FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 5, May, 2025, pp 277 - 290
DOI: https://doi.org/10.33003/fjs-2025-0905-3412

MULTI-MODAL EMOTION RECOGNITION MODEL USING GENERATIVE ADVERSARIAL NETWORKS (GANs) FOR AUGMENTING FACIAL EXPRESSIONS AND PHYSIOLOGICAL SIGNALS

Abya Newton Hegh, *Adekunle Adedotun Adeyelu, Aamo Iorliam and Samera U Otor

Department of Mathematics/Computer Science, Benue State University Makurdi, Nigeria.

*Corresponding authors' email: adeyeluadekunle@yahoo.com

ABSTRACT

Emotion recognition is a critical area of research with applications in healthcare, human-computer interaction (HCI), security, and entertainment. This study addressed the limitations of single-modal emotion recognition systems by developing a multi-modal emotion recognition model that integrates facial expressions and physiological signals, enhanced by Generative Adversarial Networks (GANs). It aims at improving accuracy, reliability, and robustness in emotion detection, particularly underrepresented emotions. The study utilized the FER-2013 dataset for facial expressions and the DEAP dataset for physiological signals. GANs were employed to augment datasets, address class imbalances and enhance feature diversity. A hybrid multi-modal model was developed, combining Convolutional Neural Networks (CNNs) for facial expression recognition and Long Short-Term Memory (LSTM) networks for physiological signal analysis. Hybrid fusion was used to integrate features at multiple levels, maximizing the complementary strengths of each modality. The results demonstrate significant improvements in emotion recognition. Without GAN augmentation, the CNN and LSTM models achieved accuracies of 62% and 76%, respectively. The hybrid model outperformed, gaining 90% across all metrics. With GAN-augmented datasets, the CNN and LSTM models improved to 81% and 86%, respectively, while the hybrid (multi-modal) model achieved state-of-the-art performance with 93% accuracy and an F1-score of 92%. These findings underscore the efficacy of GANs in enhancing data diversity and the advantages of multi-modal integration for robust emotion recognition. The study contributes to knowledge by introducing a GAN-augmented hybrid multi-modal framework, advancing methodologies in emotion recognition. Recommendations for future work include addressing ethical considerations in emotion recognition systems.

Keywords: Multimodal Emotion Recognition, Deep Learning, Facial Expression Analysis, Generative Adversarial Networks (GANs), Feature Fusion, Time-Series Classification

INTRODUCTION

Emotion recognition involves identifying and categorizing human emotions through indicators like facial expressions, voice, physiological signals, and body language (Ullah et al., 2021). It is a key area within affective computing, which seeks to simulate human emotions. Accurate emotion recognition is crucial for human interaction, enhancing empathy and understanding. Current innovations in ML and AI have notably enhanced the accuracy and efficiency of these systems (Siddiqui et al., 2022).

Emotion recognition has diverse applications across various fields. In healthcare, it helps diagnose and monitor conditions such as depression, anxiety, and autism and provides real-time emotional feedback in telemedicine (Abdulyekeen, 2025; Younis et al., 2024). In Human-Computer Interaction (HCI), it improves user experience by enabling adaptive systems such as educational platforms and customer service bots that respond to users' emotions (Muhammad et al., 2023). In security, it aids in identifying suspicious behaviour by analysing facial expressions and body language, enhancing public safety (Wei, 2024). In entertainment, it creates immersive experiences by adjusting game dynamics and enhancing interactive storytelling based on players' emotions (Siddiqui et al., 2022).

Despite progress, emotion recognition faces challenges due to variations in emotional expression influenced by cultural, social, and personal factors, making universal models difficult to develop (Jianhua et al., 2020). Recognizing subtle and complex emotions is also challenging. Collecting and annotating large emotional datasets is resource-intensive, and physiological signals can be affected by non-emotional factors, complicating the process (Alharbawee & Pugeault, 2024). Achieving real-time recognition with low latency is technically demanding, especially with multi-modal data.

Multi-modal approaches are increasingly necessary, as single-modal systems often fail to represent the full range of human emotions (Zheng et al., 2019). Multi-modal systems combine data from different sources like physiological indicators and facial emotions, providing a more accurate and thorough comprehension of emotions (Jang et al., 2019). These systems are more robust and reliable, compensating for the limitations of individual modalities, and creating a greater contextual grasp of emotions (Siddiqui et al., 2022; Younis et al., 2024).

The concept of Generative Adversarial Networks (GANs), initiated by Ian Goodfellow in 2014 (Alharbawee & Pugeault, 2024; Khan & Sarkar, 2022; Yan et al., 2021), comprise a discriminator, concurrently trained by a generator through adversarial processes (Alharbawee & Pugeault, 2024). GANs have been successful in applications like image generation and augmentation of data. When recognising emotions, GANs enhance system robustness and accuracy by generating synthetic emotional expressions and physiological signals, addressing the challenge of limited annotated data (Taisheng et al., 2020). They improve the ability to recognize subtle emotions and facilitate cross-modal translation, enriching the emotion recognition system's interpretability and completeness (Soleimani, 2024).

Integrating facial expressions and physiological signals, this approach seeks to create a strong and comprehensive method for the identification of emotions with wide-ranging applications. The goal of this study is to build a multi-modal emotion recognition model using integrated facial expressions and physiological indicators, thereby enhancing the accuracy, reliability, and applicability of emotion detection across various domains. The specific contributions of this study are as follows: It introduces a multi-modal approach using GAN-based architecture to augment facial expressions and physiological signals for emotion recognition. It also demonstrates the capability of GANs in addressing class imbalances and improving generalizability for underrepresented emotions. Furthermore, the study develops a multi-modal emotion identification model that effectively capture subtle and complex emotions; and finally, it proposes a GAN-augmented hybrid multi-modal framework that advances existing methodologies in emotion recognition.

The paper is structured into six main sections. Section 2 provides the background knowledge of existing studies in the area, establishing the knowledge gap. Section 3 delves into the materials and methods to bridge the identified research gap. In section 4, the results of the implementation are presented, while section 5 discusses the results and their implications. Section 6 concludes by summarising the study findings with recommendations for future work.

Literature Background

Emotion identification has garnered substantial attention in recent times owing to its wide range of applications in fields such as human-computer interaction, mental health monitoring, and adaptive learning systems. Traditional methods using unimodal data like facial expressions or physiological signals often miss the complexity of human emotions. Familiarizing with this existing body of knowledge is imperative to guide this study on “multimodal emotion recognition using GAN for facial expressions and physiological signals”. Thus, this section, critically examined state-of-the-art studies, methodologies, and discoveries about emotion recognition. The empirical review pinpoints research gaps eminent in the existing literature that this study seeks to address. In so doing, offers valuable insights and contributes to the cumulative knowledge base in this particular field.

Consequently, Bao et al. (2024) introduced a novel model for emotion recognition. It combines eye movement and video optical flow to indicate attention and also measures the speed of image changes. Convolutional Neural Network (CNN) was employed to extract deep features, which are then used to categorize emotions into interest, happiness, confusion, and boredom. The single-modal models yielded 64.32%, 74.67%, and 71.88% while decision-level fusion attained the best accuracy at 81.90% using a synthetic dataset. Soleimani (2024) leveraged several Deep Learning methods including CNN, RNN, GAN and Autoencoder to detect human emotional state. Experimentally, the results using the DEAP dataset show that the hybrid model achieved 65% and 68% accuracies for recognising valence and arousal emotions, respectively. Aside the the hybrid model, the study also developed a framework tagged “Contrastive Learning GAN-based Graph Neural Network” for identifying emotions from Electroencephalogram (EEG) signals. Results using the DEAP and MAHNOB datasets indicate that the DEAP dataset yielded 64% and 66% for valence and arousal emotion classification accuracies. The MAHNOB outperformed with 66% and 71% for the valence and arousal emotion classification, respectively. Zhang et al. (2024) created a multimodal emotion identification algorithm to identify learners' emotional states by combining physiological data and semantic information from videos. The outcomes of the trial demonstrate that the model greatly enhanced the ability to recognise emotions yielding 82.30% accuracy using the Video Learning Multimodal Emotion Dataset (VLMED).

Ali & Hughes (2023) designed a model tagged Unified Biosensor-Vision Multi-modal Transformer-based (UBVMT) for the classification of arousal-valence emotional states. Experimental evaluations using the MAHNOB and DEAP datasets indicate that the UBVMT model outclassed existing solutions with 50.01% and 83.84% accuracies for recognising valence and arousal respectively using the MAHNOB dataset. On the DEAP dataset, the model achieved 81.53% and 82.64% accuracies for valence and arousal emotion detection, respectively. Win et al. (2023) leveraged CNN with 3-layers and simply tagged the model 3B-Convnet model for emotion recognition. The model was evaluated using the Extended Cohn-Kanade and Japanese Female Facial Expressions dataset. Experimental results showed that the model can recognize the emotional state of compound facial expressions with an accuracy of 67.51% and 62.87% for the two datasets respectively. Muhammad et al. (2023) utilized the Deep Canonical Correlation Analysis (DCCA) based multimodal emotion recognition technique to combine electroencephalography (EEG) and facial video clips and build a multimodal framework for emotion recognition. CNN was utilized for feature extraction. Evaluations using the MAHNOB and DEAP showed 93.86% and 91.54% accuracy for the MAHNOB and DEAP datasets, respectively.

Sung-Nien et al. (2022) devised an emotion recognition scheme using ResNet, bidirectional long and short-term memory (BiLSTM) modules. Deep Convolutional Gen-Erative Adversarial Network (DCGAN) was employed for data augmentation with photoplethysmography (PPG) signals as input data. The emotions detected in the study include neutral, angry, happy, and sad emotional states with 90.34% and 86.32% for two- and four-class detection rates, respectively. Zhong et al. (2022) presented the Regularized Graph Neural Network (RGNN) for emotion identification using EEG data. An adjacency matrix derived from brain topology and neuroscience principles is used by RGNN to simulate inter-channel interactions. NodeDAT and EmotionDL, two regularizers, are included to handle noisy labelling and cross-subject variances. The better performance of RGNN is demonstrated through testing on the SEED and SEED-IV datasets. Experimental results have it that the model achieved 74.96% and 73.84% on SEED and SEED-IV datasets, respectively. Ma et al. (2022) devised an approach for a Multimodal conditional Generative Adversarial Network (GAN) used for data augmentation in audio-visual emotion recognition experiments. The system includes both audio and visual modalities generators and discriminators, sharing category information as an input to generate a variety of synthetic data. Hirschfeld-Gebelein-Rényi (HGR) maximum correlation is used to describe the dependency between the audio and visual modalities in the produced data to closely resemble real data. This synthetic data improves the data manifold and aids in resolving problems related to class imbalance. This approach used a multimodal conditional GAN for audio-visual emotion identification for the first time. The eNTERFACE'05, RAVDESS, and CMEW datasets were used in experiments, and the results were 49.48%, 65.90% and 46.19% respectively for eNTERFACE'05, RAVDESS, and CMEW datasets.

Zhang et al. (2021) used feature-level fusion, multiscale feature extraction, and hierarchical network structure, the study created a hierarchical fusion convolutional neural network model to mine data potential. The study assesses the efficacy of the model using binary classification trials on the valence and arousal dimensions of the DEAP and MAHNOB-HCI datasets. In terms of feature extraction and fusion, the findings demonstrate that the model outperformed other deep-learning emotion classification models, with accuracies of 84.71% and 89.00% on the two related data sets. Guangcheng et al. (2021) leveraged GAN to train the VAE-D2GAN data augmentation model for EEG-based emotion identification. Differential entropy (DE) topological maps are derived from EEG data that correspond to five classical frequency bands and reflect distinct emotions. This approach created synthetic training samples and learned the distributions of these characteristics for actual EEG data. To increase the variety of the simulated samples that are produced, the variational auto-encoder (VAE) architecture is incorporated into the dual discriminator GAN. The VAE design uses a latent vector to learn the geographical distribution of the real data. Evaluations using two datasets, the SEED and the SEED-IV, yielded 92.5% and 82.3% performance, respectively. Salama et al. (2021) built a novel multi-modal framework for human emotion identification that extracts spatio-temporal characteristics from human face video data and EEG signals using 3D-Convolutional Neural Networks (3D-CNN). The framework employed ensemble learning and data augmentation strategies to arrive at final fusion predictions. Three methods are established for emotion recognition: face-based, fusion-based, and EEG-based. While the face technique employs SVM classifiers and mask-RCNN for predictions, the EEG approach uses 3D-CNN. Techniques for bagging and stacking are tried for the fusion approach; stacking yields the highest accuracy. The framework outperforms existing multi-modal emotion identification techniques, with recognition accuracies of 96.13% for valence and 96.79% for arousal.

Cimtay et al. (2020) presented a unique approach for recognizing emotions based on a variety of modalities, such as electroencephalogram (EEG), galvanic skin response (GSR), and facial expressions. Utilizing a hybrid fusion approach, this technique produced a mean accuracy of 74.2% and a maximum one-subject-out accuracy of 81.2% for three different emotion classes (happy, neutral, and sad) using a synthetically generated multimodal emotion dataset (LUMED-2). On the Database for Emotion Analysis using Physiological Signals (DEAP), the method produced a mean accuracy of 53.8% and a maximum one-subject-out accuracy of 91.5% for varied numbers of emotional states. Hongli (2020) presented a deep automated encoder-based multi-modal emotion identification technique that combined EEG data with facial expressions. First, feature selection is done using a decision tree. Sparse representation is used to identify facial expression traits, which are then examined to categorize test samples. After merging facial expression and EEG data, the bimodal deep automatic encoder (BDAE) extracts features for supervised learning in the third layer. The classification task is finished by the LIBSVM classifier. The approach successfully extracts and combines high-level emotion-related characteristics, as demonstrated by experiments conducted on a created video library. The average emotion detection rate of 85.71% was attained, and the capacity to recognize emotions was greatly enhanced. Nakisa et al. (2020) devised a temporal multimodal fusion strategy using a deep learning model to capture non-linear emotional correlations inside and across EEG and blood volume pulse (BVP) signals. Both early fusion and late fusion techniques are used to assess the model's performance. In particular, after learning each modality independently, a convolutional neural network (ConvNet) long short-term memory (LSTM) model combined EEG and BVP inputs to learn and explore linked emotional representations across modalities. The model was evaluated using a dataset from smart wearable sensors. According to experimental data, human emotions are classified into four quadrants of dimensional emotions by temporal multimodal deep learning models utilizing early and late fusion techniques, with accuracies of 71.61% and 70.17%, respectively.

Song et al. (2019) built a database for physiological indicators that gathers four physiological signal types: breathing, galvanic skin reaction, EEG, and ECG. Through thorough labelling and psychological assessment, 28 movies were selected as standardized samples from a collection of over 1500 video clips to reduce cultural biases and successfully elicit desired emotions. While participants saw these movies, which depicted six distinct emotions and one neutral feeling, their physiological signals were monitored. Three distinct classification processes and a range of feature extraction techniques together with two classifiers (k-NN and SVM) were employed to identify emotional reactions and provide baseline data. To improve feature extraction, a novel Attention-Long Short-Term Memory (A-LSTM) model was utilized. Results indicated that the A-LSTM outperformed other models with accuracies of 41.88% and 42.10% respectively, using HHS and STFT as EEG features. Nemati et al. (2019) developed a hybrid multimodal data fusion approach in which a latent space linear map is utilized to fuse the audio and visual modalities, and an evidentiary fusion method based on the Dempster-Shafer (DS) theory is used to fuse the textual modality with its projected characteristics into the cross-modal space. The examination of the suggested technique using the DEAP dataset's videos demonstrates its advantages over non-latent space fusion methods as well as decision-level methods. Additionally, compared to canonical correlation analysis (CCA) and cross-modal factor analysis (CFA), the results show that feature-level audio-visual fusion improves better when using marginal fisher analysis (MFA). The model achieved 92% and 93% accuracies for audio-visual and text modalities respectively. Jang et al. (2019) assessed the validity of physiological alterations brought on by six fundamental emotions; happiness, sorrow, anger, fear, disgust, and surprise measured across a period of ten weeks. Pre- and during-emotion-provoking film clips were monitored physiologically in a group of twelve college students. After every movie, participants assessed their feelings. Ten distinct movie snippets of every emotion a total of 60 clips spread over ten weeks were employed to avoid adaption. Skin conductance level (SCL), fingertip temperature (FT), heart rate (HR), and blood volume pulse (BVP) were among the physiological characteristics that were retrieved. Results showed that Cronbach's alphas from emotion-provoking phases ranged from 0.39 to 0. 96.

Despite these advancements, many reviewed models rely on either unimodal data or simple fusion strategies, which limits their ability to accurately capture the complexity of human emotions. A recurring limitation across studies is the underperformance in recognising less prominent emotions due to dataset imbalances. While some works incorporate GANs, they are typically used in isolation for either image or signal augmentation not both simultaneously. Furthermore, most fusion approaches do not effectively preserve both spatial and temporal features from diverse modalities. This study bridges these gaps by proposing a GAN-augmented, hybrid multi-modal framework that integrates CNN and LSTM networks to capture both facial and physiological features. The model applies feature-level fusion and evaluates its performance on both original and augmented datasets, achieving higher accuracy in recognising subtle and underrepresented emotions.

MATERIALS AND METHODS

Deep learning techniques were utilised to recognize emotions using multimodal data: facial images and heart rate signals. The methodology consists of: (i) dataset source, (ii) preprocessing, (iii) feature extraction, (iv) synthetic data generation using Generative Adversarial Networks (GANs), (v) model development, (vi) multi-modal fusion and (vii) model evaluation. Figure 1 depicts the architectural diagram of the multi-modal emotion identification model.

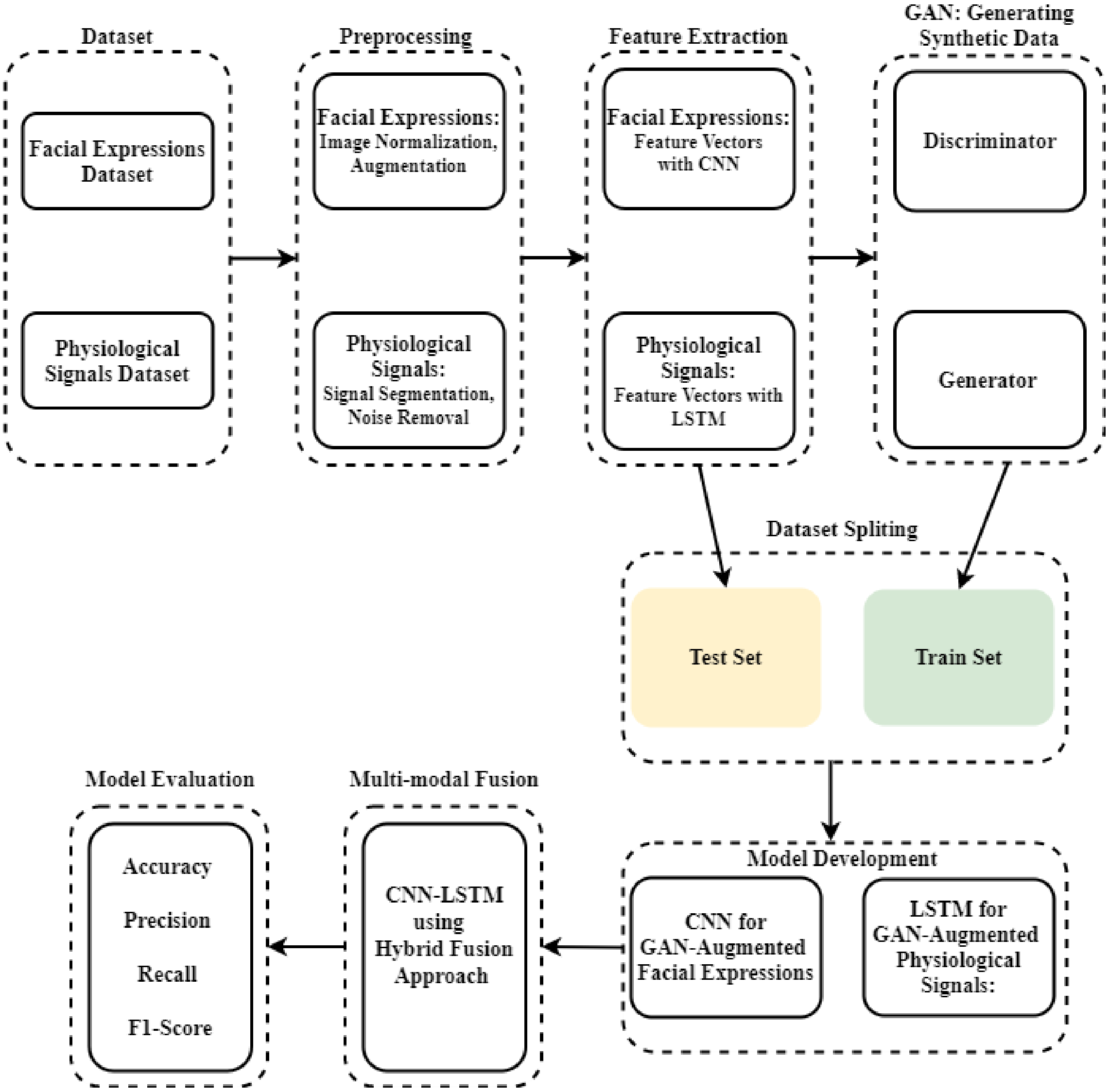

Dataset
Facial Expressions Dataset
Physiological Signals Dataset
Preprocessing
Facial Expressions: Image Normalization, Augmentation
Physiological Signals: Signal Segmentation, Noise Removal
Feature Extraction
Facial Expressions: Feature Vectors with CNN
Physiological Signals: Feature Vectors with LSTM
GAN: Generating Synthetic Data
Discriminator
Generator
Dataset Spliting
Test Set
Train Set
Model Development
CNN for GAN-Augmented Facial Expressions
LSTM for GAN-Augmented Physiological Signals:
Multi-modal Fusion
CNN-LSTM using Hybrid Fusion Approach
Model Evaluation
Accuracy
Precision
Recall
F1-Score

Figure 1: Architecture of the Model (Bao et al., 2024)

This study advances the unimodal CNN-based approach by Bao et al. (2024) by integrating LSTM networks for temporal analysis of physiological signals and employing GANs for data augmentation. Building on techniques from prior works, it proposes a novel hybrid CNN-LSTM architecture with feature-level fusion, resulting in a more robust and generalisable multi-modal emotion recognition system. The phases in the methodology are briefly discussed thus:

Dataset

Two types of datasets were utilized in this study: the Facial Expression Dataset and the Physiological Signal Dataset. The FER-2013 dataset contains 35,887 grayscale facial images with each image having a resolution of 48×48 pixels, representing facial expressions mapped to specific emotions. The images are mapped into seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. It is accessible at "https://www.kaggle.com/datasets/msambare/fer2013". The dataset is class-imbalanced, with some emotions being overrepresented while others are underrepresented. The physiological dataset utilized is the Dataset for Emotion Analysis using Physiological signals (DEAP). The DEAP dataset contains heart rate signals recorded from subjects experiencing different emotions, stored as time-series data, with each sequence representing a physiological response to an emotional stimulus. These include anger, disgust, fear, happiness, sadness, surprise, and neutral. It is available at "http://www.eecs.qmul.ac.uk/mmv/datasets/deap/".

Data Preprocessing

The facial expression images were standardized to ensure consistency across the dataset. This involves resizing all images to a uniform dimension, typically 48x48 pixels for compatibility with common neural network architectures. Additionally, pixel values were normalized between 0 to 1, which accelerates the convergence of the learning algorithm and enhances model performance. To boost the diversity of the training dataset and the generalisation capacity of the model, data augmentation approaches like flipping, rotation, and brightness adjustments were employed. The augmentation helps the model to become more robust to variations in facial expressions and environmental conditions, such as lighting and occlusions. Emotions were mapped (encoded) to numerical values accordingly to enhance classification. Physiological signals are often contaminated with noise and artefacts. Like the facial image dataset, this dataset also suffers from class imbalance, affecting model performance. To address this, all heart rate sequences were resampled to a fixed length. Min-max scaling was applied to standardize signal values and Signals were split into smaller overlapping time windows for better feature extraction. By applying these pre-processing techniques, the quality and consistency of the data were enhanced and ready for feature extraction.

Feature Extraction

CNNs were employed to automatically extract high-level features from facial images. These features are hierarchical, starting from low-level edges and textures to high-level facial components and expressions. Typically, the CNN architecture consists of fully linked, pooling, and convolutional layers, fine-tuned on the facial expression dataset to extract robust and discriminative features. In the physiological dataset, Long Short-Term Memory (LSTM) networks were used to extract the temporal dependencies in the physiological signals, capturing how heart rate variations correlate with emotions. Mean, variance, and frequency-domain features (Fourier Transform) were also extracted to enhance model understanding.

Data Generation with GAN

To address class imbalance, Generative Adversarial Networks (GANs) were used to generate synthetic facial images and heart rate signals. Deep Convolutional GAN (DCGAN) was trained on real facial images to generate synthetic samples. The GAN was trained using the underrepresented emotion classes to create new, high-quality images. Approximately 20% of the original dataset size was synthetically generated to balance class distributions. Similarly, Recurrent GAN (R-GAN) was used to synthesize new heart rate sequences by learning the temporal patterns of real signals.

Model Development

Three deep learning models: CNN, LSTM and a hybrid fusion of CNN-LSTM were implemented using the augmented and the original datasets. For the facial dataset, CNN was used for classification by learning hierarchical features from the pixel intensities. It comprised of 3 Convolutional Layers (ReLU activation, MaxPooling), Flatten & Fully Connected Layers with Softmax Classifier. LSTM was used for heart rate sequence classification by capturing temporal dependencies. Its architecture consisted of 2 Layers;

Dense Layer with Softmax Activation. Lastly, the Hybrid CNN-LSTM Model integrated CNN (for images) and LSTM (for heart rate) to make emotion predictions based on multimodal inputs. The extracted CNN features were combined with LSTM outputs in a dense fusion layer before final classification. Categorical Cross-Entropy (loss function) was used for multi-class classification. While Adam optimizer with an adaptive learning rate, Dropout of 0.3 and batch normalization were applied to prevent overfitting.

All the models were trained on both original and GAN-augmented datasets. Early stopping was implemented to avoid overfitting.

Multi-Modal Fusion Module

The Multi-Modal Fusion Module integrates spatial and temporal features extracted from Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, respectively. This fusion strategy enhances the model's ability to capture both visual and physiological cues for improved emotion recognition. The hybrid feature-level fusion approach was employed, where feature vectors from both modalities were concatenated before classification. This method ensures that both spatial and temporal patterns are preserved in a unified representation. Specifically, CNN architecture was used to extract spatial features from facial images. Mathematically, given X_{images} as inputs, the CNN transformation can be formulated as:

$$F_{CNN} = \text{CNN}(X_{image}) \tag{1}$$

Where; F_{CNN} is the image feature vectors extracted from the CNN.

Similarly, LSTMs were used to extract temporal features from heart rate signals. Mathematically, given X_{HR} as inputs, the LSTM transformation can be written as:

$$F_{LSTM} = \text{LSTM}(X_{HR}) \tag{2}$$

Where; F_{LSTM} is the feature vector extracted from the LSTM.

After extracting feature vectors from both CNN and LSTM, hybrid feature fusion was performed by concatenating F_{CNN} and F_{LSTM} into a single feature vector:

$$F_{fused} = \text{Concat.}(F_{CNN}, F_{LSTM}) \tag{3}$$

Where;

$F_{CNN} = R^{d1}$ represents the spatial feature vector.

$F_{LSTM} = R^{d2}$ represents the temporal feature vector.

$F_{fused} = R^{d1+d2}$ is the final fused feature vector used for classification.

Classification Layer

The fused feature vector F_{fused} was then passed through a fully connected (dense) layer, followed by a Softmax classifier to predict the emotion category:

$$\bar{y} = Softmax\,(W.F_{fused} + b) \tag{4}$$

Where;

W represents the weight of the matrix for classification, b is the bias term and ȳ is the predicted emotion category. This hybrid approach preserves the high-dimensional information from both modalities, enabling the model to make more informed predictions.

Model Evaluation

To assess the performance of the multi-modal emotion recognition system, several standard performance metrics will be used:

Accuracy: Assesses the ratio of correctly accurately identified emotions out of the total emotions. It is a basic metric for overall performance but is only suitable for a balanced dataset (Eke et al., 2021). The mathematical representation of accuracy is given as;

$$Acy = \frac{T_P+T_N}{T_P+T_N+F_P+F_N} \tag{5}$$

Where; T_P= True Positive, T_N= True Negative, F_N= False Negative, and F_P= False Positive.

Precision: Indicates the proportion of true positive predicted emotions out of all positive predicted emotions. It is useful for evaluating the exactness of positive recognition (Kwaghtyo & Eke, 2022). Precision is written statistically as follows:

$$Prec. = \frac{T_P}{T_P+F_P} \tag{6}$$

Recall: Takes account of the ratio of true positive detected emotions out of all actual positive emotions. It is important to assess how well the model recognizes positive emotions. The mathematical representation of recall is:

$$Rec. = \frac{T_P}{T_P + F_N} \quad (7)$$

F-Score: The harmonic mean provides a single metric that strikes a balance between recall and precision. Statistically, it is expressed as:

$$F1 - score = 2 \times \frac{P \times R}{P + R} \quad (8)$$

Where; P = Precision; R = Recall.

RESULTS AND DISCUSSION

In this section, the results of the implementation are presented. The results is presented in two phases. Phase 1 presents the evalution results with the original dataset and the second phase focuses on the evaluation result with the GAN-augmented dataset. Additionally, the performance of the study is further compared first with the utilised models and to existing emotion recognition models.

Results Phase 1: Evaluation With Original Datasets

In this section, the results of the experiments conducted without augmenting the datasets using GAN are presented.

Evaluation of the CNN Model for Facial Expression Recognition

The CNN model trained using the original dataset achieved moderate accuracy, demonstrating challenges in detecting certain emotions fear, sad and neutral as shown in Figure 2.

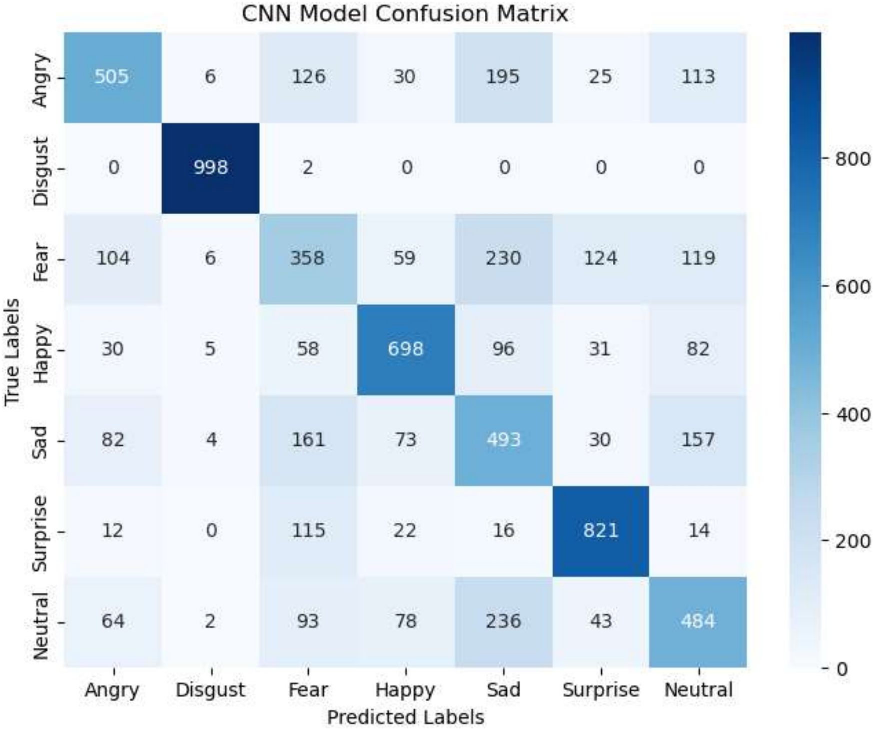


Figure 2: Confusion Matrix for CNN Model

The CNN model achieved a moderate and promising performance result of 62% accuracy, with 62% F1-score, Precision 63%, and 62% Recall. Figure 3 demonstrates the model's training and validation accuracy/loss over 100 epochs. The training loss reduced steadily, but the validation accuracy plateaued early, indicating potential overfitting.

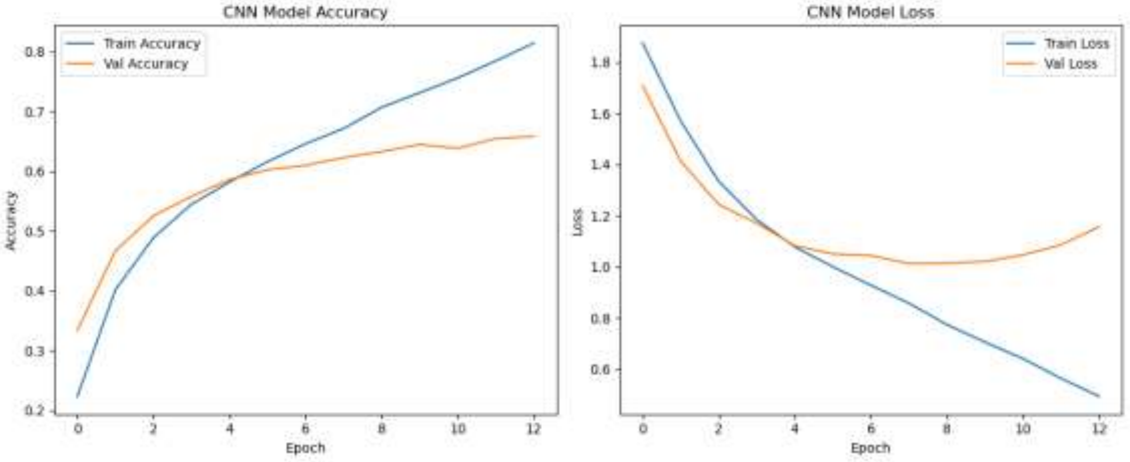


Figure 3: Training Accuracy Vs Loss for the CNN Model

Evaluation of the LSTM Model for Physiological Signal Analysis

The LSTM model showed moderate performance, particularly in detecting emotions like sadness and fear, that is class 3 and class 4. Misclassifications were observed to be high in class 0 and 5 of the emotional categories as evident in Figure 4.

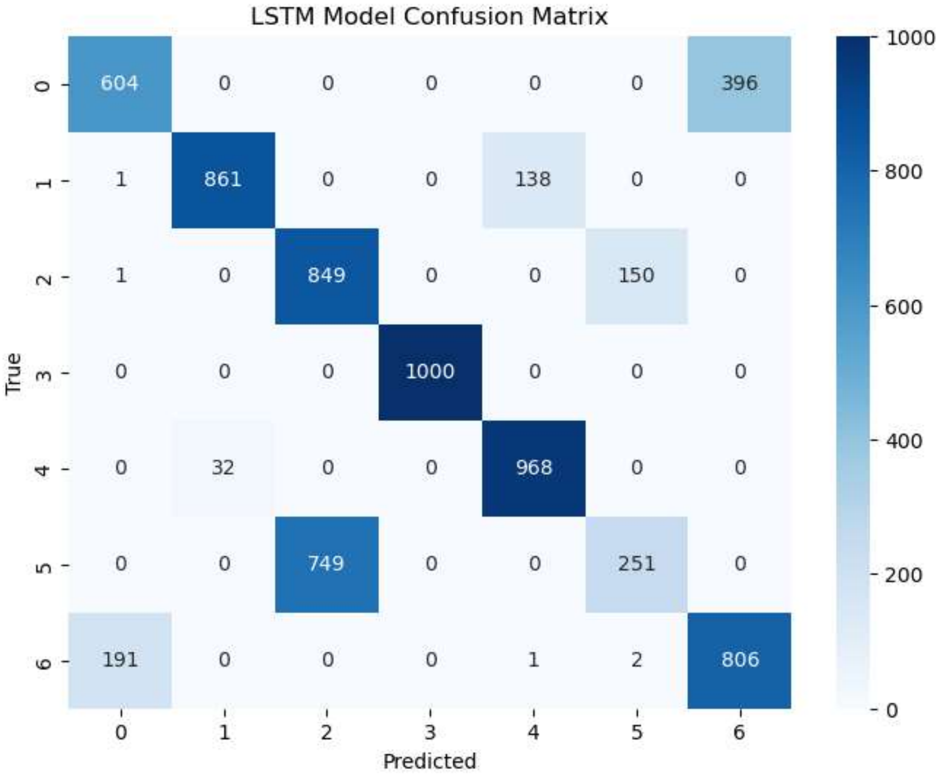


Figure 4: The Confusion Matrix for LSTM Model

Performance metrics for the LSTM model without data augmentation yielded 76% accuracy, F1-Score of 75%, Precision achieved 77% and Recall attained 76%. Figure 5 shows the training accuracy and loss trends for the LSTM model.

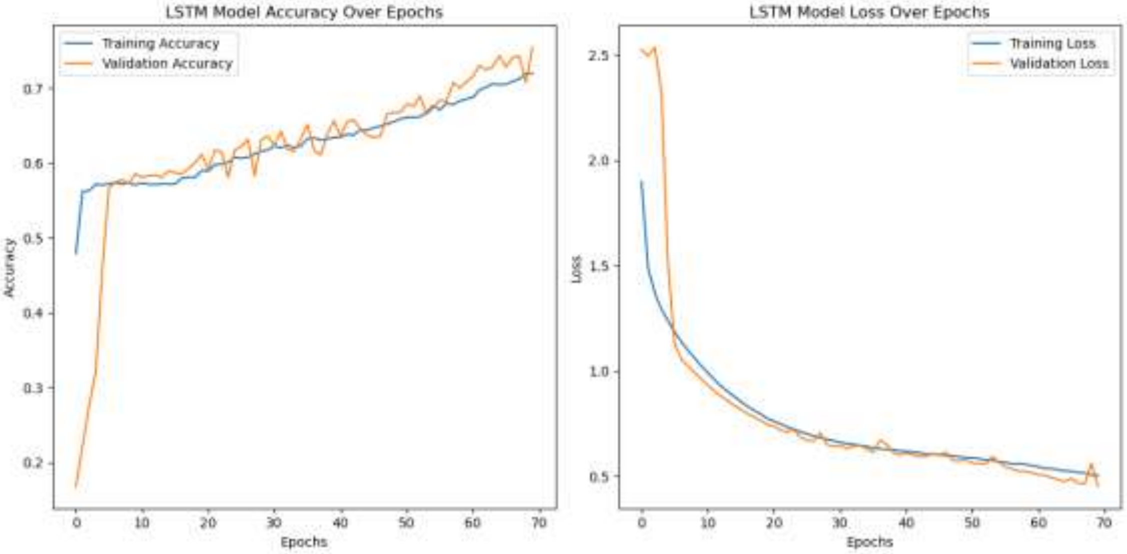
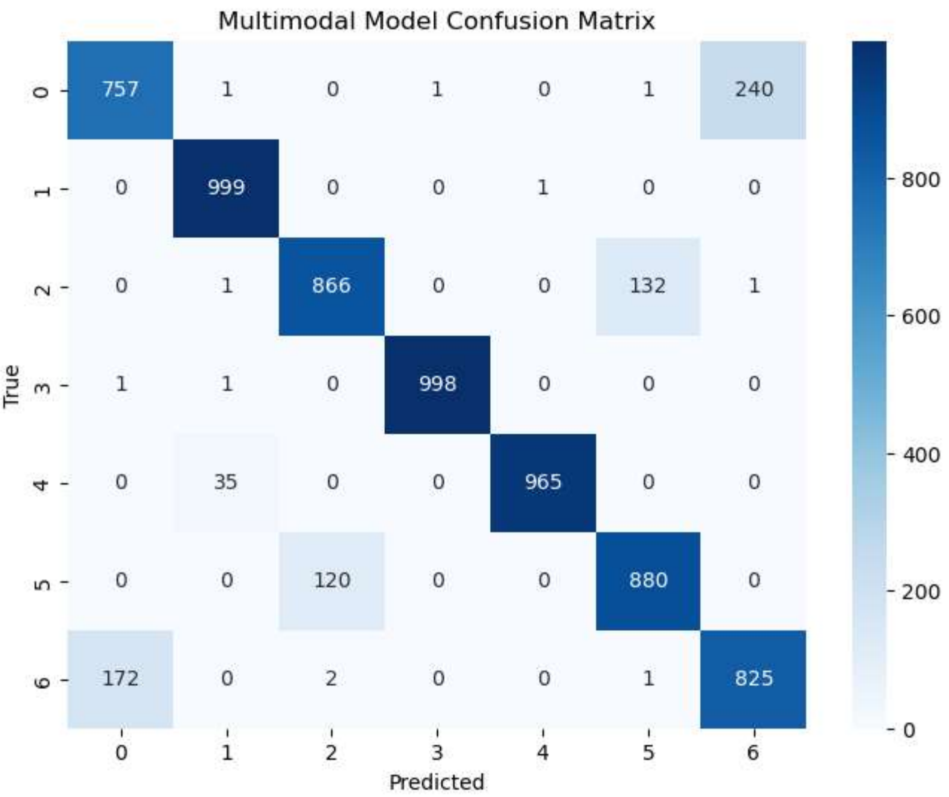


Figure 5: Training Accuracy Vs Loss for the LSTM Model

Evaluation of the Hybrid Multi-Modal Model

The hybrid model, combining CNN and LSTM outputs, outperformed the individual models demonstrating improved accuracy across all emotion categories. Figure 6 depicts the confusion matrix for the hybrid model.



	0	1	2	3	4	5	6
0	757	1	0	1	0	1	240
1	0	999	0	0	1	0	0
2	0	1	866	0	0	132	1
3	1	1	0	998	0	0	0
4	0	35	0	0	965	0	0
5	0	0	120	0	0	880	0
6	172	0	2	0	0	1	825

Figure 6: Confusion Matrix for the Hybrid Model

The multi-modal model outperformed the individual models having achieved a flat result of 90% across accuracy, F1-score, Precision and Recall metrics. Figure 7 illustrates the hybrid model's training and validation accuracy/loss over 100 epochs. The two graphs show the training loss and validation accuracy trends.

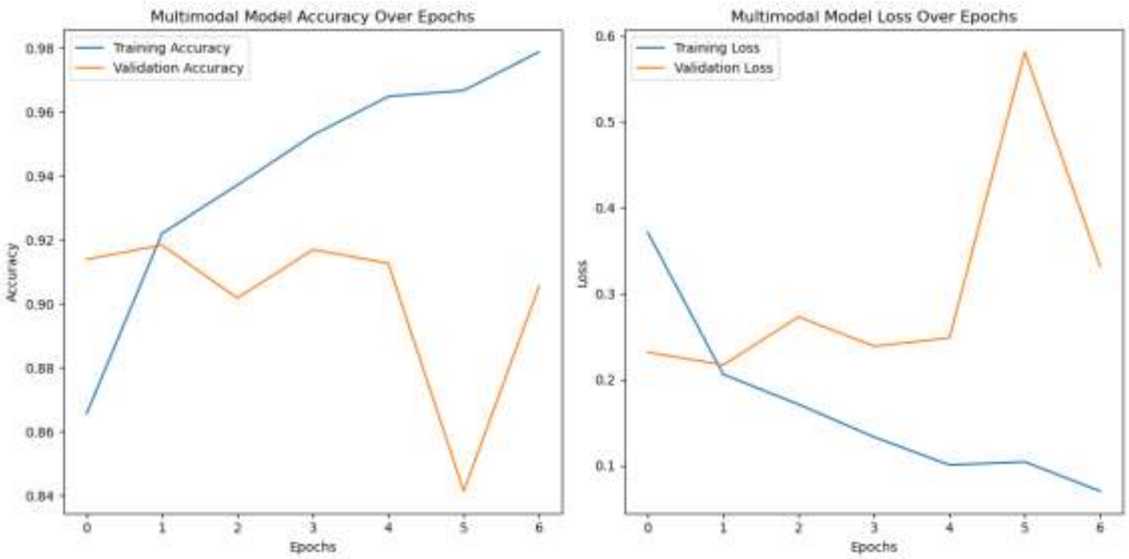


Figure 7: Training Accuracy Vs Loss for the Hybrid Model

Results Phase 2: Evaluation With GAN-Augmented Datasets

This section presents the experimental outcomes of the study conducted using the GAN-augmented datasets.

Evaluation of the CNN Model with GAN Augmented Facial Data

The CNN model when trained on GAN-augmented facial expression data showed significant improvement. Figure 8 illustrates the confusion matrix for the GAN-augmented CNN model.

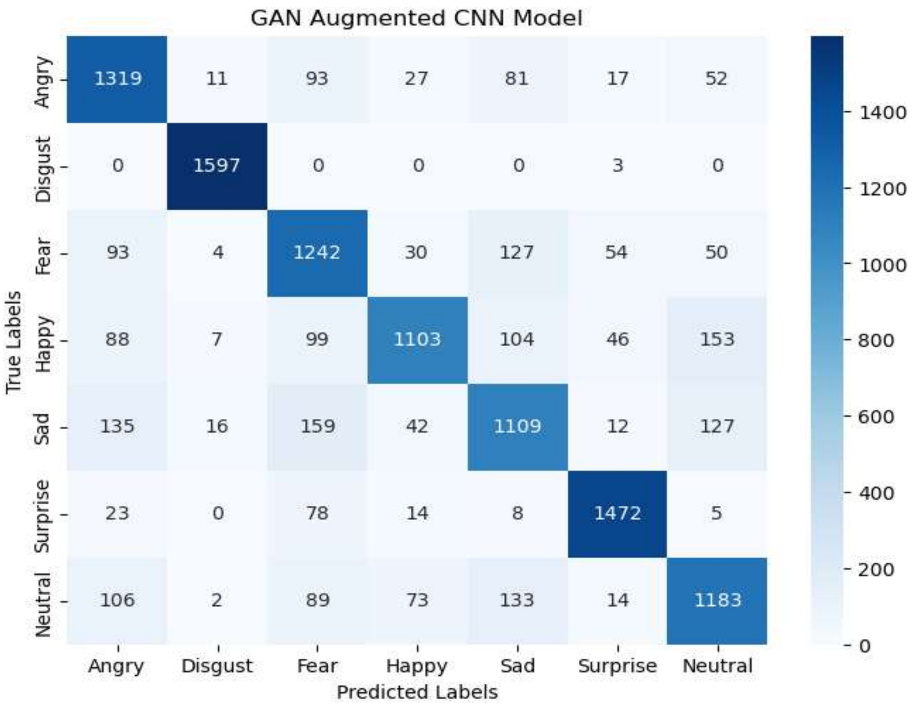


Figure 8: The Confusion Matrix for the GAN Augmented CNN Model

The GAN-augmented CNN model outclassed the CNN model trained with the original. The GAN augmented model yielded 81% across all the utilised metrics (accuracy, F1-score, Precision and Recall). Figure 9 shows the training and validation accuracy/loss over 100 epochs trends for the GAN-augmented CNN model.

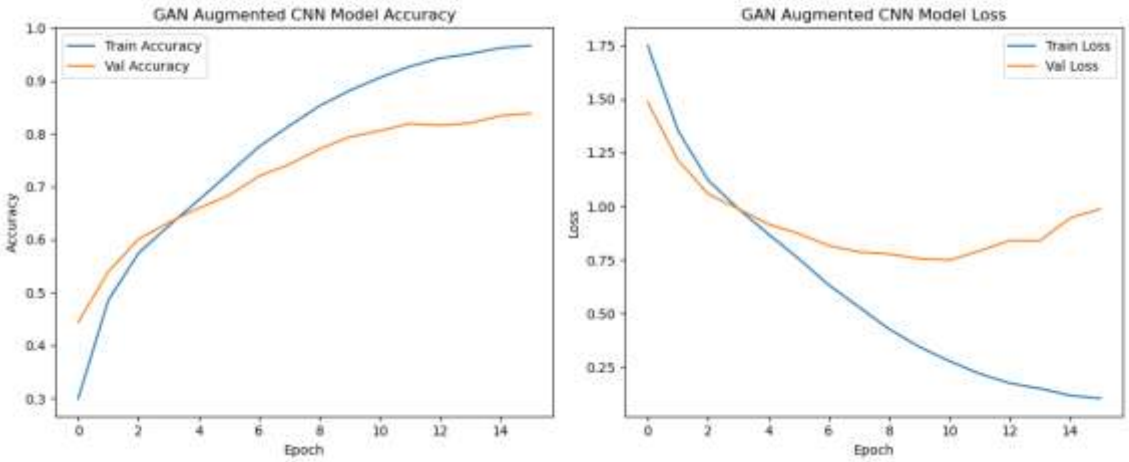


Figure 9: Training Accuracy Vs Loss for the GAN-Augmented CNN Model

Evaluation of the LSTM Model with GAN Augmented Physiological Data

The LSTM model achieved substantial improvements after training with GAN-augmented physiological signals. Figure 10 presents the confusion matrix for the GAN-augmented LSTM model.

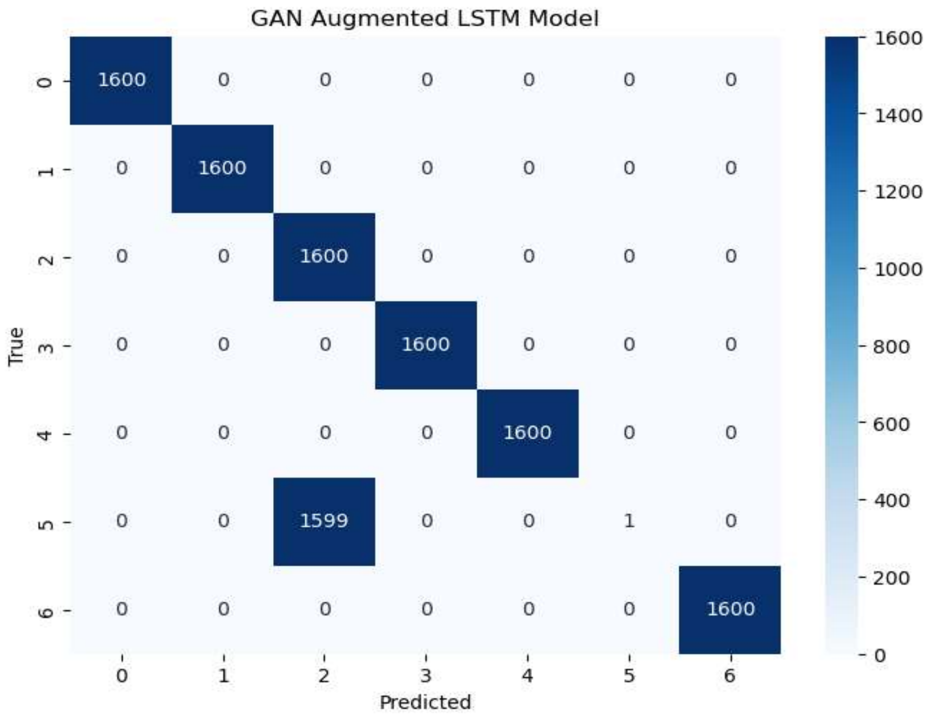


Figure 10: Confusion Matrix for the GAN-Augmented LSTM Model

The performance metrics for the GAN-augmented LSTM model yielded 86% accuracy, F1-Score of 86%, Precision achieved 93% and Recall attained 86%. Figure 11 shows the training trends for the GAN-augmented LSTM model.

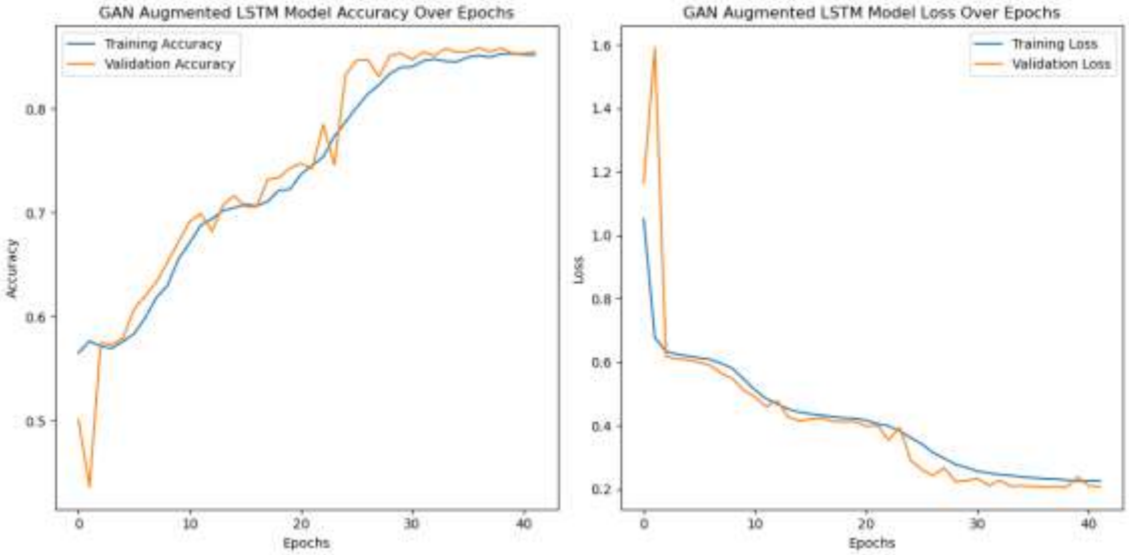


Figure 11: Training Accuracy vs. Loss for the GAN-Augmented LSTM Model

Evaluation of the GAN Augmented Hybrid Model

The hybrid model benefited the most from GAN augmentation, leveraging complementary features from both modalities. Figure 12 shows the confusion matrix for the GAN-augmented hybrid model.

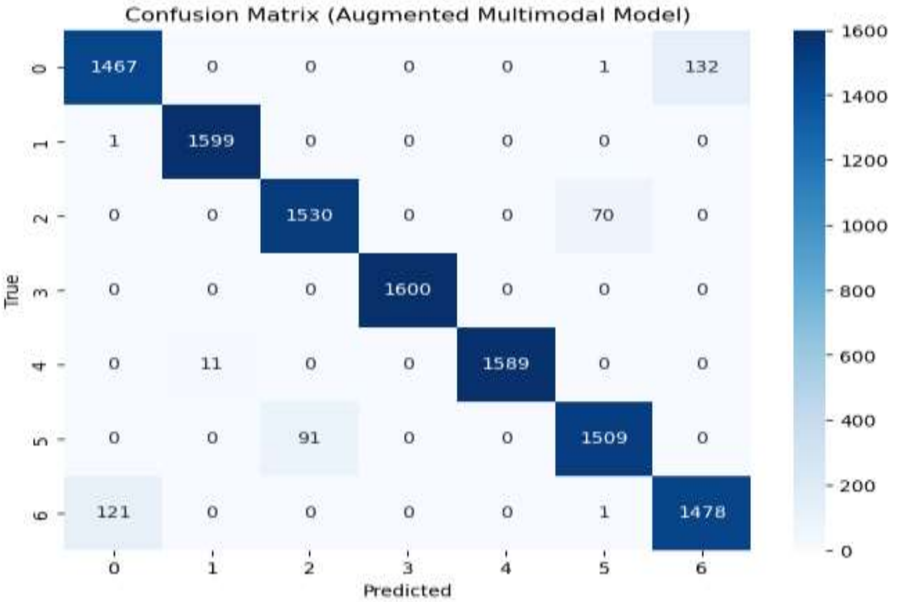


Figure 12: Confusion Matrix for GAN-Augmented Hybrid Model

The performance metrics after data augmentation with GAN outperformed all other individual models including the hybrid model without GAN augmentation. With GAN augmentation, the hybrid model achieved 93% accuracy, 92% F1-score, Precision 94%, and 91% Recall. Figure 13 demonstrates the hybrid model's improved training patterns.

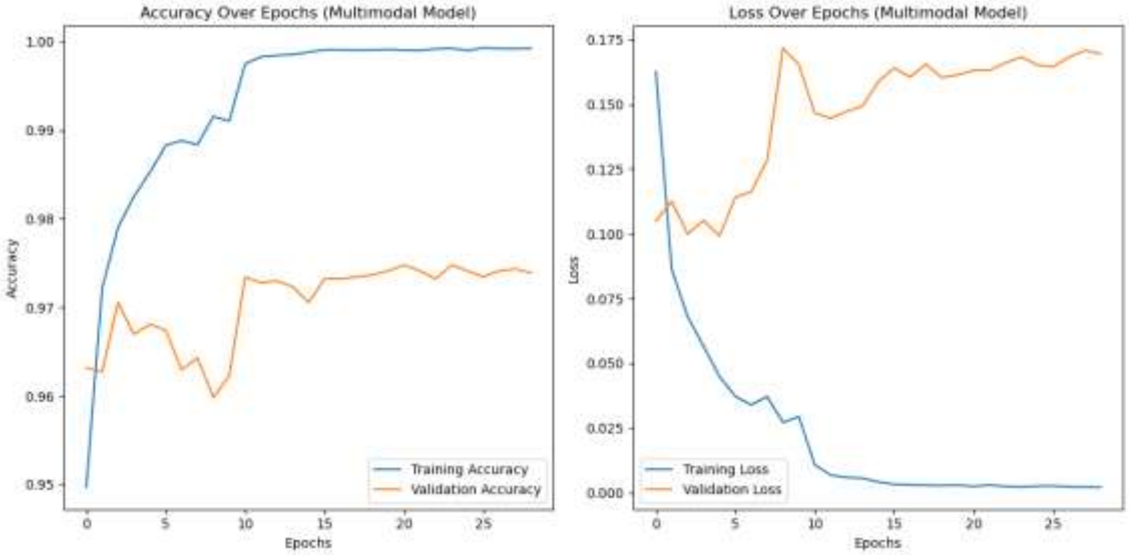


Figure 13: Training Accuracy vs. Loss for the GAN-Augmented Hybrid Model

Performance Comparison

The performance of the utilised models was compared using accuracy, precision, recall, and F1-score across the three main frameworks: CNN (for facial expressions), LSTM (for physiological signals), and the Hybrid Multi-Modal model. The metrics were compared for both evaluation phases: Without GAN-augmented datasets (original data only); and with GAN-augmented datasets (enhanced data for training). This comparison highlights the impact of GAN augmentation and the performance advantages of integrating facial and physiological signals. The results are presented in Table 1.

Table 1: Performance Comparison of Utilised Models

Metrics	Models	Without GAN (%)	With GAN (%)	Improvement (%)
Accuracy	CNN	62	81	19
	LSTM	76	86	10
	Hybrid Model	90	92	02
Precision	CNN	63	81	18
	LSTM	77	93	16
	Hybrid Model	90	92	02

Recall	CNN	62	81	19
	LSTM	76	86	10
	Hybrid Model	90	92	02
F1-Score	CNN	62	81	19
	LSTM	75	86	09
	Hybrid Model	90	92	02

GAN-augmented datasets significantly improved all models, with the CNN and LSTM showing accuracy gains of 19% and 10%, respectively, particularly in recognizing underrepresented emotions like fear and disgust. The hybrid model outperformed single-modal approaches, achieving the highest accuracy of 92% across all the utilised evaluation metrics through integrating complementary features from facial expressions and physiological signals. Overall, GAN augmentation addressed dataset imbalances, enhanced precision and recall, and improved the robustness and generalizability of the models, establishing the hybrid multi-modal architecture as a state-of-the-art solution for emotion recognition.

When compared to three existing works, the hybrid model demonstrated superior performance over these methods using accuracy, precision, recall, and F1 metrics. Specifically, Bao et al. (2024) achieved 81.9% accuracy with decision-level fusion, and Ali & Hughes (2023) reported a precision of 83.84% using a transformer-based model. Ma et al. (2022) achieved a recall of 65.9% with GAN-augmented data. The hybrid model's recall of 92% and its ability to handle subtle emotions underscore its advancements over these methods. A detailed comparison is presented in Table 2.

Table 2: Performance Comparison with the Existing Models

Metric	Bao et al. (2024)	Ali & Hughes (2023)	Ma et al. (2022)	Hybrid Model
Accuracy	81.9%	-	-	92%
Precision	-	83.84%	-	92%
Recall	-	-	65.9%	92%
F1-score	-	-	-	92%

As evidenced in Table 2, the hybrid multi-modal model demonstrates significant advancements over existing works. It achieved an overall performance of 92% accuracy, outperforming Bao et al. (2024) by 10.1%. Its precision of 92% surpassed Ali & Hughes (2023) by 8.16%, reflecting reduced false positives. With a recall of 92%, the hybridized multi-modal model outperformed Ma et al. (2022) by 26.1%, excelling in recognizing subtle emotions. An F1 score of 92% further highlights its balanced and robust performance. These results underscore the hybrid model's effectiveness in integrating GAN-augmented data, addressing dataset imbalances, and advancing research in the field of emotion recognition or computer vision at a broader look.

The observed performance improvement up to 26.1% in recall over Ma et al. (2022) can be attributed to the dual-modality approach combined with GAN-based augmentation. Unlike Ma et al. (2022), who focused on audio-visual data and applied augmentation on a single modality, this model enriches both facial and physiological datasets, which enhances representation for underrepresented emotional states like fear and disgust. This dual-augmentation strategy, coupled with hybrid fusion, enables better generalisation and more comprehensive emotion classification. Furthermore, some existing models underperform due to the use of simpler network architectures and lack of attention to class imbalance, which this study addresses effectively.

Discussion

The CNN model demonstrated moderate performance when trained on the original dataset, achieving an accuracy of 62%. It faced challenges in correctly classifying subtle emotions like fear and disgust, resulting in frequent misclassifications as seen in Figure 2. This limitation highlights the difficulty of relying solely on facial expression data, particularly for underrepresented emotions. The LSTM model, trained on physiological signals, performed better than the CNN, with an accuracy of 76%. Its ability to capture temporal dynamics from physiological signals improved recognition of emotions like sadness. However, high misclassification rates for categories like neutral and surprise revealed the limitations of using physiological signals alone for nuanced emotions. The hybrid model integrating CNN and LSTM outputs achieved significantly better performance, with an 80% score across all metrics. This improvement underscores the effectiveness of multi-modal integration, which leverages complementary features from facial and physiological data, reducing misclassifications observed in the single-modal models.

With GAN augmentation, the CNN model's accuracy increased to 81%, showing a 19% improvement. This enhancement demonstrates GANs' ability to generate diverse and representative facial expressions, addressing class imbalances and improving the model's generalizability. GAN-augmented physiological signals led to a 10% improvement in the LSTM model's accuracy, reaching 86%. The augmented dataset allowed the model to better capture subtle variations in physiological signals, enhancing recognition of emotions like fear and sadness. The hybrid model, benefiting from GAN-augmented data, achieved the highest overall performance with a 93% accuracy and an F1-score of 92%. The hybrid architecture effectively combined the strengths of both modalities, while the GAN-generated data enhanced the robustness and reliability of the system.

The hybrid multi-modal model demonstrates significant advancements over existing methods in emotion recognition. Bao et al. (2024) achieved an accuracy of 81.9% using decision-level fusion, while this study achieved 92%, representing an 11.1% improvement. Similarly, Zhang et al. (2021) reported an F1-score of 89% using a hierarchical fusion method, which the hybrid model outperformed with an F1-score of 92%. Furthermore, Ali & Hughes (2023) achieved a precision of 83.84% using a transformer-based model, whereas the hybrid model's precision of 92% demonstrates a 10.16% improvement. The significant recall gain of 26.2% over Ma et al. (2022), which employed GANs with single-modal inputs, highlights the advantage of multi-modal integration. Finally, Soleimani et al. (2020) reported an accuracy of 78.5%, further emphasizing the superior performance of the multi-modal method, especially when addressing class imbalances in underrepresented emotions such as fear and disgust. These comparisons illustrate the hybrid model's ability to surpass existing benchmarks, primarily due to its effective combination of GAN-augmented datasets and multi-modal architecture. This positions the study as a key contributor to advancing emotion recognition technologies.

CONCLUSION

This study successfully demonstrated the transformative potential of Generative Adversarial Networks (GANs) for augmenting datasets and the power of multi-modal integration in enhancing emotion recognition systems. The study achieved a state-of-the-art performance with the hybrid model attaining 93% accuracy, 92% F1-score, 94% precision, and 91% recall, significantly outperforming earlier models such as Bao et al. (2024) (accuracy: 81.9%) and Ma et al. (2022) (recall: 65.9%). The use of GAN-augmented data resulted in a 19% improvement in CNN performance and a 10% improvement in LSTM performance over the original datasets. The hybrid model improved by 2% post-augmentation, confirming that even high-performing multi-modal models benefit from GAN-generated synthetic data. This quantitative evidence underscores the robustness, generalisability, and practical utility of the proposed framework in emotion recognition systems. The challenges posed by single-modal models such as limited feature representation and difficulty in recognizing underrepresented emotions were effectively addressed by combining complementary information from facial expressions and physiological signals. Furthermore, GAN augmentation overcame the issue of dataset imbalances, generating synthetic data that enriched the diversity of training samples and improved model generalization. The hybrid model (multi-modal) achieved state-of-the-art performance, with significant improvements in accuracy, precision, recall, and F1-score compared to existing methods. Its ability to robustly recognize subtle emotions, such as fear and disgust, is proof of the effectiveness of the integrated approach. The results confirm that leveraging GANs for data augmentation, in conjunction with multi-modal architectures, is a powerful strategy for advancing emotion recognition technologies. By addressing key challenges such as dataset imbalances and the inherent weaknesses of single-modal systems, the hybrid approach paves the way for more accurate, reliable, and versatile emotion recognition systems that are well-suited for real-world applications. The findings underscore the potential of this approach to redefine state-of-the-art solutions in emotion recognition, making it a valuable contribution to the field and a strong foundation for future research and applications.

Future research should explore real-time deployment in interactive systems. Expand datasets to include diverse demographic and cultural variations to improve robustness. This would enhance its ability to generalise across different populations and ensuring more accurate recognition of emotions in varied contexts. Additionally, incorporating domain adaptation and transfer learning could enhance generalisation across diverse environments.

REFERENCES

Abdulyekeen, R. (2025). Artificial Intelligence Driven and Comparative Analysis of Pulmonary Disease Prediction Employing Random Forest for Accurate Diagnosis. *FUDMA Journal of Sciences (FJS)*, *9*, 229–235. https://doi.org/https://doi.org/10.33003/fjs-2025-09(AHBSI)-3433

Alharbawee, L., & Pugeault, N. (2024). Generative Adversarial Networks for Facial Expression Recognition in the Wild. *International Journal of Computing and Digital Systems*, *15*(1), 1259–1282. https://doi.org/10.12785/ijcds/160193

Ali, K., & Hughes, C. E. (2023). *A Unified Transformer-based Network for multimodal Emotion Recognition*. *14*(8). http://arxiv.org/abs/2308.14160

Bao, J., Tao, X., & Zhou, Y. (2024). An Emotion Recognition Method Based on Eye Movement and Audiovisual Features in MOOC Learning Environment. *IEEE Transactions on Computational Social Systems*, *11*(1), 171–183. https://doi.org/10.1109/TCSS.2022.3221128

Cimtay, Y., Ekmekcioglu, E., & Caglar-Ozhan, S. (2020). Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access*, *8*, 168865–168878. https://doi.org/10.1109/ACCESS.2020.3023871

Eke, C. I., Norman, A. A., & Shuib, L. (2021). Context-Based Feature Technique for Sarcasm Identification in Benchmark Datasets Using Deep Learning and BERT Model. *IEEE Access*, *9*, 48501–48518. https://doi.org/10.1109/ACCESS.2021.3068323

Guangcheng, B., Yan, B., Tong, L., Shu, J., Wang, L., Yang, K., & Zeng, Y. (2021). Data Augmentation for EEG-Based Emotion Recognition Using Generative Adversarial Networks. *Frontiers in Computational Neuroscience*, *15*(December). https://doi.org/10.3389/fncom.2021.723843

Hongli, Z. (2020). Expression-eeg based collaborative multimodal emotion recognition using deep autoencoder. *IEEE Access*, *8*, 164130–164143. https://doi.org/10.1109/ACCESS.2020.3021994

Jang, E. H., Byun, S., Park, M. S., & Sohn, J. H. (2019). Reliability of Physiological Responses Induced by Basic Emotions: A Pilot Study. *Journal of Physiological Anthropology*, *38*(1), 1–12. https://doi.org/10.1186/s40101-019-0209-y

Jianhua, Z., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, *59*(March 2019), 103–126. https://doi.org/10.1016/j.inffus.2020.01.011

Khan, N., & Sarkar, N. (2022). *Semi-Supervised Generative Adversarial Network for Stress Detection Using Partially Labeled Physiological Data*. http://arxiv.org/abs/2206.14976

Kwaghtyo, K. D., & Eke, C. I. (2022). Smart farming prediction models for precision agriculture : a comprehensive survey. In *Artificial Intelligence Review* (Issue 0123456789). Springer Netherlands. https://doi.org/10.1007/s10462-022-10266-6

Ma, F., Li, Y., Ni, S., Huang, S., & Zhang, L. (2022). Data Augmentation for Audio–Visual Emotion Recognition with an Efficient Multimodal Conditional GAN. *Applied Sciences (Switzerland)*, *12*(1). https://doi.org/10.3390/app12010527

Muhammad, F., Hussain, M., & Aboalsamh, H. (2023). A Bimodal Emotion Recognition Approach through the Fusion of Electroencephalography and Facial Sequences. *Diagnostics*, *13*(5). https://doi.org/10.3390/diagnostics13050977

Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F., & Chandran, V. (2020). Automatic Emotion Recognition Using Temporal Multimodal Deep Learning. *IEEE Access*, *8*, 225463–225474. https://doi.org/10.1109/ACCESS.2020.3027026

Nemati, S., Rohani, R., Basiri, M. E., Abdar, M., Yen, N. Y., & Makarenkov, V. (2019). A Hybrid Latent Space Data Fusion Method for Multimodal Emotion Recognition. *IEEE Access*, *7*, 172948–172964. https://doi.org/10.1109/ACCESS.2019.2955637

Salama, E. S., El-Khoribi, R. A., Shoman, M. E., & Wahby Shalaby, M. A. (2021). A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition. *Egyptian Informatics Journal*, *22*(2), 167–176. https://doi.org/10.1016/j.eij.2020.07.005

Siddiqui, M. F. H., Dhakal, P., Yang, X., & Javaid, A. Y. (2022). A Survey on Databases for Multimodal Emotion Recognition and an Introduction to the VIRI (Visible and InfraRed Image) Database. *Multimodal Technologies and Interaction*, *6*(6). https://doi.org/10.3390/mti6060047

Soleimani, S. (2024). *Deep Learning Architectures for Enhanced Emotion Recognition from EEG and Facial Expressions Sareh Soleimani*.

Song, T., Zheng, W., Lu, C., Zong, Y., Zhang, X., & Cui, Z. (2019). MPED: A multi-modal physiological emotion database for discrete emotion recognition. *IEEE Access*, *7*, 12177–12191. https://doi.org/10.1109/ACCESS.2019.2891579

Sung-Nien, Y., Shao-Wei, W., & Chang, Y. P. (2022). Improving Distinguishability of Photoplethysmography in Emotion Recognition Using Deep Convolutional Generative Adversarial Networks. *IEEE Access*, *10*(November), 119630–119640. https://doi.org/10.1109/ACCESS.2022.3221774

Taisheng, Z., Song, L., Wang, J., Teng, W., Xu, X., & Ma, C. (2020). Data synthesis using dual discriminator conditional generative adversarial networks for imbalanced fault diagnosis of rolling bearings. *Measurement: Journal of the International Measurement Confederation*, *158*(January). https://doi.org/10.1016/j.measurement.2020.107741

Ullah, A., Wang, J., Anwar, M. S., Whangbo, T. K., & Zhu, Y. (2021). Empirical Investigation of Multimodal Sensors in Novel Deep Facial Expression Recognition In-the-Wild. *Journal of Sensors*, *2021*. https://doi.org/10.1155/2021/8893661

Wei, W. (2024). Targeted generative adversarial network (TWGAN-GP)-based emotion recognition of ECG signals. *E3S Web of Conferences*, *522*. https://doi.org/10.1051/e3sconf/202452201042

Win, S. S. K., Siritanawan, P., & Kotani, K. (2023). Compound facial expressions image generation for complex emotions. *Multimedia Tools and Applications*, *82*(8), 11549–11588. https://doi.org/10.1007/s11042-022-14289-7

Yan, X., Zhao, L. M., & Lu, B. L. (2021). Simplifying Multimodal Emotion Recognition with Single Eye Movement Modality. *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, 1057–1063. https://doi.org/10.1145/3474085.3475701

Younis, E. M. G., Mohsen, S., Houssein, E. H., & Ibrahim, O. A. S. (2024). Machine learning for human emotion recognition: a comprehensive review. In *Neural Computing and Applications* (Vol. 36, Issue 16). Springer London. https://doi.org/10.1007/s00521-024-09426-2

Zhang, Y., Cheng, C., & Zhang, Y. (2021). Multimodal Emotion Recognition Using a Hierarchical Fusion Convolutional Neural Network. *IEEE Access*, *9*, 7943–7951. https://doi.org/10.1109/ACCESS.2021.3049516

Zhang, Y., Tao, X., Ai, H., Chen, T., Zhang, Y., Tao, X., Ai, H., Chen, T., & Gan, Y. (2024). *Multimodal Emotion Recognition by Fusing Video Semantic in MOOC Learning Scenarios Multimodal Emotion Recognition by Fusing Video Semantic in MOOC Learning Scenarios*.

Zheng, W. L., Zhu, J. Y., & Lu, B. L. (2019). Identifying stable patterns over time for emotion recognition from eeg. *IEEE Transactions on Affective Computing*, *10*(3), 417–429. https://doi.org/10.1109/TAFFC.2017.2712143

Zhong, P., Wang, D., & Miao, C. (2022). EEG-Based Emotion Recognition Using Regularised Graph Neural Networks. *IEEE Transactions on Affective Computing*, *13*(3), 1290–1301. https://doi.org/10.1109/TAFFC.2020.2994159