Joint IEEE Conference on Neural Networks (IJCNN),* IEEE, pp. 1-7, 2020.
[67] S. Alhagry, A. Fahmy and R. A. El-Khoribi, "Emotion recognition based on EEG using LSTM recurrent neural network," *International Journal of Advanced Computer Science and Applications (ijacsa),* vol. 8, 2017.

[68] D. Nath, M. Singh, D. Sethia, D. Kalra, and S. Indu, "A comparative study of subject-dependent and subject-independent strategies for EEG-based emotion recognition using LSTM network," *Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis (ICCDA)*, Association for Computing Machinery, p. 142–147, 2020.

[69] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang and Z. Cui, "MPED: A Multi-Modal physiological emotion database for discrete emotion recognition," in *IEEE Access*, vol. 7, pp. 12177-12191, 2019.

[70] Y. Lecun, Y. Bengio, G. Hinton, "Deep learning", *Nature*, pp. 436-444, 2015.

[71] P. Pandey and K. R. Seeja, "Subject-independent emotion detection from EEG signals using deep neural network," *International Conference on Innovative Computing and Communications*, Springer Singapore, pp. 41–46, 2019.

[72] S. Jirayucharoensak, S. Pan-Ngum and P. Israsena, "EEG-based emotion recognition using deep learning network with principal component-based covariate shift adaptation," *The Scientific World Journal*, 2014.

[73] J. Zhang, M. Chen, S. Zhao, S. Hu, Z. Shi and Y. Cao, "Relief-based EEG sensor selection methods for emotion recognition," *Sensors*, vol. 16, no. 10, p. 1558, 2016.

[74] F. Shen, G. Dai, G. Lin, J. Zhang, W. Kong and H. Zeng, "EEG-based emotion recognition using 4D convolutional recurrent neural network," *Cognitive Neurodynamics*, vol. 14, pp. 815–828, 2020.

[75] Y. Li, J. Huang, H. Zhou, and N. Zhong, "Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks," *Applied Sciences*, vol. 7, no. 10, 2017.

[76] J. Yang, X. Huang, H. Wu, and X. Yang, "EEG-based emotion classification based on bidirectional long short-term memory network," *Procedia Computer Science*, pp. 491-504, 2020.

[77] The McGill Physiology Virtual Lab, Available online: https://www.medicine.mcgill.ca/physio/vlab/biomed_signals/eeg_n.htm (accessed on 9 February 2020).

[78] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 327-339, 2014.

[79] R. Nawaz, K.H. Cheah, H. Nisar, and V.V. Yap, "Comparison of different feature extraction methods for EEG-based emotion recognition," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 3, pp.910-926, 2020.

[80] L. Shi, Y. Jiao and B. Lu, "Differential entropy feature for EEG-based vigilance estimation," *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6627-6630, 2013.

[81] Y. Li, L. Wang, W. Zheng, Y. Zong, L. Qi, Z. Cui, T. Zhang, and T. Song, "A Novel Bi-Hemispheric Discrepancy Model for EEG Emotion Recognition," *IEEE Trans. Cogn. Develop.*, vol. 13, no. 2, pp. 354-367, 2021.

[82] Y. Yang, Q. Wu, Y. Fu, and X. Chen, "Continuous convolutional neural network with 3D input for EEG-based emotion recognition," *In International Conference on Neural Information Processing*, pp. 433-443. Springer, Cham, 2018.

[83] W. L. Zheng, and B.L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on autonomous mental development*, vol. 7, no. 3, pp.162-175, 2015.

[84] D.W. Chen, R. Miao, W.Q. Yang, Y. Liang, H.H. Chen, L. Huang, C.J. Deng, and N. Han, "A feature extraction method based on differential entropy and linear discriminant analysis for emotion recognition," *Sensors*, vol. 19, no. 7, p.1631, 2019.

[85] J. Cheng et al., "Emotion recognition from multi-channel EEG via deep forest," *in IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 453-464, Feb. 2021.

[86] H. Cui, A. Liu, X. Zhang, X. Chen, K. Wang, and X. Chen, "EEG-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network," *Knowl. Based. Syst.*, vol. 205, no. 106243, 2020.

[87] D. Maheshwari, S. K. Ghosh, R. K. Tripathy, M. Sharma, and U. R. Acharya, "Automated accurate emotion recognition system using rhythm-specific deep convolutional neural network technique with multi-channel EEG signals," *Computers in Biology and Medicine*, 134, 104428, 2021.

[88] W. Tao, C. Li, R. Song, J. Cheng, Y. Liu, F. Wan, and X. Chen, "EEG-based emotion recognition via channel-wise attention and self-attention," *IEEE Transactions on Affective Computing*, 2020.

[89] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[90] T. Song, W. Zheng, P. Song and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *in IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 532-541, July-Sept. 2020.

[91] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans. Affect. Comput.*, 2020.

[92] G. Zhang, M. Yu, Y.J. Liu, G. Zhao, D. Zhang, and W. Zheng, "SparseDGCNN: recognizing emotion from multichannel EEG signals," *IEEE Transactions on Affective Computing*, 2021

[93] Y. Yin, X. Zheng, B. Hu, Y. Zhang, and X. Cui, "EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM," *Applied Soft Computing*, p.106954, 2021.

[94] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging based bidirectional LSTM-RNNs for multi-modal emotion recognition," *Information Processing & Management*, vol. 57, no. 3, p.102185, 2020.

[95] X. Du, C. Ma, G. Zhang, J. Li, Y.K. Lai, G. Zhao, X. Deng, Y.J. Liu, and H. Wang, "An efficient LSTM network for emotion recognition from multichannel EEG signal," *IEEE Transactions on Affective Computing*, 2020

[96] Y. Yang, Q. Wu, M. Qiu, Y. Wang, and X. Chen, "Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network," *In 2018 international joint con-ference on neural networks (IJCNN)*, pp. 1-7, 2018.

[97] G. Zhang, and A. Etemad. "Deep recurrent semi-supervised EEG representation learning for emotion recognition," *In 2021 9th International Conference on Affective Computing and Intelligent In-teraction (ACII)*, pp. 1-8, 2021.

[98] J. Liu, G. Wu, Y. Luo, S. Qiu, S. Yang, W. Li, Y. Bi, "EEG-based emotion classification using a deep neural network and sparse autoencoder," *Frontiers in Systems Neuroscience*, 14, 43, 2020.

[99] A. S. Rajpoot, and M. R. Panicker, "Subject independent emotion recognition using EEG signals employing attention driven neural networks," *Biomedical Signal Processing and Control,* 75, 103547, 2022.

[100] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, "Deep feature learning for medical image analysis with convolutional autoencoder neural network," *IEEE Transactions on Big Data,* vol. 7, no. 4, pp.750-758, 2017.

[101] X. Chai, Q. Wang, Y. Zhao, X. Liu, O. Bai, and Y. Li, "Unsupervised domain adaptation techniques based on auto-encoder for non-stationary EEG-based emotion recognition," *Computers in biology and medicine,* pp.205-214, 2016.

[102] P. H. Le-Khac, G. Healy and A. F. Smeaton, "Contrastive representation learning: A framework and review," *in IEEE Access,* vol. 8, pp. 193907-193934, 2020.

[103] A. Saeed, D. Grangier and N. Zeghidour, "Contrastive learning of general-purpose audio representations," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics,* Speech and Signal Pro-cessing (ICASSP), pp. 3875-3879, 2021.

[104] P. Li, J. Wang, Y. Qiao, H. Chen, Y. Yu, X. Yao, P. Gao, G. Xie, and S. Song, "An effective self-supervised framework for learning expressive molecular global representations to drug discovery," *Brief. Bioinformatics,* vol. 22, no. 6, 2021.

[105] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," *In Proc. Conf. North American Chapter of the Association for Computational Linguistics (NAACL),* pp. 4171- 4186, 2019.

[106] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *In International conference on machine learning,* pp. 1597-1607, 2020

[107] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems,* pp.18661-18673, 2020.

[108] A. Jaiswal, A.R. Babu, M.Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies,* vol. 9, no. 1, p.2, 2020.

[109] M.N. Mohsenvand, M.R. Izadi, and P. Maes, 2020, "Contrastive representation learning for electroencephalogram classification," *In Machine Learning for Health,* pp. 238-253, 2020.

[110] X. Shen, X. Liu, X. Hu, D. Zhang, and S. Song, "Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition," *IEEE Transactions on Affective Computing,* 2022.

[111] F. Wang, S.H. Zhong, J. Peng, J. Jiang, and Y. Liu, "Data augmentation for EEG-based emotion recognition with deep convolutional neural networks," *In MultiMedia Modeling: 24th International Conference (MMM),* Bangkok, Thailand, February 5-7, pp. 82-93, 2018.

[112] E.S. Salama, R.A. El-Khoribi, M.E. Shoman, and M.A.W. Shalaby, "EEG-based emotion recognition using 3D convolutional neural networks," *International Journal of Advanced Computer Science and Applications,* vol. 9, no. 8, 2018.

[113] R. Shankar, A.H. Kenfack, A. Somayazulu, and A. Venkataraman, "A comparative study of data augmentation techniques for deep learning based emotion recognition," *arXiv preprint arXiv:2211.05047,* 2022.

[114] K.G. Hartmann, R.T. Schirrmeister, and T. Ball, "EEG-GAN: generative adversarial networks for electroencephalograhic (EEG) brain signals," *arXiv preprint arXiv:1806.0187,* 2018.

[115] Y. Luo, L.Z. Zhu, Z.Y. Wan, and B.L. Lu, "Data augmentation for enhancing EEG-based emotion recognition with deep generative models," *Journal of Neural Engineering*, vol. 17, no. 5, p.056021, 2020.

[116] Y. Luo, L.Z. Zhu, Z.Y. Wan, and B.L. Lu, "A GAN-based data augmentation method for multimodal emotion recognition," *In International Symposium on Neural Networks*, pp. 141-150, Springer, Cham, 2019.

[117] Y. Luo, and B.L. Lu, "EEG data augmentation for emotion recognition using a conditional Wasserstein GAN," *In 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pp. 2535-2538, 2018.

[118] D. Berthelot, T. Schumm, L. Metz, "BEGAN: boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717,* 2017.

[119] A. Mehrabian, A., "Communication without words," *In Communication theory*, pp. 193-200, Routledge, 2017.

[120] W. Mellouk, and W. Handouzi, "Facial emotion recognition using deep learning: review and insights," *Procedia Computer Science,* 175, pp.689-694, 2020.

[121] P. Tarnowski, M. Kołodziej, A. Majkowski, and R.J. Rak, "Emotion recognition using facial expressions," *Procedia Computer Science,* 108, pp.1175-1184, 2017.

[122] Z. Yu, G. Liu, Q. Liu, and J. Deng, "Spatio-temporal convolutional features with nested LSTM for facial expression recognition," *Neurocomputing*, 317, pp.50-57, 2018.

[123] D.K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognition Letters,* 120, pp.69-74, 2019.

[124] U.M. Lpez-Gil, J. Virgili-Gom, R. Gil, T. Guilera, I. Batalla, J. Soler-Gonzlez, and R. Garca, "Method for improving EEG based emotion recognition by combining it with synchronized biometric and eye tracking technologies in a non-invasive and low cost way," *Frontiers in Computational Neuroscience,* 2016.

[125] H. Al Osman, T.H. Falk, "Multimodal affect recognition: current approaches and challenges," *in Emotion and Attention Recognition Based on Biological Signals and Images (InTech),* pp. 59–86, 2017.

[126] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE,* vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[127] A. Metallinou, A. Katsamanis, and S. Narayanan, "A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs," *in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2012, pp. 2401–2404

[128] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: databases, features, and data fusion strategies," *APSIPA Trans. signal Inf. Process.*, vol. 3, 2014.

[129] S. Chen and Q. Jin, "Multi-modal conditional attention fusion for dimensional emotion prediction," i*n Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 571–575.

[130] J.-S. Lee and C. H. Park, "Robust audio-visual speech recognition based on late integration," *IEEE Trans. Multimedia,* vol. 10, no. 5, pp. 767–779, Aug. 2008.

[131] Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, and X. Xue, "Exploring inter-feature and inter-class relationships with deep neural networks for video classification," *in Proceedings of the 22nd ACM international conference on Multimedia,* 2014, pp. 167–176.

[132] Y. Huang, J. Yang, P. Liao, and J. Pan, "Fusion of facial expressions and EEG for multimodal emotion recognition," *Computational Intelligence and Neuroscience,* vol. 2017, Article ID 2107451, 8 pages, 2017.

[133] X. Huang, J. Kortelainen, G. Zhao, X. Li, A. Moilanen, T. Seppänen, M. Pietikäinen, "Multi-modal emotion analysis from facial expressions and electroencephalogram," *Computer Vision and Image Understanding,* Volume 147, 2016, Pages 114-124.

[134] J. Guo, R. Zhou, L. Zhao and B. Lu, "Multimodal emotion recognition from eye image, eye movement and EEG using deep neural networks," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* Berlin, Germany, 2019, pp. 3071-3074.

[135] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, "Combining eye movements and EEG to enhance emotion recognition," *in International Joint Conference on Artificial Intelligence*, vol. 15, 2015, pp. 1170–1176.

[136] Q. Zhu, G. Lu, and J. Yan, "Valence-arousal model based emotion recognition using EEG, peripheral physiological signals and facial expression," *In Proceedings of the 4th International Conference on Machine Learning and Soft Computing (ICMLSC 2020),* Association for Computing Machinery, New York, NY, USA, 81–85, 2020.

[137] Y. Huang, J. Yang, S. Liu, and J. Pan, "Combining facial expressions and Electroencephalography to enhance emotion recognition," *Future Internet,* vol. 11, no. 5, p. 105, May 2019.

[138] M. Soleymani, S. Asghariesfeden, M. Pantic, Y. Fu, "Continuous emotion detection using EEG signals and facial expressions," *In Proceedings of the IEEE International Conference on Multimedia and Expo,* Chengdu, China, 14–18 July 2014, pp. 1–6.

[139] M. Ponti, Jr. M.P., "Combining classifiers: from the creation of ensembles to the decision fusion," *In Proceedings of the 2011 24th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T),* Alagoas, Brazil, 28–30 August 2011, pp. 1–10.

[140] Z. Wang, Y. Wang, C. Hu, Z. Yin and Y. Song, "Transformers for EEG-based emotion recognition: A hierarchical spatial information learning model," *in IEEE Sensors Journal,* vol. 22, no. 5, pp. 4359-4368, 1 March1, 2022.

[141] J.Y. Guo, Q. Cai, J.P. An, P.Y. Chen, C. Ma, J.H. Wan, and Z.K. Gao, "A Transformer based neural network for emotion recognition and visualizations of crucial EEG channels," *Physica A: Statistical Mechanics and its Applications,* 603, p.127700, 2022.

[142] C. Li, Z. Zhang, X. Zhang, G. Huang, Y. Liu and X. Chen, "EEG-based emotion recognition via transformer neural architecture search," *in IEEE Transactions on Industrial Informatics,* vol. 19, no. 4, pp. 6016-6025, April 2023.

[143] J. Liu, H. Wu, L. Zhang, and Y. Zhao, Y., "Spatial-temporal transformers for EEG emotion recognition," *In Proceedings of the 6th International Conference on Advances in Artificial Intelligence*, pp. 116-120, 2022.

[144] J. Huang, J. Tao, B. Liu, Z. Lian and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 3507-3511.

[145] H. Chao, L. Dong, Y. Liu, and B. Lu, "Emotion recognition from multiband EEG signals using CapsNet," *Sensors*, vol. 19, p. 2212, 2019.

[146] P. Bashivan, R. Irina, Y. Mohammed, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," *arXiv preprint arXiv:1511.06448, 2015.*

[147] M. A. Ozdemir, M. Degirmenci, E. Izci, and A. Akan, "EEG-based emotion recognition with deep convolutional neural networks," *Biomedical Engineering*, vol. 66, pp. 43-57, 2021.

[148] H. M. Shao, J. G. Wang, Y. Wang, Y. Yao, and J. Liu, "EEG-based emotion recognition with deep convolution neural network," *In 2019 IEEE 8th Data Driven Control and Learning Systems Conference (DDCLS)*, pp. 1225-1229, 2019.

[149] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, Dragomir D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1-9. 2015.

[150] Y. Lavinia, H. H. Vo and A. Verma, "Fusion based deep CNN for improved large-scale image action recognition," *2016 IEEE International Symposium on Multimedia (ISM)*, San Jose, CA, pp. 609-614, 2016.

[151] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[152] I. Ayoub, Master thesis, University of Ottawa, Ottawa, ON, Canada, 2019.

[153] C. E. Rasmussen, "Gaussian processes for machine learning," *Summer School on Machine Learning*, 2003.

[154] A. Karpati, "Convolutional neural networks for visual recognition," http:/cs231n.github.io/linear-classify/, 2016.

[155] R. Gupta, K. Laghari, and T. Falk, "Relevance vector classifier decision fusion and EEG graph-theoretic features for automatic affective state characterization," *Neurocomputing*, pp. 875-884, 2015.

[156] A. Paszke, et al. "Automatic differentiation in pytorch," 2017.

[157] M. Tan, and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *In International conference on machine learning*, pp. 6105-6114, 2019.

[158] Q. Lao, T. Fevens and B. Wang, "Leveraging disease progression learning for medical image recognition," *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Madrid, Spain, pp. 671-675, 2018.

[159] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE transactions on affective computing*, vol. 3, no. 1, pp.42-55, 2011.

[160] S. Katsigiannis, and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp.98-107, 2017.

[161] P. Viola, M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, 137–154, 2004.

[162] D. Hendrycks, and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415,* 2016.

[163] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," *Advances in neural information processing systems*, 34, pp.30392-30400, 2021.

[164] Y. Lu, H. Zhang, L. Shi, F. Yang, and J. Li, "Expression-EEG bimodal fusion emotion recognition method based on deep learning," *Computational and Mathematical Methods in Medicine,* pp.1-10, 2021.

[165] U. Chinta, J. Kalita and A. Atyabi, "Soft voting strategy for multi-modal emotion recognition using deep-learning- facial images and EEG," *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC),* Las Vegas, NV, USA, pp. 0738-0745, 2023.

Appendix A

The performance of the second proposed model based on the amount of artificially generated data appended to the training dataset for DEAP and MAHNOB databases over different nodes are presented in this section.

Table A.1. Performance evaluation on the volume of appended artificially generated EEG data in DEAP dataset over different nodes.

× 0									
Node	Valence		Arousal		Node	Valence		Arousal	
	Train	Test	Train	Test		Train	Test	Train	Test
1	47.66%	54.69%	49.22%	57.81%	17	49.80%	60.94%	54.69%	61.33%
2	57.32%	61.33%	**69.33%**	**62.10%**	18	85.84%	58.59%	60.45%	61.33%
3	76.56%	59.77%	59.67%	62.50%	19	50.88%	58.59%	82.23%	62.89%
4	57.32%	60.16%	81.45%	62.10%	20	50.68%	58.20%	59.77%	64.06%
5	48.83%	58.98%	57.52%	60.55%	21	56.54%	58.59%	57.52%	60.94%
6	77.93%	61.33%	51.46%	61.33%	22	55.27%	58.98%	57.71%	61.33%
7	56.05%	59.38%	60.55%	61.33%	23	56.25%	59.38%	61.13%	62.10%
8	52.05%	58.98%	56.74%	60.55%	24	87.99%	59.77%	57.71%	60.55%
9	66.50%	58.98%	57.13%	60.94%	25	52.15%	58.20%	47.66%	60.94%
10	55.66%	58.20%	57.42%	60.55%	26	86.23%	60.55%	58.98%	62.89%
11	57.81%	58.59%	61.04%	63.28%	27	90.33%	60.16%	57.71%	60.55%
12	54.30%	58.59%	58.69%	60.94%	28	**66.01%**	**59.37%**	72.95%	60.94%
13	56.54%	58.98%	55.27%	61.72%	29	57.71%	58.59%	61.33%	62.10%
14	52.15%	58.98%	57.32%	60.94%	30	53.22%	58.20%	63.96%	64.06%
15	53.91%	59.77%	55.66%	61.72%	31	53.81%	59.38%	58.20%	61.72%
16	55.86%	58.98%	59.77%	61.72%	32	61.04%	60.16%	70.51%	62.10%

Table A.2. Performance evaluation on the volume of appended artificially generated EEG data in DEAP dataset over different nodes.

× 0.5									
Node	Valence		Arousal		Node	Valence		Arousal	
	Train	Test	Train	Test		Train	Test	Train	Test
1	52.67%	58.20%	55.21%	62.50%	17	89.78%	60.16%	52.02%	60.94%
2	80.21%	63.67%	60.09%	62.10%	18	50.00%	58.59%	55.40%	61.33%
3	85.16%	62.50%	54.56%	61.33%	19	84.96%	63.67%	55.92%	63.67%
4	53.84%	58.20%	56.64%	63.67%	20	91.86%	60.55%	60.74%	62.50%
5	86.85%	60.16%	51.50%	60.94%	21	52.28%	58.20%	55.14%	60.94%
6	51.30%	60.16%	52.21%	61.72%	22	88.41%	60.94%	52.80%	60.55%
7	88.80%	60.94%	53.26%	62.10%	23	80.73%	61.33%	53.84%	60.55%
8	85.55%	60.55%	52.02%	60.94%	24	68.95%	58.98%	50.65%	61.72%
9	80.47%	60.94%	54.95%	60.94%	25	71.03%	59.38%	58.72%	62.50%
10	91.08%	59.38%	53.06%	60.55%	26	90.04%	62.10%	50.91%	62.50%
11	86.72%	62.89%	57.42%	61.72%	27	89.00%	60.55%	52.93%	60.55%
12	**72.39%**	**62.10%**	51.30%	60.94%	28	92.19%	60.55%	60.68%	62.89%
13	74.87%	62.89%	52.41%	60.55%	29	56.45%	58.98%	59.90%	64.06%
14	52.60%	58.98%	54.17%	60.55%	30	51.56%	58.20%	50.78%	61.72%
15	89.13%	64.06%	55.27%	60.94%	31	84.51%	63.67%	55.92%	61.72%
16	54.30%	59.38%	**70.89%**	**63.67%**	32	55.79%	58.59%	53.84%	60.94%

Table A.3. Performance evaluation on the volume of appended artificially generated EEG data in DEAP dataset over different nodes.

× 1									
Node	Valence		Arousal		Node	Valence		Arousal	
	Train	Test	Train	Test		Train	Test	Train	Test
1	89.01%	61.33%	47.27%	39.45%	17	89.50%	62.50%	60.01%	60.94%
2	85.01%	59.77%	53.61%	60.94%	18	60.64%	60.16%	55.81%	61.33%
3	88.72%	58.98%	54.15%	61.72%	19	88.13%	60.94%	57.03%	63.67%
4	78.81%	61.33%	53.71%	61.72%	20	51.76%	58.20%	**74.21%**	**66.40%**
5	85.64%	62.10%	50.15%	60.55%	21	50.63%	58.20%	69.34%	64.06%
6	88.72%	59.38%	53.27%	61.72%	22	**74.65%**	**64.84%**	52.54%	62.50%
7	87.40%	60.16%	56.64%	62.10%	23	57.47%	62.50%	53.86%	61.33%
8	89.65%	58.59%	54.88%	60.94%	24	81.40%	60.94%	54.10%	60.55%
9	82.57%	62.50%	63.33%	63.67%	25	78.96%	57.81%	55.76%	60.94%
10	56.20%	59.38%	50.10%	60.55%	26	77.84%	64.06%	55.32%	62.89%
11	79.69%	59.38%	54.74%	62.10%	27	82.81%	58.59%	51.95%	60.55%
12	88.67%	63.67%	55.13%	60.94%	28	51.90%	58.59%	51.17%	60.55%
13	88.87%	61.72%	54.30%	60.94%	29	52.44%	58.59%	67.82%	61.33%
14	88.04%	61.72%	53.96%	61.33%	30	53.56%	58.59%	54.35%	63.28%
15	86.96%	60.55%	55.76%	62.10%	31	77.39%	59.38%	58.01%	62.10%
16	82.37%	62.10%	67.53%	62.50%	32	87.55%	60.94%	76.95%	60.55%

Table A.4. Performance evaluation on the volume of appended artificially generated EEG data in DEAP dataset over different nodes.

× 1.5									
Node	Valence		Arousal		Node	Valence		Arousal	
	Train	Test	Train	Test		Train	Test	Train	Test
1	84.84%	62.89%	54.06%	60.94%	17	88.05%	60.55%	52.27%	61.72%
2	89.26%	60.55%	50.63%	61.33%	18	**78.09%**	**63.28%**	55.51%	60.94%
3	88.13%	58.59%	52.81%	60.55%	19	86.25%	58.59%	57.81%	62.10%
4	86.88%	59.77%	53.83%	60.94%	20	89.77%	60.16%	55.70%	63.67%
5	88.87%	59.38%	49.14%	60.55%	21	89.02%	60.16%	55.43%	60.94%
6	83.44%	60.16%	49.38%	62.10%	22	86.72%	61.33%	54.65%	62.10%
7	86.76%	59.38%	52.50%	60.55%	23	55.04%	61.33%	54.69%	63.28%
8	88.05%	58.98%	52.54%	60.94%	24	89.30%	60.94%	51.95%	60.94%
9	59.88%	60.94%	52.66%	61.33%	25	80.04%	60.16%	50.74%	62.10%
10	86.60%	62.10%	53.52%	60.94%	26	84.88%	59.77%	55.43%	63.67%
11	88.40%	60.94%	52.89%	61.33%	27	90.86%	59.38%	52.89%	60.55%
12	85.04%	61.72%	54.02%	60.94%	28	88.95%	62.50%	55.78%	61.72%
13	83.75%	59.77%	55.12%	60.94%	29	51.80%	58.20%	54.69%	63.28%
14	89.38%	59.38%	49.49%	60.55%	30	86.95%	62.10%	59.06%	64.06%
15	88.36%	62.50%	52.93%	60.94%	31	52.23%	58.59%	52.77%	60.55%
16	87.81%	62.10%	78.09%	64.84%	32	87.89%	61.33%	**73.63%**	**64.84%**

Table A.5. Performance evaluation on the volume of appended artificially generated EEG data in DEAP dataset over different nodes.

× 2									
Node	Valence		Arousal		Node	Valence		Arousal	
	Train	Test	Train	Test		Train	Test	Train	Test
1	47.69%	50.39%	58.07%	60.94%	17	89.36%	60.16%	52.28%	61.33%
2	89.84%	60.55%	53.35%	60.94%	18	51.99%	60.55%	50.29%	62.10%
3	86.88%	58.98%	52.15%	61.72%	19	88.48%	61.72%	56.67%	64.06%
4	89.16%	61.33%	51.37%	61.33%	20	89.10%	57.81%	52.73%	60.94%
5	81.18%	59.77%	54.56%	61.33%	21	88.02%	62.89%	54.00%	60.94%
6	90.89%	62.50%	51.50%	62.89%	22	89.03%	60.16%	54.49%	62.10%
7	77.11%	58.59%	51.50%	60.55%	23	66.83%	61.72%	52.99%	61.72%
8	88.93%	60.94%	52.77%	60.55%	24	89.42%	62.89%	53.19%	61.72%
9	88.93%	60.94%	55.01%	61.72%	25	86.82%	60.55%	49.51%	60.55%
10	85.77%	62.50%	52.28%	58.98%	26	88.57%	61.72%	53.71%	62.10%
11	86.69%	58.98%	54.10%	63.28%	27	86.62%	59.38%	78.26%	60.94%
12	88.57%	62.89%	52.38%	60.94%	28	85.74%	61.33%	50.03%	60.94%
13	88.93%	62.50%	51.27%	60.55%	29	52.18%	58.98%	56.22%	62.50%
14	89.13%	60.94%	53.48%	62.50%	30	81.54%	59.77%	76.95%	63.67%
15	88.57%	61.33%	52.25%	62.10%	31	88.02%	61.33%	51.01%	60.55%
16	**77.53%**	**64.84%**	**74.27%**	**64.45%**	32	88.25%	60.55%	67.09%	61.72%

Table A.6. Performance evaluation on the volume of appended artificially generated EEG data in MAHNOB-HCI dataset over different nodes.

× 0									
Node	Valence		Arousal		Node	Valence		Arousal	
	Train	Test	Train	Test		Train	Test	Train	Test
1	76.72%	68.87%	90.97%	61.32%	17	85.27%	66.98%	85.75%	60.38%
2	81.00%	62.26%	85.27%	63.20%	18	89.31%	67.92%	81.24%	61.32%
3	61.05%	64.15%	50.83%	56.60%	19	82.19%	59.43%	56.06%	67.92%
4	51.78%	58.49%	85.99%	62.26%	20	57.96%	65.09%	85.51%	60.38%
5	61.05%	62.26%	89.07%	62.26%	21	89.31%	70.75%	89.55%	61.32%
6	88.36%	66.98%	63.42%	57.55%	22	78.38%	59.43%	81.00%	63.20%
7	87.17%	70.75%	88.60%	61.32%	23	73.40%	68.87%	88.36%	62.26%
8	**72.68%**	**66.03%**	**73.15%**	**61.32%**	24	89.31%	63.20%	80.29%	63.20%
9	69.60%	63.20%	76.96%	61.32%	25	86.46%	64.15%	83.37%	58.49%
10	86.70%	67.92%	85.99%	65.09%	26	49.64%	60.38%	88.60%	67.92%
11	53.68%	60.38%	71.73%	59.43%	27	59.14%	64.15%	46.32%	66.98%
12	80.29%	57.55%	92.16%	60.38%	28	88.12%	63.20%	53.92%	54.72%
13	58.43%	66.03%	88.60%	58.49%	29	63.66%	66.03%	79.10%	62.26%
14	50.59%	59.43%	62.95%	66.98%	30	87.17%	68.87%	85.27%	54.72%
15	78.62%	62.26%	80.05%	51.89%	31	80.05%	69.81%	82.90%	58.49%
16	86.22%	62.26%	84.09%	60.38%	32	63.42%	64.15%	85.75%	61.32%

Table A.7. Performance evaluation on the volume of appended artificially generated EEG data in MAHNOB-HCI dataset over different nodes.

× 0.5									
Node	Valence		Arousal		Node	Valence		Arousal	
	Train	Test	Train	Test		Train	Test	Train	Test
1	85.10%	63.20%	89.06%	66.98%	17	76.23%	66.98%	55.63%	59.43%
2	88.27%	62.26%	72.90%	65.09%	18	55.31%	60.38%	90.33%	59.43%
3	91.60%	69.81%	53.57%	57.55%	19	51.35%	63.20%	88.43%	65.09%
4	60.38%	64.15%	51.82%	56.60%	20	87.16%	61.32%	80.35%	58.49%
5	90.81%	65.09%	79.24%	66.98%	21	**74.32%**	**66.98%**	90.02%	62.26%
6	52.77%	65.09%	55.63%	66.03%	22	87.96%	58.49%	67.51%	60.38%
7	91.44%	72.64%	61.81%	61.32%	23	49.92%	61.32%	56.89%	65.09%
8	90.49%	62.26%	90.49%	70.75%	24	93.03%	64.15%	83.04%	71.69%
9	46.43%	41.51%	63.87%	63.20%	25	86.85%	61.32%	91.92%	61.32%
10	90.49%	67.92%	63.55%	66.03%	26	82.09%	62.26%	88.91%	65.09%
11	55.15%	62.26%	52.46%	57.55%	27	52.77%	58.49%	51.66%	58.49%
12	90.17%	64.15%	89.06%	66.98%	28	87.96%	66.98%	87.80%	54.72%
13	51.03%	59.43%	79.24%	65.09%	29	55.94%	61.32%	82.25%	68.87%
14	57.69%	64.15%	87.00%	60.38%	30	71.47%	68.87%	77.50%	64.15%
15	51.35%	63.20%	64.34%	63.20%	31	88.27%	64.15%	**78.12%**	**71.69%**
16	76.70%	65.09%	52.14%	62.26%	32	89.06%	60.38%	74.80%	58.49%

Table A.8. Performance evaluation on the volume of appended artificially generated EEG data in MAHNOB-HCI dataset over different nodes.

× 1									
Node	Valence		Arousal		Node	Valence		Arousal	
	Train	Test	Train	Test		Train	Test	Train	Test
1	87.65%	63.20%	87.29%	65.09%	17	89.19%	67.92%	54.16%	58.49%
2	54.63%	60.38%	61.52%	66.03%	18	52.26%	63.20%	81.24%	57.55%
3	83.25%	60.38%	52.49%	57.55%	19	64.73%	66.03%	88.12%	61.32%
4	89.19%	62.26%	89.07%	60.38%	20	89.31%	63.20%	83.37%	60.38%
5	86.34%	61.32%	58.67%	65.09%	21	84.32%	63.20%	70.55%	61.32%
6	**72.92%**	**66.03%**	90.50%	64.15%	22	80.17%	62.26%	89.43%	66.03%
7	79.10%	61.32%	78.98%	65.09%	23	55.70%	61.32%	79.69%	66.98%
8	88.48%	64.15%	59.62%	65.09%	24	56.41%	60.38%	89.55%	68.87%
9	90.50%	66.03%	90.02%	63.20%	25	86.70%	61.32%	89.67%	64.15%
10	69.48%	66.03%	53.80%	63.20%	26	69.48%	65.09%	55.94%	62.26%
11	49.76%	60.38%	61.88%	57.55%	27	51.19%	60.38%	86.70%	62.26%
12	88.12%	65.09%	**71.25%**	**62.26%**	28	79.57%	63.20%	89.90%	65.09%
13	83.14%	63.20%	85.87%	66.03%	29	67.34%	66.03%	88.24%	66.03%
14	93.11%	62.26%	52.85%	61.32%	30	87.41%	63.20%	52.97%	65.09%
15	90.26%	61.32%	55.23%	60.38%	31	86.34%	62.26%	73.28%	57.55%
16	55.46%	60.38%	67.22%	66.98%	32	92.64%	62.26%	80.76%	63.20%

Table A.9. Performance evaluation on the volume of appended artificially generated EEG data in MAHNOB-HCI dataset over different nodes.

× 1.5									
Node	Valence		Arousal		Node	Valence		Arousal	
	Train	Test	Train	Test		Train	Test	Train	Test
1	91.92%	68.87%	92.40%	64.15%	17	89.92%	63.20%	64.35%	58.49%
2	91.35%	61.32%	62.45%	60.38%	18	89.16%	66.03%	72.53%	59.43%
3	66.06%	62.26%	61.12%	61.32%	19	54.94%	64.15%	66.35%	62.26%
4	84.51%	61.32%	58.75%	58.49%	20	**74.32%**	**63.20%**	85.46%	57.55%
5	91.92%	59.43%	60.08%	66.98%	21	55.13%	59.43%	80.51%	64.15%
6	85.08%	60.38%	52.09%	64.15%	22	50.48%	60.38%	84.41%	66.03%
7	58.94%	61.32%	91.06%	65.09%	23	87.74%	62.26%	61.50%	64.15%
8	83.27%	61.32%	86.60%	66.98%	24	65.49%	66.98%	83.94%	66.98%
9	68.35%	64.15%	62.07%	62.26%	25	86.50%	60.38%	64.92%	60.38%
10	80.89%	63.20%	79.94%	66.03%	26	92.68%	66.98%	84.98%	71.69%
11	80.32%	63.20%	88.88%	63.20%	27	57.32%	60.38%	59.98%	61.32%
12	90.30%	63.20%	74.52%	63.20%	28	59.22%	61.32%	77.57%	59.43%
13	47.34%	62.26%	85.46%	63.20%	29	62.55%	65.09%	84.98%	66.98%
14	88.21%	64.15%	91.06%	64.15%	30	58.08%	65.09%	**76.42%**	**67.92%**
15	49.81%	60.38%	54.56%	60.38%	31	92.02%	67.92%	69.58%	60.38%
16	52.85%	63.20%	69.20%	62.26%	32	56.84%	62.26%	51.33%	65.09%

Table A.10. Performance evaluation on the volume of appended artificially generated EEG data in MAHNOB-HCI dataset over different nodes.

× 2									
Node	Valence		Arousal		Node	Valence		Arousal	
	Train	Test	Train	Test		Train	Test	Train	Test
1	90.50%	65.09%	91.69%	68.87%	17	52.89%	59.43%	61.92%	61.32%
2	90.02%	60.38%	81.24%	64.15%	18	54.95%	66.03%	90.58%	61.32%
3	87.89%	62.26%	83.06%	59.43%	19	89.39%	61.32%	90.74%	63.20%
4	56.22%	62.26%	90.34%	62.26%	20	54.79%	60.38%	89.63%	61.32%
5	55.03%	65.09%	67.46%	62.26%	21	50.36%	62.26%	82.66%	65.09%
6	90.10%	60.38%	49.56%	66.03%	22	52.57%	61.32%	73.56%	62.26%
7	50.91%	58.49%	63.50%	61.32%	23	62.87%	60.38%	86.70%	67.92%
8	58.99%	60.38%	68.33%	66.03%	24	62.95%	63.20%	72.60%	65.09%
9	57.40%	61.32%	58.04%	62.26%	25	55.50%	61.32%	88.92%	65.09%
10	52.18%	63.20%	65.72%	66.98%	26	59.14%	63.20%	81.55%	68.87%
11	57.96%	62.26%	52.18%	58.49%	27	76.78%	56.60%	84.32%	61.32%
12	52.02%	58.49%	80.21%	65.09%	28	50.20%	58.49%	**74.50%**	**64.15%**
13	51.54%	60.38%	88.99%	62.26%	29	80.76%	58.49%	91.45%	68.87%
14	54.63%	62.26%	78.54%	65.09%	30	52.26%	63.20%	87.02%	66.98%
15	91.05%	57.55%	86.14%	65.09%	31	**77.88%**	**64.15%**	50.36%	58.49%
16	91.13%	63.20%	87.17%	60.38%	32	53.92%	63.20%	88.44%	65.09%