

**FIRAT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
TÜRKİYE**



**GRAPH NEURAL NETWORK-BASED MULTIMODAL
EMOTION RECOGNITION**

Hussein Farooq Tayeb ALSAADAWI

Doctoral Thesis

DEPARTMENT OF SOFTWARE ENGINEERING

JULY 2024

FIRAT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
TÜRKİYE

Department of Software Engineering

Doctoral Thesis

GRAPH NEURAL NETWORK-BASED MULTIMODAL
EMOTION RECOGNITION

Thesis Author

Hussein Farooq Tayeb ALSAADAWI

Supervisor

Prof. Dr. Resul DAŞ

JULY 2024

ELAZIG

TURKEY REPUBLIC
FIRAT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

Department of Software Engineering
Doctoral Thesis

Title: Graph Neural Network-Based Multimodal Emotion Recognition
Author: Hussein Farooq Tayeb ALSAADAWI
Initial Sub. Date: 10.06.2024
Defense Date: 18.07.2024

THESIS APPROVAL

This thesis, prepared according to the Firat University Institute of Science's thesis writing rules, has been evaluated by the jury members whose signatures are below and has been unanimously accepted as a result of the defense being made open to academic listeners.

Supervisor:	Prof. Dr. Resul DAŞ Firat University, Faculty of Technology	<i>Signature</i> Approved
Chairman:	Prof. Dr. Mehmet KAYA Firat University, Faculty of Engineering	Approved
Member:	Asist Prof. Dr. Murat AYDOĞAN Firat University, Faculty of Technology	Approved
Member:	Assoc. Prof. Dr. Harun BİNGÖL Turgut Özal University, Engineering and Natural Sciences Faculty	Approved
Member:	Assoc. Prof. Dr. Serpil ASLAN Turgut Özal University, Engineering and Natural Sciences Faculty	Approved

This thesis was registered at the meeting of the Institute Board of Directors date/...../2024.

Prof. Dr. Burhan ERGEN
Institute Director

DECLARATION

As a PhD student at Fırat University, I, Hussein Farooq Tayeb, under the supervision of Professor Resul Das from the Software Engineering Department, confirm that my thesis titled "Graph Neural Network-Based Multimodal Emotion Recognition" is prepared under the thesis writing guidelines of the Fırat University Institute of Science. I affirm that all information included in my thesis is accurate, and the production and presentation of this information comply with scientific and ethical standards. I have correctly cited all sources and acknowledged all individuals and organizations providing financial and moral support. Furthermore, I declare that the data and information presented here have never been used to obtain any academic title before this submission.

18.07.2024

Hussein Farooq Tayeb ALSAADAWI

PREFACE

The successful completion of this dissertation is primarily attributed to the grace and mercy of ALLAH. Professor Dr. Resul DAS's exceptional guidance and steadfast support have also been invaluable. His expert insights into the selection of research topics and methodologies were indispensable to the success of this endeavor. His unwavering encouragement and meticulous feedback continuously inspired the refinement and advancement of this research.

I extend profound thanks to the members of the thesis defense jury. Their incisive and insightful comments have enhanced the quality of this dissertation and profoundly deepened my understanding of the specialized field. My heartfelt appreciation extends to the distinguished faculty members of the Department of Software Engineering at Firat University. Their exemplary instruction and guidance have been fundamental to my academic growth.

I also sincerely thank my esteemed official office and administration for their generous support during my study leave. I extend heartfelt and profound appreciation to my parents and brothers for their relentless support and motivation throughout my academic journey. Special acknowledgment is reserved for my wife and children, whose enduring support and patience have been the cornerstone of my resilience. I am deeply grateful to all my friends for their sage advice, which has greatly enriched my academic and personal life. Lastly, I would like to express my gratitude to the Turkish government for providing me with the opportunity to pursue higher education and develop scientific skills.

Hussein Farooq Tayeb ALSAADAWI

ELAZIĞ, 2024

CONTENTS

	Page
PREFACE	iv
CONTENTS	vii
ABSTRACT	viii
ÖZET	ix
LIST OF FIGURES	xi
LIST OF TABLES	xiii
SYMBOLS AND ABBREVIATIONS	xiv
1. INTRODUCTION	1
1.1. Problem Overview	2
1.2. Main objectives of Thesis	3
1.3. Main Contributions and Scope of Thesis	4
1.4. Outline of the Thesis	5
2. LITERATURE REVIEW	7
2.1. AI-Enhanced Affective Computing	14
2.1.1. Sentiment Analysis	16
2.1.2. Emotion Recognition	19
2.1.3. Textual Analysis	27
2.1.4. Audio Analysis	31
2.1.5. Visual Analysis	35
2.1.6. Graph Neural Networks an Overview	39
2.2. Comprehensive Analysis of Trimodal	43
3. MULTIMODAL EMOTION RECOGNITION	50
3.1. Overview of Affective Computing	50
3.1.1. Trimodal Emotion Recognition	52
3.2. Artificial Intelligence	55
3.2.1. Neural Networks	56
3.2.2. Graph Neural Networks	57
3.2.3. Evolution of Research in General Graph Neural Networks	60
3.3. Aproach of General Graph Neural Networks Frameworks	63
3.3.1. Emotive Source	64
3.3.2. Exploring Deep Learning Architectures	65
3.3.3. Dimensionality Reduction and Graph Embeddings:	66
3.4. Fusion Techniques	67
3.4.1. Feature Layer Fusion	68
3.4.2. Decision-level Fusion	69
3.4.3. Mixed Layer Fusion	70
3.5. CA-WGNN Algorithm	72
3.5.1. Implementation	72
3.5.2. Dataset	73
3.5.3. Text Emotive Source	73
3.5.4. Long Short-Term Memory Architecture	74
3.5.5. Audio Emotive Source	74
3.5.6. Methods for Analyzing Emotions in Audio Data	75

3.5.7. Deep neural Network Architectures	76
3.5.8. Video Emotive Source using Convolutional Neural Networks	76
3.5.9. Application of Convolutional Neural Networks	77
3.5.10. Convolutional Neural Network Architectures	77
3.5.11. Cumulative Attribute-Weighted Graph Neural Network	78
3.5.12. Model Evaluation Metrics	79
3.5.13. Improved Accuracy and User Experience	80
3.5.14. Enhanced Performance and Scalability	81
3.5.15. Robustness, Reliability, and Real-Time Processing	82
3.5.16. Adaptive Learning and User Interaction	82
3.5.17. Navigating Challenges	82
3.6. Affective Computing in Society	82
3.6.1. Health	83
3.6.2. Business	83
3.6.3. Security	83
3.6.4. Education	84
3.7. Challenges and Solutions	84
3.7.1. Obstacles and Alternate Approaches	85
3.7.2. Privacy	86
3.7.3. Open Directions and Challenges Remaining	86
3.7.4. The Current State and Prospects	88
3.7.5. Emerging Technologies and Applications	88
3.7.6. Ethical Considerations and Cultural Variances	89
3.7.7. Real-World Instances of Affective Computing	90
3.8. Classificatory and Predictive Dimensions	90
3.8.1. Classification	92
3.8.2. Prediction	92
3.8.3. Recommendations and Development	93
4. TER-CA-WGNN: TRIMODEL EMOTION RECOGNITION USING CUMULATIVE ATTRIBUTE-WEIGHTED GRAPH NEURAL NETWORK	97
4.1. Literature Review	98
4.2. Proposed Approach for Emotion Recognition	102
4.2.1. Dataset Description	105
4.2.2. Speaker Embedding	106
4.2.3. Classification Using CA-WGNN	110
4.2.4. Problem Statement	114
4.2.5. Assessment Metrics for Classifier Performance	114
4.3. Experimental Results and Evaluation	116
4.3.1. Accuracy Analysis	117
4.3.2. Precision Analysis	120
4.3.3. Recall Analysis	122
4.3.4. F1-Score Analysis	124
4.3.5. Advancements Brought by CA-WGNN	126
5. MULTIMODAL EMOTION RECOGNITION USING BI-LG-GCN FOR MELD DATASET	133

5.1. Contributions of Bi-LG-GCN	134
5.2. Methodology	139
5.2.1. Data Collection	140
5.2.2. Data Preprocessing	141
5.2.3. Feature Extraction Using Kernel Principle Component Analysis	142
5.2.4. Implementation of Bilateral gradient graphical Conventional Network	144
5.3. Experimental Results and Evaluation	146
5.3.1. Experimental Setup	146
5.3.2. Existing Methods	147
5.3.3. Comparison Phase	148
6. EMOTION RECOGNITION FROM TEXT AND AUDIO DATASET USING CUMULATIVE ATTRIBUTE GRAPH NEURAL NETWORKS	154
6.1. Study contributions and novelties	155
6.2. Literature review	155
6.3. Methodology and Implementation	156
6.3.1. Dataset	156
6.3.2. Emotion Classification Using BERT	157
6.3.3. Visualization and Dimensionality Reduction: Creating a Valence-Arousal Graph	158
6.3.4. Dimensionality Reduction in Text Data Using t-SNE	159
6.3.5. Data Fusion and Feature Integration	160
6.3.6. Bi-CAGNN Approach for Emotion Classification	161
6.4. Experimental Results and Evaluation	164
7. CONCLUSION	167
REFERENCES	169
CURRICULUM VITAE	

ABSTRACT

Graph Neural Network-based Multimodal Emotion Recognition

Hussein Farooq Tayeb ALSAADAWI

Doctoral Thesis

FIRAT UNIVERSITY

Graduate School of Natural and Applied Sciences

Department of Software Engineering

July 2024, Page: xv + 197

The growing interest in emotion recognition highlights its vital role in improving human-computer interactions. This thesis introduces a new multimodal emotion recognition algorithm that represents a significant advancement by uniquely integrating textual, facial, and vocal data. However, it's important to note that the algorithm may face challenges in accurately interpreting certain emotional cues, such as sarcasm or irony, often conveyed through tone and speech. Despite these potential limitations, this comprehensive analysis enhances the accuracy and realism of interactions between humans and computers, enabling systems to respond more effectively to user moods and intentions. Our approach leverages various Graph Neural Network architectures to develop models specialized in processing textual and visual information, utilizing level fusion. Our algorithm underwent rigorous testing on the CMU-MOSI and IEMOCAP datasets, achieving high accuracies of 94% and 93%, respectively. These results, supported by metrics such as precision, recall, and F1-score all-surpassing 92% demonstrate the superiority of our method over existing techniques. Additionally, we incorporated attention mechanisms and thoroughly reviewed datasets used in emotion recognition, emphasizing our method's robustness in practical scenarios. Moreover, we introduced the Trimodal Emotion Recognition Employing Cumulative Attribute-Weighted Graph Neural Network (TER-CA-WGNN), a practical application that inputs multiple data types and outputs emotional predictions, showcasing the direct applicability and impact of our research. This intuitive application enhances existing systems, facilitating research and development. It supports robust app creation, particularly aiding those with hearing impairments. Our algorithm's validation advances emotion recognition by demonstrating the importance of multimodal data integration for improved accuracy and robustness.

Keywords: Emotion recognition, Sentiment analysis, Deep learning, Graph neural network, Trimodal emotion, CA-WGNN algorithm.

ÖZET

Graph Neural Network-Based Multimodal Emotion Recognition

Hussein Farooq Tayeb ALSAADAWI

Doktora Tezi

FIRAT ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü

Yazılım Mühendisliği Anabilim Dalı

Doktora Programı

July 2024, Sayfa: xv + 197

Duygu tanıma alanındaki artan ilgi, insan-bilgisayar etkileşimlerini geliştirmedeki hayati rolünü vurgulamaktadır. Bu tez, metin, yüz ifadeleri ve ses verilerini benzersiz bir şekilde entegre eden yeni bir çoklu modlu duygu tanıma algoritması sunmaktadır. Ancak, algoritmanın ton ve konuşma yoluyla sıklıkla ifade edilen alaycılık veya ironi gibi belirli duygusal ipuçlarını doğru bir şekilde yorumlama konusunda zorluklarla karşılaşabileceğini belirtmek önemlidir. Bu potansiyel sınırlamalara rağmen, bu kapsamlı analiz, insanlar ile bilgisayarlar arasındaki etkileşimlerin doğruluğunu ve gerçekliğini artırır, sistemlerin kullanıcı ruh hallerine ve niyetlerine daha etkili bir şekilde yanıt vermesini sağlar. Yaklaşımımız, metin ve görsel bilgileri işlemek için uzmanlaşmış modeller geliştirmek üzere çeşitli Graf Sinir Ağı mimarilerinden yararlanmaktadır ve düzey birleştirme kullanılmaktadır. Algoritmamız, CMU-MOSI ve IEMOCAP veri kümelerinde sırasıyla %94 ve %93 yüksek doğruluk oranları elde ederek, kesinlik, hatırlama ve F1 skoru gibi metriklerle %92'nin üzerindeki değerlerle yöntemimizin mevcut tekniklerden üstünlüğünü göstermektedir. Ayrıca, dikkat mekanizmaları ekledik ve duygu tanıma kullanılan veri kümelerini kapsamlı bir şekilde gözden geçirdik, yöntemimizin pratik senaryolarda sağlamlığını vurguladık. Dahası, çoklu veri türlerini girdi olarak alan ve duygusal tahminler üreten Trimodal Duygu Tanıma Kumulatif Öznitelik Ağırlıklı Graf Sinir Ağı (TER-CA-WGNN) adında pratik bir uygulama tanıttık, araştırmamızın doğrudan uygulanabilirliğini ve etkisini sergileyen. Bu sezgisel uygulama, mevcut sistemleri geliştirir, araştırma ve geliştirmeyi kolaylaştırır. Özellikle işitme engellileri destekleyen sağlam uygulama oluşturmayı destekler. Algoritmamızın doğrulaması, gelişmiş doğruluk ve sağlamlık için çoklu modlu veri entegrasyonunun önemini göstererek duygu tanıma alanında ilerleme sağlar.

Anahtar Kelimeler: Duygu tanıma, Duygu analizi, Derin öğrenme, Grafik sinir ağı, Trimodal duygu, CA-WGNN algoritması.

LIST OF FIGURES

	Page
Figure 2.1. Plutchik’s Wheel of Emotions with PAD model and Example Emotions.	8
Figure 2.2. General block diagram for Trimodal Affective Computing.	14
Figure 2.3. Architecture of BiLSTM [124]	29
Figure 2.4. (FEA) learning [179].	36
Figure 3.1. Trimodal Fusion in Affective Computing.	51
Figure 3.2. Neural Network	57
Figure 3.3. A simplistic representation of a GNN, showcasing nodes.	59
Figure 3.4. Comparison of directed and undirected edges.	59
Figure 3.5. The process of graph filtering.	61
Figure 3.6. The process of graph pooling.	62
Figure 3.7. A general GNN architecture for node-focused tasks [263].	63
Figure 3.8. Multimodal Emotion Recognition using Trimodal	64
Figure 3.9. Graph Embedding.	67
Figure 3.10. Feature Layer Fusion	69
Figure 3.11. Decision-level fusion framework with three classifiers.	70
Figure 3.12. Mixed layer fusion framework integrating different modalities at multiple layers.	71
Figure 3.13. Deep Neural Network architecture.	75
Figure 3.14. The network architecture of BIMHA for affective computing [66].	91
Figure 3.15. The semi-supervised multimodal emotion recognition framework [83].	91
Figure 4.1. Graphic flow.	103
Figure 4.2. Multimodal data processing framework.	103
Figure 4.3. Proposed approach flow chart.	104
Figure 4.4. Structure of LSTM method.	107
Figure 4.5. Architecture of CNN method.	109
Figure 4.6. Structure of CA-WGNN.	111
Figure 4.7. The accuracy of the IEMOCAP and CMU-MOSI datasets.	119
Figure 4.8. Graphical representation of Accuracy.	120
Figure 4.9. The comparison of precision between the IEMOCAP and CMU-MOSI datasets.	121
Figure 4.10. Graphical representation of Precision.	122
Figure 4.11. The recall comparison between the IEMOCAP and CMU-MOSI datasets.	123
Figure 4.12. Graphical representation of Recall.	124
Figure 4.13. The F1-scores comparison between the IEMOCAP and CMU-MOSI datasets.	125
Figure 4.14. Graphical representation of F1-score.	126
Figure 4.15. ROC Curve Analysis for Text Data from the CMU-MOSI Dataset.	130
Figure 5.1. Multimodal Emotion Recognition using Bi-LG-GCN system	140
Figure 5.2. Design for Combining Multiple Text and Audio Datasets	141
Figure 5.3. GCN Architecture	145
Figure 5.4. Accuracy comparison of Bi-LG-GCN with existing methods	149
Figure 5.5. F1-score comparison of Bi-LG-GCN with existing methods	150
Figure 5.6. Precision comparison of Bi-LG-GCN with existing methods	151
Figure 5.7. Recall comparison of Bi-LG-GCN with existing methods	152
Figure 6.1. HoH: Analysis of Bi-CAGNN Algorithm.	156
Figure 6.2. Text-Valence-Arousal Emotion Graph.	158
Figure 6.3. t-SNE Visualization of Text Data.	159

Figure 6.4. Block diagram illustrating the feature-level fusion process. 161

Figure 6.5. The CAGNN architecture [24]. 161

Figure 6.6. X,Y Plot of Test Data. 163



LIST OF TABLES

	Page
Table 2.1. Overview of Graph Neural Network Methods	12
Table 2.2. Trimodal Research Review for Sentiment Analysis.	19
Table 2.3. Trimodal Research Review for Emotion Recognition	21
Table 2.4. Textual Emotional Recognition Methods	28
Table 2.5. Text Emotional Recognition Methods Using LSTM	31
Table 2.6. Audio Emotion Recognition Methods	33
Table 2.7. Audio Emotional Recognition Methods Using DNN	34
Table 2.8. Face Emotional Recognition Methods	36
Table 2.9. Face Emotional Recognition Methods Using CNN	37
Table 2.10. Review of GNN Methods and Applications (2018 - 2020)	40
Table 2.11. Review of GNN Methods and Applications (2021 - 2024)	41
Table 2.12. Review of GNN Methods and Applications (2019 - 2022)	43
Table 2.13. Overview of Datasets in Affective Computing	45
Table 3.1. Comprehensive Review of Emotion Recognition Using DNNs	54
Table 3.2. CNN Models Used for Emotion Recognition and Sentiment Analysis	55
Table 3.3. Core Concepts in CA-WGNN	72
Table 3.4. Binary classification confusion matrix. Image adapted from.	80
Table 3.5. Comparison of Various Models for Emotion Recognition	81
Table 3.6. Overview of Unresolved Challenges in TER and Potential Solutions	85
Table 3.7. The most frequently utilized trimodal datasets among researchers.	89
Table 3.8. Top Five Accuracy in Trimodal Sentiment Analysis Research	94
Table 3.9. Top five accuracy in trimodal emotion recognition research	95
Table 4.1. Comparison of Experimental Configurations	114
Table 4.2. Hardware and software configuration for system requirements	117
Table 4.3. Performance Metrics for IEMOCAP Dataset	118
Table 4.4. Performance Metrics for CMU-MOSI Datasets	118
Table 4.5. Comparison of accuracy between the IEMOCAP and CMU-MOSI datasets.	119
Table 4.6. Numerical outcomes of Accuracy.	120
Table 4.7. The comparison of precision between the IEMOCAP and CMU-MOSI datasets.	121
Table 4.8. Numerical Outcomes of Precision.	122
Table 4.9. Comparison of Recall.	123
Table 4.10. Numerical Outcomes of Recall.	124
Table 4.11. Comparison of F1-score.	125
Table 4.12. Numerical Outcomes of F1-score.	126
Table 4.13. Performance metrics of the CA-WGNN model on the IEMOCAP dataset.	127
Table 4.14. Performance Metrics	128

Table 4.15. Comparison of Performance Metrics 131

Table 4.16. Challenges of CA-WGNN 131

Table 5.1. Hardware and Software Configuration 147

Table 5.2. Representation of TP, FP, TN, and FN Explanation Matrix 148

Table 5.3. Comparison of accuracy 149

Table 5.4. Comparison of F1-score 150

Table 5.5. Comparison of precision 151

Table 5.6. Comparison of recall 152

Table 6.1. Model Performance Metrics 164

Table 6.2. Detailed Classification Report 165

Table 6.3. Consolidated Emotion Recognition Performance on CMU-MOSI Dataset 165



SYMBOLS AND ABBREVIATIONS

Symbols

F	:	The number of layers in the GNN
C	:	Set of nodes in graph S (assumed static)
A	:	Set of edges in graph S (assumed static)
z_{c0}	:	Initial feature vector of node c , where y_c is the initial node feature
s_{a0}	:	Initial feature vector of edge a , where h_c is the initial edge feature
$M(a)$	:	Set of nodes connected to edge a
$M(c)$	:	Set of nodes connected to node c
$\rho_A^{(f)}$	:	Aggregation function for edges at layer f
$\phi_A^{(f)}$	:	Combination function for edges at layer f
$\rho_C^{(f)}$	:	Aggregation function for nodes at layer f
$\phi_C^{(f)}$	:	Combination function for nodes at layer f
ρ_S	:	Aggregation function for the readout phase
ϕ_S	:	Combination function for the readout phase
$\hat{x}$	:	Final output of the GNN after the readout phase

Abbreviations

AI	:	Artificial Intelligence
AUC	:	Area Under the Curve
CNNs	:	Convolutional Neural Networks
DL	:	Deep Learning
FLP	:	Facial Landmark Points
ANNs	:	Artificial Neural Networks
AER	:	Audio Emotion Recognition
ML	:	Machine Learning
DNNs	:	Deep Neural Networks
ER	:	Emotion Recognition
SVM	:	Support Vector Machine
GBM	:	Gradient Boosting Model
VC(LR-SGD)	:	Voting Classifier, Logistic Regression and Stochastic Gradient Descent
GNNs	:	Graph Neural Networks
HCI	:	Human-Computer Interaction
LSTM	:	Long Short-Term Memory
MFCC	:	Mel-Frequency Cepstral Coefficients
ROC	:	Receiver Operating Characteristic
SA	:	Sentiment Analysis
AC	:	Affective Computing
SER	:	Speech Emotion Recognition
TER	:	Trimodal Emotion Recognition
DCCA	:	Deep Canonical Correlation Analysis

PAD	:	Pleasure-Arousal-Dominance
FEA	:	Face Emotional Analysis
MSA	:	Multimodal Sentiment Analysis
TAC	:	Trimodal Affective Computing
NLP	:	Natural Language Processing
KNN	:	K-nearest Neighbors
GCNs	:	Graph Convolutional Networks
RNNs	:	Recurrent Neural Networks
MER	:	Multimodal Emotion Recognition
Emo-DB	:	Berlin Database of Emotional Speech
LDC	:	Linguistic Data Consortium
CA-WGNN	:	Cumulative Attribute-Weighted Graph Neural Network
CMU-MOSI	:	Carnegie Mellon University Multimodal Opinion-level Sentiment Intensity Dataset
PSRMER	:	Proactive Service Recommendation Driven by Multimodal Emotion Recognition
CREMA-D	:	Crowd-sourced Emotional Multimodal Actors Dataset
IEMOCAP	:	Interactive Emotional Dyadic Motion Capture

1. INTRODUCTION

Over the last twenty years, the : Affective Computing (AC) field, which includes : Emotion Recognition (ER) and sentiment analysis, has garnered significant attention from the research community. This domain focuses on enabling computers and devices to recognize, interpret, and respond to human emotions through software and technology. In the modern digital era, integrating human emotions into computer systems has transitioned from a novel idea to an essential requirement [1]. At the heart of this shift is the expanding field of ER analysis [2], which enhances user experiences. ER, achieved through interpreting facial expressions, voice tones, and physiological signals, is increasingly integral to our interactions with technology [3].

This field, intersecting domains such as computer science [4], psychology, and signal processing, extends beyond traditional : Human-Computer Interaction (HCI). It significantly impacts various sectors by utilizing emotional data to create more engaging content that captures human feelings through artificial intelligence [5]. Multimodal approaches, employing techniques in textual, audio, and visual : Sentiment Analysis (SA), are instrumental in mirroring the complex nature of human emotions [6, 7]. These approaches combine different forms of communication to enhance our understanding of emotional expression. Textual data, encompassing emails, social media posts, and spoken language transcripts, is a primary communication medium. Audio data, tone, volume, pitch variations, and visual data capturing facial expressions are crucial in understanding an individual's identity and emotional state. Integrating these diverse data types deepens our comprehension of human emotions and behaviors, offering a nuanced view of societal moods through social media analysis.

Numerous studies [8, 9] have investigated the application of technology in identifying human emotions through various modalities, including facial expressions, speech, language, and gestures. Research reveals that nearly 90% of affective ER literature focuses on the modalities of facial expressions, speech, and text [10]. Emotions are generally defined as "short-term, complex, multidimensional behavior responses" [11], although this definition lacks consensus due to the absence of a universally accepted definition of emotion [12]. To understand human emotions, Ekman established the six basic emotional states of happiness, anger, sadness, fear, disgust, and surprise as widely recognized [13]. In non-verbal communication, emotion detection involves analyzing the behavior correlates of emotions, such as facial actions for expressions or vocal cues for speech [14]. For verbal communication, emotion detection typically entails SA of spoken text, categorizing words as positive or negative [15]. However, fully detecting emotions from text alone is notably challenging leading to a preference for SA in verbal communication contexts [15]. ER technologies can operate in unimodal [16, 17, 18], bimodal [19, 20, 21], or multimodal [22, 23, 24] settings.

Understanding and conveying emotions involves recognizing sentiment contradictions or

emotional categories, as discussed by Zhang et al. (2021) [25]. Emotion expression is multifaceted, extending beyond facial expressions to include gestures, voice tones, and other behaviors. Multimodal emotion detection can enhance fields like AC and human-robot interaction by making systems more empathetic and intuitive. This thesis explores using deep learning to interpret complex emotional cues, addressing emotions' variability and context-sensitive nature. The ultimate objective of this research is to develop a comprehensive understanding of human emotions through multimodal approaches. This understanding could lead to emotionally intelligent systems poised to transform HCI.

1.1. Problem Overview

A significant challenge in ER is the dependency on technologies focused on a single modality or traditional classification methods. Such methods often need to represent the wide array of human emotions adequately. This study introduces advanced methodologies crucial for refining emotion detection, a field that has recently seen a marked increase in interest. ER and Sentiment analysis enhance human-technology interactions by effectively identifying emotions by amalgamating textual, visual, and auditory data. AC aims to improve these interactions by recognizing, understanding, and appropriately responding to human emotions. This approach is vital for effective emotion management and fostering behavior changes, contributing to more precise emotional state classifications. By fusing text, audio, and visual inputs, ER systems can address the shortcomings of unimodal systems, which often fail to capture the complexity of emotional expressions. The following are the significant problems that led to cart this thesis:

- Constraints of contextual understanding: Unimodal systems frequently overlook essential contextual cues critical for precise ER. For instance, text-based systems might not discern sarcasm or subtle emotional nuances, whereas systems that solely analyze audio may overlook crucial visual emotional indicators.
- Challenges in comprehensive ER: Dependence on a single sensory modality typically leads to partial detection of emotions. For example, only interpreting facial expressions can reflect the complete emotional context with the support of vocal tones or textual cues.
- Demand for in-depth emotional analysis: Emotions are inherently complex and layered, usually necessitating the amalgamation of various cues for accurate interpretation. Systems based on a single modality are insufficient to understand an individual's emotional state comprehensively.

- Complexity in integrating multimodal data: Conventional methods for ER often struggle to efficiently merge and maintain the complexity inherent in diverse modalities such as textual, auditory, and visual inputs. Each modality possesses distinct characteristics and noise profiles, and their simplistic amalgamation or preliminary integration frequently results in the dilution of vital information and the subtle interdependencies among them. Traditional methods like Convolutional Neural Networks, predominantly designed for Euclidean data, face challenges in managing the non-Euclidean structure of multimodal data.
- Scalability, generalization, and interpretability issues: Current models for ER face significant challenges in scaling to accommodate large, real-world datasets. Moreover, these models often need more generalizability across varied datasets and emotional contexts. The opacity of many of these models also poses difficulties in interpretation, which is crucial for applications that require high trust levels, such as monitoring mental health and human-computer interaction.

Furthermore, in human-computer interaction, the advancement of multimodal ER can lead to more intuitive and empathetic digital assistants, enhancing user satisfaction and productivity. As our reliance on digital communication continues to grow, addressing the shortcomings of current ER systems is crucial for improving interpersonal communication, reducing misunderstandings, and fostering more meaningful connections in both personal and professional contexts. The development of sophisticated, multimodal ER systems thus represents a technological challenge and a significant opportunity to enhance human well-being and societal progress across various domains.

1.2. Main objectives of Thesis

This research introduces a groundbreaking approach to ER through the : Cumulative Attribute-Weighted Graph Neural Network (CA-WGNN) model. Leveraging the trimodal datasets, our study assesses the model’s efficacy across diverse linguistic and emotional contexts. The significance of this study is multifaceted:

- Enhanced emotion detection:
Our study explores the CA-WGNN model’s capacity to identify nuanced and complex emotional states previously unaddressed in research. By broadening the spectrum of detectable emotions, we deepen the understanding of human affective experiences.
- Empirical validation and academic impact:
Utilizing trimodal datasets such as : Carnegie Mellon University Multimodal Opinion-level Sentiment Intensity Dataset (CMU-MOSI) and : Interactive Emotional Dyadic Motion Capture (IEMOCAP), our work not only provides rigorous empirical support for the CA-WGNN model, highlighting its robustness and reliability in handling complex, diverse emotional data but also addresses a

crucial void in ER studies. It introduces a novel analytical framework that fosters academic discourse and sets a precedent for subsequent multimodal ER research.

- Critical review and survey study:

Our research encompasses a comprehensive review and survey of the current datasets, models, and literature on trimodal ER and sentiment analysis. This approach establishes the innovative character of the CA-WGNN model and provides a consolidated view of recent advances and ongoing challenges. It sets a benchmark for future explorations, enhances the field’s accessibility and comprehension, and contributes beneficial private articles that help guide and inform readers.

- Practical applications and contributions to AI research:

The CA-WGNN model significantly enriches : Artificial Intelligence (AI) research by illustrating the effectiveness of : Graph Neural Networks (GNNs) in complex, multimodal ER scenarios. Its real-world applications are extensive, enhancing user experiences in emotion-sensitive systems across healthcare, therapeutic interventions, and accessibility tools. Specifically, it offers potential benefits in emotion monitoring and tailored communication solutions, encouraging further investigations into advanced GNNs architectures.

1.3. Main Contributions and Scope of Thesis

This study significantly advances the field of AC by exploring the integration of trimodal psychological data with AI. We introduce and rigorously test a novel algorithm employing a graph neural network. The experiments were conducted using two multimodal datasets, bolstering the credibility of our results. The key contributions of this study are as follows:

- Introduction of Trimodal Emotion Identification:

The study presents Trimodal Emotion Identification using the CA-WGNN model, which improves ER accuracy by integrating auditory, visual, and textual data. The study utilizes the CMU Multimodal Opinion Sentiment and Emotion Intensity CMU-MOSI and Interactive Emotional Dyadic Motion Capture IEMOCAP databases, specifically designed for multimodal analysis as trimodal inputs.

- Advanced Feature Extraction Techniques:

We employ state-of-the-art : Deep Learning (DL) techniques for feature extraction and embedding, including : Long Short-Term Memory (LSTM) networks for text, : Deep Neural Networks (DNNs) for audio, and : Convolutional Neural Networks (CNNs) for video. These methods ensure robust and comprehensive feature representation across modalities.

- **Innovative Model Architecture:**

The innovative CA-WGNN model combines data from multiple modalities using a weighted graph structure and fusion techniques for classification, yielding robust results. This architecture leverages the complementary nature of different data types to enhance overall ER performance.

- **Comprehensive Evaluation and Validation:**

Our extensive experiments on the CMU-MOSI and IEMOCAP datasets demonstrate the model’s effectiveness, achieving significant improvements in ER accuracy. We provide a thorough evaluation and validation, ensuring the reliability and generalizability of our results.

- **Potential Applications and Contributions:**

This research emphasizes the versatility of trimodal data integration for SA and ER in text, audio, and video domains. We further explore the algorithm’s applicability in bimodal settings within the medical field, underscoring its potential for broader impact. Additionally, our work provides theoretical insights and practical contributions to AC, advancing the understanding of multimodal ER and offering a practical framework that can be adapted to various applications, including HCI, healthcare, and psychological research.

1.4. Outline of the Thesis

The thesis structure, meticulously designed, presents a comprehensive overview of significant research and valuable findings. It commences with an introductory section that outlines the study’s framework, including the background, research problem, and main objective, along with its contributions and primary limitations.

- **Chapter 2.** reviews the historical development and current trends in AI-enhanced and AC, with a focus on ER across various data types, including trimodal data, and a special emphasis on detection through computer vision.
- **Chapter 3.** delves into various approaches to ER using multiple modalities, offering an extensive review of existing techniques and a detailed overview of GNNs. This chapter also discusses the use of multi-datasets, such as CMU-MOSI, employed by the CA-WGNN algorithm, including feature extraction, data fusion methodologies, and an analysis of various model evaluation metrics.
- **Chapter 4.** introduces the TER-CA-WGNN model, analyzing its performance metrics such as accuracy, precision, recall, and F1-score, and discussing the advancements and improvements achieved by the proposed model. This chapter also

clarifies the novelty and contributions of the TER-CA-WGNN model, highlighting its significance in the field of AC, and aids the reader in understanding the advancements made in the research.

- **Chapter 5.**, the narrative presents the Bi-LG-GCN and discusses its methodology, implementation, and evaluation. This includes details on data collection, preprocessing, and feature extraction processes, with the Result Evaluation section thoroughly detailing the experimental setup and comparisons with existing methods.
- **Chapter 6.** investigates the CA-GNN approach, explicitly addressing the challenges posed by using text and audio sources in AC and Human-Computer Interaction for individuals with hearing or speech impairments. This chapter meticulously covers the methodology, implementation, and evaluation processes, demonstrating the comprehensive nature of the research.
- **Chapter 7.** summarizes the thesis's contributions, applications, and future research directions, offering recommendations and discussing recent developments.

2. LITERATURE REVIEW

This chapter explores current research and advances in empathy assessment, focusing on the methods and analyses employed by various studies.

In the 1960s, researchers began exploring the connection between emotions and computers, marking the early days of artificial intelligence. They developed Emotional Analysis Databases, categorizing emotions into classes such as anger, disgust, fear, joy, neutrality, sadness, and confusion [26]. Interest in artificial intelligence (AI) to understand genuine human emotions has significantly increased. Affective computing (AC) leverages advanced software and technologies to enable computers to detect, comprehend, and respond to human emotions effectively [27]. While AC is closely linked to AI and HCI, it primarily focuses on ER and interpretation [28]. HCI integrates computer systems that might include practical computing aspects, while AC specifically pertains to emotional analysis. AC utilizes physiological responses to identify emotions; this field intersects with psychology and various information technology domains, emphasizing the multidisciplinary nature of ER.

Understanding emotional communication involves analyzing sentiment polarity or categorizing emotions, which helps in discerning emotional feelings [25]. Emotional expressions extend beyond facial cues to include gaze, posture, hand gestures, and vocal nuances, encompassing basic emotions like fear, anger, contempt, joy, sadness, and surprise. AC addresses two main areas in identifying and understanding these emotions and sentiments. Psychologists have developed two key models to describe emotions: the discrete (or categorical) emotion model and the dimensional emotion model. The former focuses on distinguishing specific emotional states, while the latter employs a multi-dimensional approach, considering factors such as valence and arousal. Paul Ekman further refined ER by categorizing emotions into specific types with distinct triggers, highlighting that different expressions share universal patterns [29].

Eight fundamental emotions, such as love, trust, fear, surprise, sadness, disgust, and anger, can be separated into bipolar emotions and arranged into a circumplex model, according to Plutchik's theory paradigm [30]. This paradigm is built on an emotion circle and a color wheel. Emotions color language in human communication can be divided into arousal and valence, which describe how strongly a feeling is, either positive or negative. In contrast, arousal tells how strong a sense is; valence refers to whether one feels pleasant or unpleasant, whereas arousal is often viewed as how aroused or calm one feels [31], [32]. The categorical approach, discrete emotion modeling, segments emotions into distinct, identifiable categories. A seminal model in this domain is Plutchik's emotional wheel. This model identifies eight primary emotions, presented as opposites, for instance, joy versus sadness [33]. Furthermore, Ekman's model proposes six core emotions, recognized universally yet expressed variably across different cultures. This variance highlights the intricate interplay between universal emotional constructs and cultural specificity. These

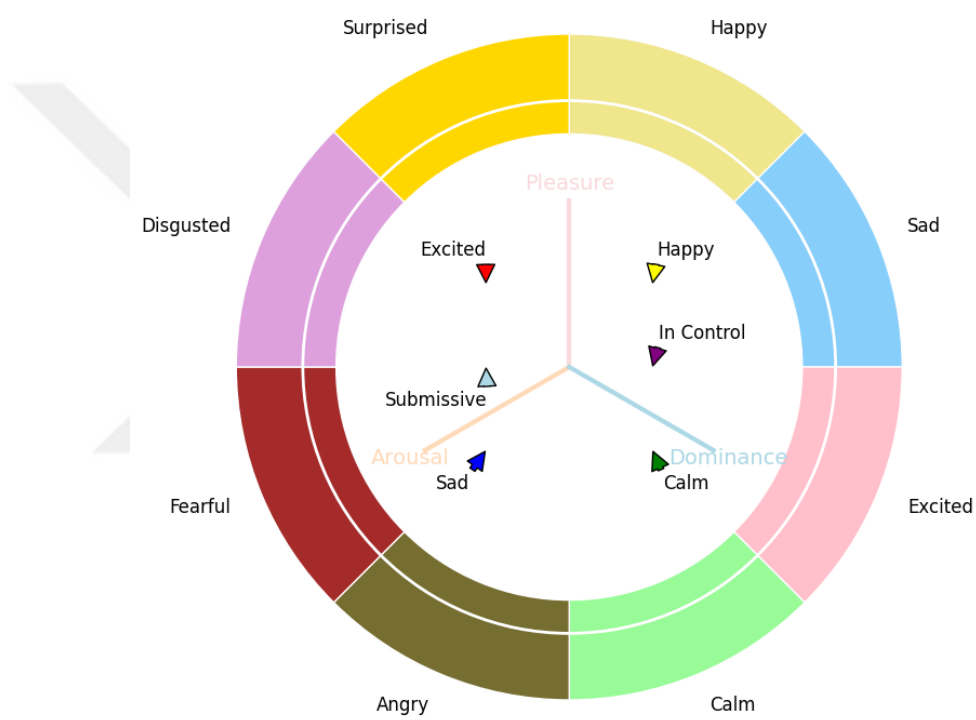


Figure 2.1. Plutchik's Wheel of Emotions with PAD model and Example Emotions.

models underpin significant theoretical frameworks in psychology and have profound applications in AI and HCI. For instance, Plutchik’s model, with its concentric arrangement of emotions based on intensity, is a valuable tool in SA, enabling nuanced interpretations beyond binary classifications. In AI, this model guides the development of algorithms capable of interpreting human emotions with greater sophistication.

Hourglass’s new two-dimensional SA model separates sentiment into two categories. Categorized is defined by horizontal valence and vertical arousal, each having four major quadrants covering different types of emotional responses, such as active and passive, which cover (pleasant and unpleasant). Based on dominance, several dimensional [34], [35] models, such as 2D and 3D models, characterize emotions. Typically, SA provides a binary output of 1 or 0, indicating a positive or negative attitude. Conversely, ER yields multiple labels representing various emotions. The field of AC encompasses both SA and ER. The domains illustrate the state of emotion and sentiment even if multimodal emotion they still has the output unimodal, bimodal, or trimodal. Text data, encompassing written or typed languages such as emails, social media posts, and spoken language transcripts, serves as one communication avenue. Voice data refers to a person’s voice’s pitch, loudness, and intonation. Visual data capture facial expressions, which offer valuable information for identifying individuals. Combining these diverse data types enhances our understanding of human emotions and behavior. Additionally, analyzing social media data enables us to gain better insights into the public mood.

Affective cognitive processing constitutes the intricate interaction between our emotional responses and cognitive operations, thereby facilitating the interpretation, response, and comprehension of our emotional states. This complex interplay fundamentally shapes human emotional experiences, exerting substantial influence over our perception of the world, interpersonal relationships, and self-identity. The influence of affective cognitive processes on our decision-making capabilities, learning processes, and memory retention can significantly affect our psychological health and overall well-being. The circumplex model of affect is a conceptual framework used to categorize emotions along two dimensions arousal and valence. This model is particularly useful for multimodal emotion analysis, which includes text, voice, and visual data sources.

- Arousal: This axis represents the state of alertness and physiological activity, ranging from low (e.g., calm, tired) to high (e.g., excited, tense).
- Valence: This axis represents the intrinsic attractiveness or averseness of an emotion, from negative (e.g., sad, angry) to positive (e.g., happy, content).

In the field of AC, particularly in the study of emotions, Multimodal Fusion Strategies are leading the way with groundbreaking methods for deciphering intricate emotional information. These strategies utilize multimodal ER and artificial intelligence to integrate various forms of data, including text, voice, and visual signals. This integration significantly improves the precision of ER. This part explores key advancements and

theoretical models that have influenced the modern approach to multimodal emotion analysis. By studying previous research on these fusion techniques, we gain insight into their sophisticated methods of enhancing our understanding and response to the complex spectrum of human emotions.

- Text Emotion: Analyzing the sentiment and emotional tone of the language used in text, where certain words and phrases can indicate the valence and arousal levels.
- Voice Emotion: Extracting emotional cues from speech characteristics, such as intonation, pitch, and volume, which can be indicative of different states on the arousal-valence spectrum.
- Visual Emotion: Interpreting emotions from facial expressions and body language, where visual cues like smiles or frowns align with points on the circumplex model to indicate the emotional state.

The circumplex model provides a structured approach to decoding emotions from diverse data sources in AC, enabling a comprehensive understanding of human emotions in a multimodal context. Gandhi et al. [6] present a detailed examination of various architectural developments in Sentiment Analysis SA fusion, comparing their strengths and limitations. The study categorizes new Multimodal Sentiment Analysis : Multimodal Sentiment Analysis (MSA) methods into ten distinct groups within a primary taxonomy, providing a comprehensive overview of recent advances in multi-modal fusion architectures. It also discusses the practical applications of these architectures and explores potential future directions in the field of AC.

Researchers have explored the relationship between emotions and computers, marking the nascent stages of artificial intelligence. AC, as Rosalind Picard described, focuses on the impact of technology on emotions [36]. Throughout its evolution, AC has introduced innovative tools and strategies for understanding and interpreting emotions, leading to its adoption in diverse sectors such as healthcare, education, and HCI. Today, AC employs advanced technologies and algorithms, with social media users enriching databases by expressing emotions through various media, including text, videos, and pictures.

Z. Sun et al. employed a deep canonical correlation network to analyze three suggested : Multimodal Emotion Recognition (MER) datasets [37], highlighting the challenge of automating emotional data extraction across different modes and domains. Similarly, N. Vryzas et al. developed SER within multi-modal semantic libraries, demonstrating that personalized approaches enhance speech ER accuracy [38]. Naveed Ahmed et al. emphasized the role of sophisticated emotion acquisition tools and classifiers, noting the necessity for enhanced privacy measures, unbiased models, and optimized techniques to improve the overall accuracy of ER systems [39].

M. Siddiqui et al. examined the inconsistencies in human-machine connections within MER systems, highlighting the challenges of real-time deployment, emotional annotation, and creating emotion-categorizing databases in restricted contexts [40]. A significant

limitation noted was the dependency on databases for collecting emotion-related data, which limits adaptability in complex environments. Moreover, recognizing various emotional states accurately and intensely is essential for effective multi-modal feature integration, a process critical for triggering emotional responses in the brain [38]. Research shows that MER systems, which utilize multiple modalities, outperform unimodal systems in accuracy and interpretation of emotions, underscoring their importance in advancing intelligent machine development [41].

Zhang et al. review the progress of DL in : Trimodal Emotion Recognition (TER), highlighting significant advancements in feature extraction and operational efficiency on devices with limited processing capabilities [42]. Their comprehensive analysis of multi-modal fusion strategies, including feature, decision, and model-level integration, addresses critical challenges such as dataset biases and the assimilation of new modalities. The paper also commends the potential of few-shot learning, though it notes the need for more practical solutions and detailed discussions on handling newer modalities and cross-corpus variations. Numerous researchers have analyzed various developments in architectural designs, comparing their advantages and limitations for ER fusion. Multimodal sentiment analysis methods have emerged as crucial advancements. This thesis provides an overview of the recent progress in multi-modal fusion architectures, categorizes the current architectures, and discusses their strengths, weaknesses, and practical applications. Additionally, it explores potential future directions in AC. To advance AC, it is essential to review the extensive body of existing research covering diverse emotion models and classification techniques.

Table 2.1. presents an extensive review of various GNNs methods focused on emotion recognition and multimodal analysis, reflecting significant academic interest in this field. The studies span from 2018 to 2022 and cover various topics, including multimodal emotion recognition, sentiment analysis, and dialogue modeling. Key contributions include integrating visual, audio, and text features for enhanced emotion recognition, achieving state-of-the-art performance, and addressing the complexities of multimodal data. Noteworthy improvements include surpassing previous methods by 3-5% on benchmark datasets and achieving an 84.50% F1 score, demonstrating advancements in this area. These studies also underscore the ongoing research efforts to refine GNN techniques, highlighting the dynamic nature of the field and the need for continued exploration to overcome practical application challenges and enhance real-time performance in emotion recognition systems.

The application of : Trimodal Affective Computing (TAC) can help identify and comprehend human emotions. Data analysis is performed using a comprehensive methodology that integrates text, visuals, and audio. The combination of sophisticated : Machine Learning (ML) methodologies, natural language processing, computer vision, and audio signal processing enables a robust system for recognizing emotional expressions. The integrated system delivers more comprehensive, precise, and contextually anchored HCI,

Table 2.1. Overview of Graph Neural Network Methods

Article Name	Year	Problems Addressed	Accuracy Results	Gap Identified
Review of GNN Methods and Applications	2018	Extensive overview of GNN methodologies, architectures, training techniques, and applications across different domains.	Comprehensive overview, no specific accuracy metrics.	Practical application challenges, scalability issues.
Multimodal Language Analysis	2018	Discusses the CMU-MOSEI dataset and challenges in analyzing language in natural settings.	No specific accuracy metrics.	Handling diverse and noisy data from real-world settings.
Multimodal Affect Analysis	2018	Integrates visual, audio, and text features for utterance-level affect analysis.	Improved emotion recognition performance.	Handling long-term dependencies, synchronization of modalities.
GCNs for Multi-Speaker Conversations	2019	Models context and speaker dependencies in multi-speaker conversations using GCNs.	No specific accuracy metrics.	Scalability and real-time application.
Dialogue-based GCNs	2019	Models context and speaker dependencies in conversations using GCNs.	Achieves state-of-the-art performance.	Handling long-term dependencies and varied dialogue structures.
GNNs for Image Understanding	2019	Uses GNNs to integrate multiple visual cues for improved image understanding.	No specific accuracy metrics.	Integration of complex visual features and scalability.
GNNs with Self-Attention	2020	Models context and speaker dependencies in conversations using GNNs and self-attention.	Improvement of 1%~2% over state-of-the-art strategies.	Handling large numbers of speakers.
Dialogical Emotion Correction Network	2020	Improves emotion recognition by correcting erroneous predictions in dialogues.	Significant improvement over existing methods, specific metrics not extracted.	Complex training process; need for large annotated datasets.
Graph-based Multimodal Emotion Recognition	2020	Addresses combining multiple modalities for emotion recognition using graph structures.	Outperforms previous methods by 3-5% on benchmark datasets.	Handling large and complex graphs efficiently.
Multimodal Fusion Network	2020	Enhances emotion recognition by integrating multiple modalities.	Outperforms state-of-the-art methods on IEMOCAP dataset.	Complexity in managing interactions between modalities.
Aspect Sentiment Analysis with Heterogeneous GNN	2021	Addresses sentiment analysis in reviews with heterogeneous data.	83.21% F1 Score on IEMOCAP.	Lack of dynamic and automatic selection of window size.
Contextualized GNN for Multimodal Emotion Recognition	2021	Integrates multiple modalities for better emotion recognition in dialogues.	Various F1 scores based on window size, highest being 84.50%.	No significant improvement over simple concatenation in some cases.
Hierarchical Graph Fusion for Multimodal Emotion Recognition	2021	Integrates hierarchical graphs for multiple modalities in emotion recognition.	Improved accuracy over baseline models.	Complexity in managing hierarchical graph structures.
Multimodal Dynamic Fusion Network	2022	Improves emotion recognition by fully understanding multimodal conversational context.	Significantly outperforms comparison methods on IEMOCAP and MELD datasets.	Complexity in modeling contextual interactions between modalities.

mental health surveillance, and user experience design.

Using data exchange improves accuracy, efficiency, and accuracy reduction. Different affective recognition systems, such as facial expression, speech, physiological signals, and body movements, are being studied to improve the accuracy and reliability of emotion identification. Besides unimodal, bimodal, and trimodal databases, VIRI is an image database that was created to represent five human emotions in the real world using visible and infrared images. MSA is a method developed by T. Kumar et al. [43]. In MSA, emotions and sentiment data can be classified using DL techniques and multivariate data fusion. The article delves into various current MSA approaches, outlining their strengths, weaknesses, and possible avenues for future research. MSA can be applied to a variety of sentiment analyses. The significance of emotion models rooted in psychological theories has been underscored in a comprehensive examination of AC literature, which included many research papers with MER. These models, often categorized as discrete or dimensional, play a pivotal role in ER systems, whether they rely on textual sentiment, speech, or physiological data. *Dis* (Discrete) Emotions are categorized into distinct, separate types, such as joy, sadness, anger, fear, surprise, and disgust. *Dim* (Dimensional) Emotions represented on a continuous scale across dimensions like valence (positive to negative) and arousal (low to high). Yan Wang et al. [44], further elucidated the technological advancements in this domain. While notable progress has been in utilizing both unimodal and multi-modal data, the field still faces challenges. Many current systems lack the robustness to predict emotions across varied scenarios accurately. Hence, future endeavors should expand benchmark databases, refine ER techniques, and enhance fusion strategies. The overarching goal is to better equip robots with emotional intelligence, ensuring they can empathetically interact within diverse environments.

A researcher has presented review papers on the study of multi-modal and unimodal frameworks. One notable study area includes a multi-modal framework for detecting abnormal behavior using various metrics such as ECG, EDA, EMG, BVP, accelerometer, respiration, and temperature [45]. However, readers may find their attention diverted by the numerous subtopics. Thus, focusing more deeply on one specific area, like the field that emphasizes TER, might be more beneficial, which aligns with our study.

Through discussions of different frameworks and categories of modern techniques, researchers gain insight into this emerging area of research. S. Poria et al. [27] provides an overview of the current state of multi-modal affect computing. It reviews benchmark datasets, advances three data sources that affect recognition, and examines the effectiveness of multi-modal fusion compared to unimodal analysis. The survey shows that multi-modal classifiers can outperform unimodal classifiers. Over the last five years, the framework of AC has seen a significant shift in its architecture, as evidenced by the computer science bibliography.

The technical aspects of this field are so deep, and the practical implications are so broad that it is impossible to display a vital model to cover all information. The study framed

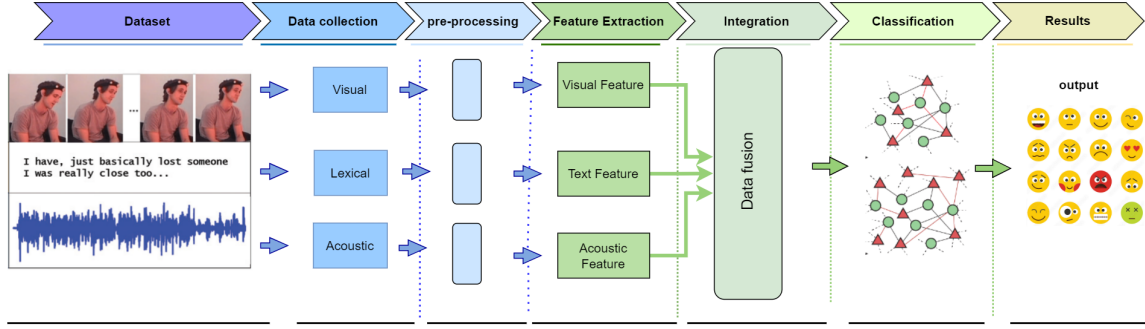


Figure 2.2. General block diagram for Trimodal Affective Computing.

the trimodal to be more understandable. This document introduces several performance standards, such as MSA and MER. Organizing and presenting previous articles on TAC in tables makes it easier to compare and contrast different studies and approaches, identify research trends, and build a foundation for future research.

2.1. AI-Enhanced Affective Computing

Affective cognitive processing constitutes the intricate interaction between our emotional responses and cognitive operations, thereby facilitating the interpretation, response, and comprehension of our emotional states. This complex interplay fundamentally shapes human emotional experiences, exerting substantial influence over our perception of the world, interpersonal relationships, and self-identity. The influence of affective cognitive processes on our decision-making capabilities, learning processes, and memory retention can have significant implications for our psychological health and overall well-being. The structure of the trimodal process is crucial for extracting significant features, as depicted in Figure 2.2.. This structure encompasses algorithms, fusion-type techniques, dataset details, and intelligent methods. Understanding the general block diagram, which outlines exploring SA and ER, is essential. These are key components of HCI using AI technology.

The initial phase involves delineating the datasets for multi-modal sentiment and ER experiments, including text, voice, and visual data pertinent to AC. Subsequent preprocessing of these datasets is vital, entailing tasks to clean, transform, and ready the data for analysis. Early ML stages necessitate feature extraction from these datasets, focusing on isolating critical features across modalities, including employing audio normalization techniques. The process advances through trimodal fusion, merging text, voice, and visual data using ML and : Natural Language Processing (NLP). This fusion encompasses feature layer fusion, decision-level fusion, and mixed-layer fusion strategies to refine emotion detection by enhancing feature relevance, minimizing noise, and applying precise fusion methods. Ultimately, this comprehensive approach leverages computer vision, speech analysis, and NLP to assign emotion labels, predict future emotions

accurately, and tailor the methodology to the data and task requirements in affective classification and prediction.

To emphasize the importance of multimodal approaches in emotion recognition, it is crucial to highlight the limitations of unimodal systems and the enhanced capabilities that bimodal and trimodal systems provide. By integrating multiple data sources, these systems can achieve a more comprehensive and accurate understanding of emotional states, thereby improving user engagement and interaction. Emotions play a critical role in human interaction. The surge in multimedia data has significantly advanced the field of emotion recognition through computing. Expressions of emotions, written or spoken, substantially impact behavior, decision-making processes, learning, and cognitive activities. Recent advancements in artificial intelligence have focused on creating systems adept at interpreting and replicating human emotions, primarily driven by progress in AC. Audio and text data integration has been enhanced, thus improving user engagement and understanding. This progression has rendered the management of isolated text-based modes more complex and intricate without such integrative techniques [46, 47].

Speech communication is challenged by the variability in accurately detecting emotions and a deficit in fundamental insights into emotional dimensions. These elements are essential for effectively conveying intentions, interpreting social cues, and fostering deeper relationships. Traditionally, emotion detection primarily relied on analyzing facial expressions [48]. Human emotional expressions are diverse, including body language, vocal tonality, and facial cues. Adopting a comprehensive approach that acknowledges these varied indicators is crucial for accurately recognizing and integrating emotions [49]. The field of multimodal emotion recognition has thus gained traction, focusing on amalgamating emotional signals from multiple sources for a more detailed and accurate depiction of human emotions. This field synergizes AC, voice analysis, computer vision, and machine learning to devise models for processing and interpreting diverse emotional inputs [50].

Improving human-computer interaction is a fundamental objective of multimodal emotion detection. By responding to users' emotional states, systems equipped with emotional intelligence can offer tailored and more engaging experiences [51]. Multimodal emotion recognition also finds applications in healthcare, contributing to evaluating emotional health and supporting personalized care approaches. Nevertheless, integrating and harmonizing data from different modalities, each with unique attributes and noise levels presents substantial challenges [52]. Deep learning algorithms are pivotal in effectively amalgamating data from diverse modalities. However, the variation in emotional expressions across individuals and cultural contexts poses difficulties in developing universally applicable models. Emotions' subjective nature and dependency on context, shaped by personal experiences, add complexity to this field.

: Graph Convolutional Networks (GCNs) have become instrumental in graphically representing complex systems. Nonetheless, the significance of textual data, especially in

sentiment analysis, remains critical. The efficiency of graphical representations, as seen in the growing popularity of infographics and knowledge graphs, contrasts the pivotal role of text in analyzing sentiments [53]. Research by Mucahit Pinar Savci et al. underscores the efficacy of pre-trained language models in multi-lingual sentiment analysis within e-commerce contexts [54]. In the domain of speech emotion extraction, the work of Pulatov et al., employing a combination of CNNs, : Mel-Frequency Cepstral Coefficients (MFCC), and Speech2Vec, shows notable advancements. However, the absence of 10-fold cross-validation raises queries about its generalizability [55]. Their findings suggest that integrating varied features and modalities can significantly augment speech emotion recognition research. While graphical representations aid in simplifying complex systems, textual data remains indispensable for a comprehensive understanding of human emotions and intentions. Future research could explore the synergies between graphical and textual analyses for broader insights.

Despite the challenges, the potential of multimodal emotion recognition spans various domains, including AC, human-robot interaction, virtual reality, gaming, market research, and psychology. Analyzing and interpreting intricate emotional signals enables the development of systems that are more empathetic, intuitive, and responsive [56]. This field aspires to offer an extensive understanding of human emotions by incorporating multiple modalities, potentially revolutionizing areas like human-computer interaction and healthcare. Our work introduces a bi-lateral gradient graph convolutional network (Bi-LG-GCN) methodology for recognizing human emotions [19]. The figure below illustrates the system for combining diverse text and audio datasets.

2.1.1. Sentiment Analysis

MSA involves utilizing several data sources, such as text, voice, and visual information, to determine the sentiment or emotion expressed by an individual. Combining these modalities allows a more comprehensive understanding of the sentiment expressed since each modality may convey different information. L. Qu [57] emphasized the effectiveness of multi-modal audio-visual strategies in eliciting emotions. Despite the significant advancements in this field, it's suggested that focusing research on specific areas could potentially enhance the efficacy of trimodal approaches. This thesis uses MSA to optimize the combination of these modalities. A trimodal analysis of multi-modal data employs MSA, and the various steps come together in the SA, as shown in Table 2.2.. The classification and prediction outputs are achieved through ML or DL, and the accuracy of the result is demonstrated by identifying which datasets and types of fusions were used. This survey provides an in-depth understanding of MSA through trimodal sources and sentiment categorization. By reviewing all papers on Trimodal that discussed various algorithms and datasets for sentiment classification, we could demonstrate the accuracy and intelligence of MSA. For future updates, researchers need to be aware of the usefulness and disadvantages of different methods.

To analyze MSA proposes a new method that explicitly incorporates trimodal behaviors. When compared with multi-modal approaches, the CMU-MOSI dataset experiments achieved state-of-the-art performance. In addition to text-based SA, Tensor Fusion Network also aims at generalizing this approach to opinionated videos, in which three forms of communication exist language (spoken words), visuals (expressions), and acoustics (speech). This dataset was built by A. Zadeh et al. [58] in 2018. It includes trimodal, and it is the largest multi-modal dataset at present. Twenty-three thousand videos are available with labels, covering many hot topics with different speakers. It uses hand transcription, and the audio and voice statuses of each video are aligned. Amazon Machinery Turkey platform judges scored the footage according to crowdsourcing.

When contrasted with other models, trimodal models exhibit notably superior performance than their unimodal or bimodal counterparts. By implementing an attention-based fusion technique, S. Poria et al. [59] established a system that uniformly integrates spoken phrases. To extract textual features from utterances, CNNs takes statements represented by Google and uses deep networks to identify voices and images. In a study by the same researcher P. Liang [60], each stage addresses a subset of multi-modal signals using Recurrent Multistage Fusion Networks. Through the use of trimodal datasets, extensive experiments have proven that the method can effectively model human multi-modal language and achieve modern results.

In the domain of embedded Sentiment Analysis, Quentin et al. [59] introduced a multi-modal Neural Network for use in NN. The three datasets involved are embedded systems, implying they possess restricted computational capabilities. Due to the large datasets, CMU-MOSI, ICT-MMMO, and IEMOCAP could not be utilized. The MOSI dataset was selected as it was anticipated to display distinct behavior from natural speakers compared to other datasets. In G. Huddar et al. [61], attention-based contextual fusion is used to classify AC through multi-modal fusion by simultaneously comparing and visualizing multiple stages. As a first step in multi-modal contextual fusion, contextual information is extracted from utterances, followed by voice features extracted from openSMILE and visual elements extracted from 3D-CNN. According to the results, attention-based contextual fusion is an effective method for classifying sentiments and emotions in the study conducted by J. Wu et al. [62] emphasize the importance of explaining the relationship between different modalities and the internal structures of unimodal features. To tackle this issue, a three-module BERT model is introduced. The first module employs a hierarchical multi-head self-attention mechanism, while the second module is responsible for filtering and substituting information using a gate channel. Finally, features are extracted by utilizing cross-modal signals. According to N. Majumder et al. [63], a hierarchical fusion strategy was used to evaluate two types of datasets in a trimodal scenario. Emotion reflects only some words in the text or specific regions in the picture or audio, based on the theory of attention. As a result, noise and redundancy are introduced. T. Wu et al. [86] demonstrate that the Bimodal Information-Augmented

Multi-Head Attention (BIMHA) architecture significantly enhances video sentiment analysis. Utilizing an attention-based multi-head fusion, this model effectively integrates data from textual, visual, and acoustic modalities, prioritizing critical features to boost prediction accuracy. As depicted in figure 3.14., BIMHA dynamically adapts to the relevance of different modalities, achieving superior accuracy on benchmark datasets compared to traditional methods.

The AMSAER framework excels in MSA, addressing data heterogeneity and modality synchronization. It uses attention mechanisms in a hierarchical model to interpret visual, auditory, and textual data, combining four analytical models. Its prowess, especially in identifying 'angry' and 'positive' sentiments, surpasses traditional methods, proven by the IEMOCAP dataset performance. Yet, AMSAER's reliance on specific datasets and controlled settings questions its adaptability. The study's limited comparison with recent research, like A. Aslam et al. [64], restricts broader understanding. Emphasizing accuracy neglects a holistic model analysis. Future versions should broaden analytical reach, refine metrics, and emphasize ethics for a comprehensive, ethical SA approach.

The article delves into MSA, spotlighting the innovative Fine-grained modal label-based Multi-Stage Network (FmlMSN). Diverging from conventional techniques, it leverages an advanced architecture, enhancing semantic alignment with attention mechanisms and Multi-Task Learning. Despite its prowess in complex tasks, the article lacks comprehensive experimental detail, impeding result reproducibility. Performance evaluation needs deeper comparative analysis, reminiscent of gaps in J. Peng et al. [65]. Moreover, with limited clarity on data noise handling and adaptability, concerns arise about the model's real-world utility. MSA is an effective method for organizing and presenting sentiment data. This approach enables researchers and data scientists to swiftly and efficiently identify trends, patterns, and relationships among various sentiments. Further observations indicate that pairwise MSA plays a vital role in predicting sentiments and significantly influences the design of system classification and prediction.

Table 2.2. Trimodal Research Review for Sentiment Analysis.

Authors	Year	Dataset	Fusion Type	Method	Language	Accu (%)
Aslam A. [64]	2023	IEMOCAP	Hybrid Fusion	AMSAER	English	93.14
Peng J. [65]	2023	CMU-MOSI	Multi-modal Fusion	AMSAER	English	83.59
Huddar M. [61]	2021	CMU-MOSI	Attention Fusion	Bi-LSTM	English	80.0
Wu T. [66]	2021	CMU-MOSEI, CH-IMS, MOSEI, IEMOCAP	Bimodal Fusion	BIMHA	Chinese, English	86.2
Wu J. [62]	2020	MOUD, CMU-MOSI	Attention Fusion	MMHA	English	82.71
Zadeh A. [67]	2018	CMU-MOSEI	Word-Level Fusion	MFN+DFG	English	85.2
Majumder N. [63]	2018	CMU-MOSI	Hierarchical Feature Fusion	CNN, RNN	Persian, English	80.0
Poria S. [68]	2017	CMU-MOSI	Bimodal Fusion	CATF-LSTM	English	81.30
Liang P. [60]	2018	CMU-MOSI	Word-Level Fusion	LSTHM	English	85.4
Cambria E. [34]	2017	MOUD, MOSI	Feature-level	CNN	Spanish, English	78.80
L. Poria [69]	2017	MOUD	Feature-level	CNN	English	80.35
Praveen K. [59]	2017	MELD	Non-hierarchical, Hierarchical	LSTM	English	57.50
Yu W. [70]	2020	CH-SIMS	Memory Fusion Network	DNN, LSTM	English, Chinese	64.74

Identifying weak or strong sentiment areas can significantly affect the MSA and help researchers more effectively target future work service efforts. Finally, table-based information MSA can be used to track sentiment changes over time, enabling the identification of sentiment trends and potentially significant problems.

2.1.2. Emotion Recognition

Advanced algorithms are pivotal in refining automation within platforms such as Facebook and YouTube, particularly for ER in extensive social data. Despite data management hurdles, progress in AC continues, anchored by robust foundational data. As discussed in subsequent sections, the evolution in ER technology is attributed to varied research approaches and frameworks that examine human sentiments using multi-modal data. This document presents these scholarly contributions chronologically, emphasizing the unique technical nuances and methodologies each paper proposes. Poria et al. [69] developed an innovative framework for Multimodal Sentiment Analysis MSA and Multimodal ER. Their approach leverages the temporal dynamics captured by : Recurrent Neural Networks (RNNs), spatial characteristics recognized through CNNs, and acoustic features extracted using openSMILE. This methodology is particularly effective in discerning emotions and thoughts expressed by users in response to video content. Notably, their late feature fusion strategy, which integrates these diverse data streams

later, has demonstrated superior performance over existing state-of-the-art models, especially when applied to the Interactive Emotional Dyadic Motion Capture IEMOCAP dataset. This capability to effectively analyze complex affective data underscores the robustness and adaptability of their model.

S. Dham et al. [71] presented a trimodal recognition task based on data from the Depression Evaluation Panel, utilizing a decision-level fusion technique. The methodology involved categorizing fusions using statistics based on various human interaction indicators, visualized using Fisher vectors and GMMs. J. Liang et al. [60] identified and matched utterances cross-modally to reconstruct trimodal emotion features. Their novel semi-supervised learning method leveraged unlabeled data and multi-modal fusion, enhancing classification accuracy across various datasets. M. Akhtar et al. [72] proposed a multi-task learning approach for recognizing and analyzing multi-modal emotions and sentiments. This method addressed challenges like data scarcity, mode alignment, noise, ambiguity, and computational complexity by incorporating data augmentation, modality fusion, and attention mechanisms.

Y. Wang et al. [73] introduced the Recurrent Attended Variation Embedding Network models, emphasizing the dynamic capture of subtle differences in nonverbal contexts, solidifying the foundations of multi-modal language SA and ER. D. Zhang et al. [74] proposed a dynamic dual influence network for real-time emotion detection during conversations. The approach simultaneously modeled both intramodal and intermodal influences, employing bidirectional information propagation for current and historical utterances, thereby significantly enhancing existing techniques. Mittal et al. [75] integrated a MER algorithm with a data-driven multiplicative early-fusion method, improving upon prior models using IEMOCAP and CMU-MOSI datasets.

Y. Mao et al. [76] described the "DialogueTRM," incorporating the Hierarchical Transformer and the Multi-Grained Interactive Fusion. These modules could capture conversations in both intramodal and intermodal contexts, outperforming other fusion approaches in ER scenarios. L. Stappen IEMOCAP et al. [77] introduced the MuSe and MuSe-CaR datasets, launching the MuSe 2021 challenge with four sub-challenges. They used a straightforward fusion method for MSA, incorporating continuous valence, DeepSpec-trim features, and MuSe-Sent classification.

B. Xie et al. [78] employed cross-modal transformers to merge trimodal input information, adapting different sources of representation vectors from separate prediction models for identifying trimodal emotion inputs. In 2024, Alsaadawi and Das demonstrated how GNNs significantly enhances ER, marking a major progress in AI by improving detection accuracy [19]. Additionally, the CA-WGNN model emerged as a key development, exhibiting exceptional skill in accurately identifying emotions from multimodal data [24]. M. Caschera et al. [79] proposed a trimodal feature sequence for classifying video data, testing Hidden Markov Models for each emotion to uncover complex communication and behavior characteristics.

Using a dynamic memory network, S. Xing et al. [80] introduced a novel approach for modeling self and interspeaker influences in dialogue videos. This model captured utterance-level dependencies using separate RNNs for each speaker, synthesizing these influences to create an adaptive memory network. V. Chudasama et al. [81] proposed a Multi-modal fusion network, the M2F-Net, for extracting emotion-relevant features from trimodal inputs.

Table 2.3. Trimodal Research Review for Emotion Recognition

Authors	Dataset	Year	Language	Fusion Type	Method	EL	Accuracy (%)
S. Dham [71]	DAIC-WZ	2017	Eng.	DL	GMM, SVM	8C	87.01%
S. Poria [82]	IEM	2017	Spanish	DL	CNN-SVM	4C	79.14%
L. Poria [69]	IEM, MOUD	2017	Eng.	FL	CNN	4C	80.35%
M. Caschera [79]	SE	2018	Eng.	HG	ML, HMMs	7C	76.12%
M. Akhtar [72]	MOSEI	2019	Eng.	MB	CIM-attention	6C	62.80%
Y. Wan [73]	MOSI	2019	Spanish	NM	RAVEN	3C	86.7%
D. Zhang [74]	AVEC, IEM	2020	Eng.	FL	BiDDIN	6C	63.6%
T. Mittal [75]	MOSEI, IEM	2020	Eng.	FL	M3ER	4C	82.70%
J. Liang [83]	IEM, MELD	2020	Eng.	MMF	MMD	4C	72.5%
S. Xing [80]	IEM, MELD	2020	Eng.	FL, DL	CSixN	6C	64.3%
Y. Mao [76]	IEM, MELD	2020	Eng.	MGI	Dialo-TRM	6C	69.23%
Y. Jiang [84]	FM	2020	Eng.	FL, DL	ML, DL	6C	49.17%
B. Xie [78]	MELD	2021	Eng.	FL	GPT, RNN	7C	65.00%
S. Tang [85]	IEM, MELD	2021	Eng.	HF	CNN, GFourN	4C	74.68%
L. Stappen [77]	MuSe-CaR	2021	Eng.	DL	LSTM, RNN	5C	35.12%
P. Kumar [86]	IIT-RM	2022	Eng.	HY	DNNs	4C	95.99%
V. Chu. [81]	IEM	2022	Eng.	MH	M2FNet	7C	69.69%
S. Mai [87]	MOSI, IEM	2023	Eng.	HY	HyCon	5C	83.8%
H. F. T. , R. Das [24]	CMU-MOSI, IEMOCAP	2024	Eng.	WGNN	CA-WGNN	3C	94%

Note: **AVEC:** Audio/Visual Emotion Challenge, **IEM:** IEMOCAP, **MOSEI:** CMU Multimodal Opinion Sentiment and Emotion Intensity, **MOSI:** CMU Multimodal Sentiment Analysis, **DAIC-WZ:** DAIC-WOZ Depression Study, **IIT-RM:** IIT Roorkee Multimodal, **MOUD:** Multimodal Opinion Understanding Dataset, **MELD:** Multimodal EmotionLines Dataset, **SE:** Spontaneous Emotions, **FM:** Five Multi-modal, **FL:** Feature Level, **MB:** Model-Based, **NM:** Not Mentioned, **HY:** Hybrid, **DL:** Decision Level, **MH:** Multi-Head, **MMF:** Multi-Modal Fusion, **MGI:** Multi-Grained Interactive, **HF:** Hierarchical Fusion, **HG:** Heterogeneous, **WGNN:** Weighted Graph Neural Network, **CA-WGNN:** Cumulative Attribute-Weighted Graph Neural Network, **EL:** Emotion Label, **3C:** Three Classes, **4C:** Four Classes, **5C:** Five Classes, **6C:** Six Classes, **7C:** Seven Classes, **8C:** Eight Classes, **Eng:** stands for English. Other abbreviations are standard within emotion recognition research.

The dynamic field of ER is rapidly evolving, with significant advancements in

Multi-modal data analysis, fusion techniques, and model development. These studies have broadened our understanding of human emotions, highlighting the vast potential of AI in this complex area. Utilizing multiple modalities, as shown in Table 2.3., not only improves the accuracy and robustness of ER but also deepens the comprehension of attitudes and feelings. A logically structured composition enhances the user’s engagement and understanding, making it easier for readers to grasp the research’s key findings and implications.

The 3D-CNN output characteristics of the face chunks were then classified using the : Support Vector Machine (SVM) classifier. Similarly, Wu and Liang [88] proposed a fusion-based approach for detecting emotions in affective speech. This method integrated acoustic-prosodic information and semantic labels using multiple classifiers.

The research outlined in [89] introduced a novel multi-modal approach for recognizing human emotions. This method employed a DL architecture, specifically the 3D-Convolutional Neural Network (3D-CNN), to extract spatiotemporal features from electroencephalogram (EEG) signals and facial recordings. The approach combined data enhancement and ensemble learning techniques to achieve final fusion projections. The multi-modality fusion in the proposed scheme was realized using both data and score fusion methods.

AC, as described by Rosalind Picard, impacts emotions [36]. Throughout the history of AC, researchers and developers have introduced new tools and strategies for understanding and interpreting emotions. AC has gained attention due to its foundational ideas, evolving frameworks, measurements, and extensive research. Its popularity has increased in the healthcare, education, and HCI sectors. This field increasingly employs sophisticated technologies and intelligent algorithms.

Additionally, social media users commonly express their emotions through text messages and frequently share videos, pictures, and other forms of textual communication. For each sentiment expressed, corresponding audio, text, and video files are collected, compiled, and stored in respective databases. S. Poria et al. [27] provide an overview of the current state of multi-modal affect computing, reviewing benchmark datasets and examining the effectiveness of multi-modal fusion compared to unimodal analysis. The survey demonstrates that multi-modal classifiers can outperform unimodal classifiers. Over the last five years, the AC framework has undergone significant architectural shifts, as evidenced in the computer science bibliography.

This field’s technical depth and practical breadth are so extensive that it is challenging to display a comprehensive model covering all information. The study emphasized the trimodal approach for a clearer understanding. This document introduces several performance standards, such as Multi-Modal Sentiment Analysis MER and MER. Organizing and presenting previous articles on TAC in tables facilitates comparing and contrasting different studies and approaches, identifying research trends, and establishing a foundation for future research. Speech ER, as developed by N. Vryzas et al. [38], aimed

to broadcast and communicate effectively within multi-modal semantic libraries. They tested the hypothesis that personalized speech ER would yield more accurate results. The findings suggest that the use of multi-modal databases enhances the accuracy of speech ER.

Gathering multi-modal features effectively is challenging, yet crucial, as it triggers emotional processing in the brain [38]. MER systems have proven more successful than unimodal systems in recognizing and interpreting emotions. Recognizing multiple emotional states enhances the accuracy of human-machine interactions, which is essential for developing intelligent machines.

In their study, Liu et al. [90] : Deep Canonical Correlation Analysis (DCCA) for recognizing emotions across multiple modalities. DCCA operates by independently transforming each modality and then correlating the various modalities into a unified hyperspace, following predefined canonical statistical constraints. The effectiveness of DCCA was evaluated across five different datasets, with results showing state-of-the-art detection accuracy rates. The study also revealed that each of the five emotions exhibited distinct functional conjunction characteristics. Additionally, the researchers in [72] proposed a deep multi-task learning system for simultaneous emotion analysis. A video’s multi-modal inputs provide unique and distinct information and typically do not contribute equally to decision-making. They suggested a context-level inter-modal attention method for concurrently evaluating an utterance’s sentiment and conveying emotions. The suggested method was tested for multi-modal sentiment and emotion evaluation using the CMU-MOSI database.

Nemati et al. [91] proposed an integrated multimodal data fusion approach where audio and visual models were fused using a latent space linear map, and their presented features in the cross-modal space were fused with the textual modality using a Dempster-Shafer (DS) theory-based evidential fusion approach. To enhance the efficiency of the emotional identification framework, a multimodal emotion detection model based on voice and text was presented [92]. The study confirmed that combining sentiment analysis of textual data with artificial intelligence yields effective outcomes for classification and prediction. The authors of [93] introduced the Visual Aspects Attention Network (VistaNet), integrating text and visual elements. They showed that visuals often support text in emotion identification, emphasizing significant elements of an entity rather than conveying emotions independently. As a result, VistaNet uses visual information as alignment to highlight relevant phrases in a text using attention, rather than employing visual data solely as features.

The field of AC is an emerging research area [41]. Lian et al. [94] proposed domain adversarial neural networks (DANN) for emotion detection, with the main objective being to forecast emotion categories and develop a common representation where the speaker’s identity could not be discriminated. Various neural network architectures, including CNNs, LSTM, and GRU, have demonstrated their efficacy in this regard [95]. To learn

acoustic emotion features, CNNs and LSTM were combined in the form of binary channels; meanwhile, an effective Bi-LSTM (bidirectional long short-term memory) network was used to capture textual elements. In addition, a Deep Neural Network DNNs was employed to classify and understand the fusion characteristics. The results of both voice and text emotion analysis were used to establish the final emotional state.

The paper examines emotion models, databases, and recent advances in AC. ML algorithms are usually trained using five types of baseline databases. Multi-modal emotion surveys still need to be included. A study proposed by N. Shoumy [96] was designed to fill this gap. Many studies have shown that businesses and researchers need emotional recognition of multi-modal, extensive data. Using physiological and other modalities, the impact of computing can construct more robust systems, resist spoofing, and be used in various environments. Through discussions of different frameworks and categories of modern techniques, researchers gain insight into this emerging area of research.

The depictions of different speakers were closer together thanks to their use of this technique. Using unlabeled data during the training phase, they reduced the effect of low-resource sample collections. The researchers of [97] introduced a DL based method for accessing and combining text and perceptual data to classify emotions. To extract acoustic characteristics from unprocessed audio, they added a DCNN layer after using a SincNet layer based on a customizable sinc value with band-pass filtering. Compared to performing convergence over the raw voice signal, the method develops filter banks modified to recognize emotions and provides superior aspects. To infer N-gram level association on hidden models derived from the Bi-RNN, they used two parallel streams for text analysis (a DCNN and a Bidirectional RNNs followed by a DCNN) with cross attentiveness.

The researchers of [75] demonstrated M3ER, a learning-based approach for emotion identification from several input modalities. Their system incorporated inputs from many co-occurring models and is more resistant to any of the sensors are noisy various approaches than previous methods. M3ER models an innovative, data-driven multiplicative fusion strategy for combining the models, which learns to enhance the more accurate cues. M3ER resisted sensor noise by incorporating a check step that used Canonical Correlational Assessment to distinguish between ineffective and effective models. Mittal et al. introduced EmotiCon, a learning-based system designed for identifying perceived emotions from photos and video clips by considering context [98]. This approach incorporates three different perspectives of context for emotion detection. The initial interpretation was based on using various senses for emotion identification. In the second analysis, they obtained semantic information from the input image and used a self-attention-based CNNs to encode it. Lastly, depth maps were employed to simulate the third interpretation, connected to socio-dynamic interactions and agent proximity. They showed the effectiveness of their network by running tests on datasets.

The authors of [99] suggested an innovative emotion-relevant crucial subnetwork

identification method and examined 3 EEG functional connectivity network features. The capacity of the EEG connectivity characteristics in emotion detection was identified using three open emotion EEG databases. By single-channel analysis, the strength feature surpassed the state-of-the-art differential entropy feature in classification accuracy.

The method for MER was then suggested. The HED dataset was far bigger than previous datasets and comprises emotion-aligned face, body, and text samples that express many emotions, including happiness, sadness, disgust, anger, and fear. The researchers of [100] utilized Muse-CaR, a vehicle review database, to make continuous emotional predictions. They initially extract handmade characteristics and complex depictions from several modes to do this. They next used the LSTM RNNs and the self-attention system to represent the sequence’s complicated temporal relationships. The Concordance Correlation Coefficient (CCC) loss was used to direct the model’s learning of local fluctuations and the global trend of emotion. Lastly, two fusion procedures, early fusion and late fusion, were used to further enhance the model’s performance by complementing information from various modalities.

The researchers [101] suggested an innovative cross-representation speech model for detecting emotions on wav2vec 2.0 voice characteristics and also trained a CNNs based model to distinguish emotions from text data extracted with Transformer-based models. An approach to identifying age, gender, and state of mind from audio was developed by Zaman et al. [102]. All audio recordings in our system were transformed into 20 statistical attributes, and the transformed numerical dataset was utilized to build several prediction models in order to achieve the goal. : K-nearest Neighbors (KNN), XGBoost, AdaBoost, and Decision Tree, : Artificial Neural Networks (ANNs), Naive Bayes, and SVM were some of the models used in this approach.

Data exchange enhances accuracy, efficiency, and reduces inaccuracies in affective recognition systems, which include facial expression, speech, physiological signals, and body movements. These systems are being developed to improve the accuracy and reliability of emotion identification. VIRI, an image database representing five human emotions using visible and infrared images, complements unimodal, bimodal, and trimodal databases. T. Kumar et al. developed the MSA method [43]. Naveed Ahmed et al. [39] emphasized the importance of emotion acquisition tools and feature classifiers for improving ER accuracy. Their study includes a detailed analysis of human emotions, reliable emotion acquisition tools, and an overview of MER datasets, highlighting gaps in privacy policies and model bias. M. Siddiqui et al. [40] discussed the inconsistencies in the connection between humans and machines in MER systems, focusing on aspects like real-time deployment and emotional annotation. Their work underlines the limitations of database-dependent data collection in challenging environments. The HED dataset, created by Fang et al. [103], is a sizable multimodal emotion dataset for aiding emotion detection challenges.

MSA, as outlined, categorizes emotions and classifies sentiment data using DL techniques

and multivariate data fusion. Yan Wang et al. [44] discussed the significance of emotion models rooted in psychological theories and the challenges faced in ER, suggesting the expansion of benchmark databases and refinement of recognition techniques. Ragheb et al. [104] developed a model that recognizes six emotions using a two-phase approach involving tokenization, encoders, and Bi-LSTM units trained via average stochastic gradient descent.

The article by Hussein Farooq Tayeb Al-Saadawi and Resul Das [24] was published in 2024. This study introduces a novel approach for emotion recognition that integrates trimodal datasets, textual, audio, and visual, using the Cumulative Attribute-Weighted Graph Neural Network (CA-WGNN). The model effectively combines these diverse modalities through a weighted graph structure, significantly enhancing sentiment analysis's accuracy and robustness. The experimental results demonstrate impressive performance metrics, achieving 94% accuracy on the CMU-MOSI dataset and 93% on the IEMOCAP dataset. These high accuracy rates indicate the model's superior ability to capture nuanced emotional states, making it a promising tool for sentiment analysis. This innovative methodology marks a notable advancement in affective computing, emphasizing the value of integrating multimodal data for improved sentiment recognition. The study of multi-modal and unimodal frameworks includes a multi-modal framework for detecting abnormal behavior using metrics such as ECG, EDA, EMG, BVP, accelerometer, respiration, and temperature [45].

Mucahit Soylu et al. [105] investigated attitude markers in academic writing, using data cleaning with Java and a radial knowledge graph for visualization, highlighting the potential for ER in diverse linguistic landscapes. As discussed in [106], verbal and non-verbal communication in nursing significantly influences patient healing. The research article "Exploring The Role of Verbal and Non-Verbal Communication Skills of Nursing in Patients' Healing" elucidates the integral role of emotional facets in nurse-patient communication. It underscores that verbal communication, characterized by the tone, pitch, and choice of words, extends beyond mere information exchange, serving as a conduit for emotional support and reassurance, uplifting patient morale and fostering a positive treatment outlook. Concurrently, the study accentuates the profound impact of non-verbal cues, such as body language, facial expressions, and tactile interactions, in forging an emotional connection with patients, offering solace and a sense of security.

The research further delves into the nuances of cultural sensitivity in emotional expression, highlighting nurses' need to navigate the diverse interpretations of emotional cues across different cultural contexts to avoid miscommunication and provide efficacious emotional support. Collectively, these insights illuminate the direct correlation between effective nurse-patient communication, addressing emotional needs, and the consequent enhancement of patients' emotional well-being, which is pivotal in their healing journey.

Jin et al. [107] explored the fusion of acoustic and lexical features to enhance ER in speech. They introduced novel representations such as Gaussian supervectors and the

"emotion vector" (eVector) feature, demonstrating significant accuracy improvements. Zhang et al.'s [42] review on TER delves into the evolution of DL techniques in feature extraction and multi-modal fusion strategies. They highlight the challenges in dataset biases and integrating new modalities, offering insights into few-shot learning and practical applications in the TER landscape.

Identifying and comprehending human emotions through TAC involves integrating text, visuals, and audio. This approach combines ML methodologies, natural language processing, computer vision, and audio signal processing, delivering a comprehensive system for ER and HCI. The field of artificial intelligence encompasses a broad spectrum of studies aimed at problem-solving, notably through the use of Graph Neural Networks GNNs. These networks are distinguished by their potent capacity to streamline and resolve complex issues, offering notable advantages over other AI models. Concurrently, a diverse range of methodologies and algorithms has been devised for classification and prediction tasks in the domain of AC. Despite these technological strides, the field grapples with substantial hurdles, particularly in effectively processing and interpreting both Euclidean and non-Euclidean data frameworks. Addressing these obstacles is imperative for improving the precision of emotion analysis. This is especially pertinent in the case of trimodal emotional data, which presents as irregular and multifaceted, posing additional layers of complexity to the analysis process.

2.1.3. Textual Analysis

The increased number of online social networking sites and e-commerce websites has contributed to a significant rise in user-generated textual data. The introduction of Textual Emotional Recognition aimed at identifying contextual sentiments or emotions that are represented directly or indirectly with user-generated data. The conventional methodologies employed in TER depend on the utilization of feature engineering to identify relevant elements that are associated with sentiment. Performing this activity might be perceived as demanding and time-consuming. DL-based models have the capability to do sentiment analysis directly from textual input [108] without the need for further preprocessing or feature engineering. Table 2.4. provides an overview of many significant techniques employed in the field of TER. The table includes information such as the year of publishing, the level of statistical detailing, the method of presenting features, the classifier employed, the database employed, and the corresponding performance parameters. Subsequently, a comprehensive exposition of the works of the TER is provided. From the TER methods analysis, we can see CNNs, LSTM, and Tokenization models performed better than other gathered ones. CNNs, LSTM achieved 95.90(%); 96.89(%) and Tokenization models achieved 98.7(%).

Table 2.4. Textual Emotional Recognition Methods

Authors	Year	accuracy(%)	Databases	Classifiers	Feature Representations
Maithri [109]	2022	86.21%	Movie Reviews	SVM-ELM	Multi-feature fusion
Onan [110]	2022	95.90%	Reviews	RCNNGWE	CNN, LSTM
Araque [111]	2023	86.32%	Emotions in Italian (EMit)	Logistic Regression	Linguistic features (text)
Lutsai [112]	2023	66.00%	Twitter Tweets	Metric-based	Multilingual features
Wang [113]	2023	86.35%	Laptops; Hotels	Softmax	PF-CNN
Li and Mao [114]	2023	87.52%	AM-30	Auxiliary Classifier	FA-GAN
Xiong [115]	2023	87.35%	Laptops; Hotels	ICMDP	Linguistic features (text)
Mahmoud [116]	2023	87%	Arabic Text Bank	Cross-Language Independent	Tokenization models

DL algorithms have substantial abilities to acquire and identify distinctive feature representations directly from the data. The effectiveness of DL sentiment analysis has been demonstrated by its effective application of word incorporation and the "expression of training data through multi-class categorization." Several DL methods for sentiment analysis have been proposed, including "deep convolutional neural network (ConvNet) learning and deep ConvNet-RNN learning."

LSTM-based learning: Authors [117][118] recommend applying LSTM to improve Emotion Detection and Action Recognition and classification. Their LSTMclassifier for ED was trained using text written in Urdu. They established the Urdu Nastalique Emotions Dataset (UNED) to identify sentiments in written Urdu. Many investigations into ED have been conducted in English alongside various languages. Figure 4.4. shows their suggested methodology for Urdu emotion identification using the UNED database.

Deep ConvNet learning: CNN-derived approaches have been utilized in various stages ofTER, which include "document-level, sentence-level, and aspect-level (or word-level)," employing various filters to acquire specific characteristics from the input data. In the study [119], the authors introduced a structure for sentence-level sentiment categorization that utilized the semantic lexical-augmented CNNs (SCNN) model. The model successfully integrated word details into the classification process. The study [120] utilized a highly advanced CNNs known as the Very Deep CNN (VDCNN). The network was designed to acquire hierarchical depictions of documents and effectively capture long-range dependencies in text analysis. In the field of aspect-level evaluation of sentiment, authors [121] introduced an innovative method that involves the integration of parameterized filters and parametrized gates to develop an aspect-specific convolutional CNNs.

Deep ConvNet-RNN learning: Researchers employed the integration of ConvNet-RNN to improve the effectiveness of sentiment analysis. The study [108] demonstrated the

integration of "CNNs and Bidirectional extended Short-Term Memory (BiLSTM)" models to enhance the performance of sentiment analysis. This approach combines the benefits of CNNs in capturing local information and BiLSTM in handling extended sequences, resulting in improved sentiment analysis findings. In the study [122], authors proposed the utilization of a "Gated Convolutional Network with Aspect Embedding (GCAE)" for enhancing the efficiency of Aspect-Category and Aspect-Term Sentiment Analysis (ACSA and ATER). The approach reduced the complexity of the training process.

The study [123] focused on the prediction of sentiment intensities. The researchers suggested a stacked ensemble approach that combines the outputs of multiple models, including LSTM, SVR, and GRU.

TheTER depends on conventional ML approaches, which encompass knowledge-based methodologies and statistical approaches. The first method involves the utilization of dictionary modeling techniques to handle comprehensive emotional terms, while the second approach depends on the presence of sizable databases that are characterized by polarities or emotional characteristics.

TER-based BiLSTM: The research [124] evaluated the multi-label emotion categorization challenges in both English and Chinese text using DL, including transfer learning-based algorithms. To address the issues of multi-label classification, they suggested a "BiLSTM in addition to n (the number of emotions in the dataset) multiple attentions and transfer learning multiple attentions-based models like XLNet-MA, DistilBERT-MA, and RoB-ERTa-MA." Their proposed methodology framework is mentioned in Figure 2.3..

TER based on Statistical: The statistical-based approach in TER involves a higher

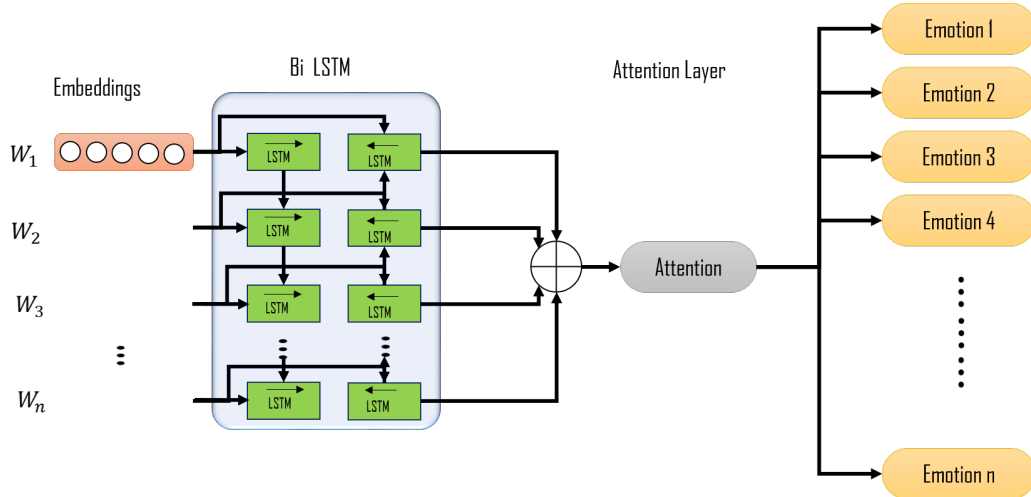


Figure 2.3. Architecture of BiLSTM [124]

priority on utilizing an identifiable dataset for training a ML classifier. This was accomplished by integrating previous statistics or posterior probabilities. Statistical-based methods for sentiment evaluation are considered more appropriate in comparison to lexicon-based techniques [125] due to their capacity to handle significant volumes of data.

Study [126] employed the semantic orientation of words as a basis for constructing a feature space, which was classified using an SVM that was created for the purpose. Research [127] introduced a novel model based on sub-graphs. The document can be described as an integration of sub-graphs, and the characteristics extracted from these sub-graphs are supplied into an SVM. The "Naive Bayes (NB) classifier" is a widely utilized and reliable method that operates under the assumption of feature independence in the dataset. This tool has the capability to select or exclude sentences that are inconsistent with comparative viewpoints.

TER based on Knowledge: The sentiment analysis technique known as knowledge-based TER utilizes dictionaries and linguistic principles, such as SentiWordNet, MPQA, and SenticNet. The effectiveness of lexicons in categorizing words as positive or negative is enhanced through the use of cultural standards. Authors of [128] introduced a comprehensive lexicon-based methodology that integrates external evidence and language rules. The researchers of [129] developed a sentiment analysis methodology that was specific to a particular region. The framework utilized lexical association data in its construction. Study [130] proposed an integrative methodology that incorporated computational intelligence, linguistics, and common-sense computing, with a particular emphasis on the significance of negation. Authors of [131] incorporated Twitter-specific attributes into a sentiment analysis system that utilized a knowledge-based approach and accommodated many languages. Conventional and precisely defined concepts limit knowledge-based models.

Numerous methods are available for textual analysis, ranging from traditional ML approaches to advanced DL models. LSTM networks, a type of RNNs, have proven to be particularly effective in handling the sequential nature of text data, capturing long-term dependencies, and contextual information, which is crucial for tasks like sentiment analysis and emotion recognition [132]. LSTM models excel in processing and understanding the nuances of human language over long sequences, making them indispensable for tasks requiring contextual memory. Their ability to remember previous words and phrases while analyzing current input enables them to outperform many other models in textual analysis [133]. This importance is reflected in various methods and impressive accuracy results, as shown in Table 2.5..

Table 2.5. Text Emotional Recognition Methods Using LSTM

Authors	Year	Method	Accuracy Result
Wang et al.	2016	LSTM for Sentiment Analysis on Twitter Data [133]	82.9%
Huang et al.	2017	LSTM with Attention Mechanism for Emotion Detection [134]	85.7%
Tay et al.	2018	LSTM with Hierarchical Attention for Aspect-Level Sentiment Analysis [135]	87.3%
Chatterjee et al.	2019	Multimodal Emotion Detection Using LSTM [136]	89.4%
Zhang et al.	2020	BiLSTM with Attention for Fine-Grained Sentiment Analysis [137]	90.2%
Yang et al.	2021	Hybrid CNN-LSTM for Sentiment Classification in Social Media Texts [138]	88.1%
Kim et al.	2022	Transformer-LSTM Model for Emotion Recognition in Text [139]	91.5%
Iyer et al.	2023	CNN-LSTM Ensemble for EEG-based Emotion Recognition [140]	97.16%
Hosseini et al.	2024	Multimodal Emotion Recognition Using Sound, Image, and Text Fusion [141]	82.9%

2.1.4. Audio Analysis

The process of emotional analysis of audio data entails the utilization of advanced computer algorithms to identify and evaluate the range of emotions present in audio recordings. Through the analysis of many characteristics, such as pitch, timbre, and syntax of speech, this methodology attempts to recognize and classify the emotions communicated in linguistic or sound expressions. The applications of this technology encompass an extensive range of tasks, such as sentiment analysis in calls from consumers and ER in audio. This emerging discipline combines the principles of ML and audio processing to provide a deeper understanding of human emotions by analyzing sound data.

: Audio Emotion Recognition (AER) is a computational technique designed to detect and classify emotions communicated in language signals. AER technologies utilize the capabilities of both ML and DL techniques to analyze retrieved features. The ML approach to AER emphasizes the analysis of sound characteristics and the selection of suitable classifications. The DL approach to AER utilizes a CNNs framework that operates in an end-to-end manner, eliminating the need for manual feature engineering. Table 2.6. presents a description of significant AER methods, encompassing significant studies, year of publication, feature representation, classification algorithm, database, and effectiveness. The subsequent discussion delves into these methodologies, which exemplify

progress in the field of emotional analysis from various viewpoints. From the AER methods evaluation, we can see 1D-CNN-LSTM-GRU performed efficiently, compared with other gathered methodologies, which achieved 90.47(%) in the CREMA-D database. Databases for voice recordings distinguish between natural and artificial voices. At first, experts favored less unstructured databases, particularly those collected from experienced actors like those included on "Belfast Induced Natural Emotion (Belfast)." Forty participants took part in five different tests relating to emotions, which were recorded at Queen's University in Belfast. Each trial includes recordings of people expressing anger, grief, joy, anxiety, and passivity [142]. These databases, which record professional outcomes, provide essential information for emotion analysis and associated applications by providing insight into well-defined emotional qualities.

The dataset (MVPD) possessed a total of 382 people, with 252 having regular voices and 130 having dysphonic vocals. These individuals were considered in the organized dataset called MVPD. Comprehensive data, such as voice examples, voice recordings, and acoustic evaluation, were collected for both groups. The OSELM's effectiveness was assessed and contrasted with other classifications based on its accuracy, sensitivity, and specificity in recognizing and distinguishing dysphonic vocals [143]. DL-based AER systems provide the capability of recognizing and analyzing con-textual elements and distinctive characteristics of emotional communication without requiring the development of a customized feature extractor. CNNs that incorporate auto-encoder structures are widely recognized among the employed methods for DL-based AER systems. RNNs and their variations, such as Bi-LSTM, have been used for the purpose of capturing temporal data. The approach for sentiment analysis incorporated a combination of ConvNets and RNNs, along with attention mechanisms. The adversarial training can be employed in DL-based AER to address the challenges presented by restricted data availability and low database quality. The approach involved enhancing the training data through methods of enhancement and reducing instabilities.

RNNs learning: RNNs have a significant competency in handling consecutive sound inputs. These networks are capable of preserving internal states throughout sequences by capturing and integrating short-term sound properties into comprehensive representations at the level of pronouncements. RNNs variations are capable of reducing label uncertainty by addressing the unpredictability of emotions in lengthy utterances. The utilization of an Extreme learning machine (ELM) as an utterance-level classification algorithm, in combination with Bi-LSTM, results in improved comprehension of temporal changes. Authors of [144] employed pretrained denoising AE and Bi-LSTM-RNN for identifying emotions. Researchers of [145] have introduced attention mechanisms in their work on Attention-3D-RNNs, which have demonstrated superior performance compared to the state-of-the-art AER models on the IEMOCAP and Emo-DB datasets. Notably, these models have shown outstanding results in terms of unweighted mean re-call. These improvements demonstrate the efficient application of emotional analysis techniques to

audio data in various circumstances.

Table 2.6. Audio Emotion Recognition Methods

Authors	Year	Performance (%)	Databases	Classifiers	Feature Representations
ML-based AER					
Xu et al. [146]	2021	58.80	IEMOCAP	Pooling	DNN/RNN
Ahmed et al. [147]	2023	90.47	CREMA-D	1D-CNN-LSTM-GRU	Audio
Kaur and Singh [148]	2023	43.80/79.10	LDC/EmoDB	SVM	Audio features
Bhosale [149]	2023	68.00	Emo-DB	SVM	MFCC
DL-based AER					
Gao et al. [150]	2023	58.38	IEMOCAP	SVM	Adversarial AE
Mustaqeem et al. [151]	2023	84.30	Emo-DB	FC	CNN

Two-Stream DL: The research [152] compared the effectiveness of a two-stream-based emotion recognition system utilizing Bi-LSTM and CNNs using a Korean-utilized emotion dataset. "Bi-LSTM and YAMNet, a CNN-based transfer learning model," were combined in a two-stream configuration to create an emotion identification model. The effectiveness of voice extraction of features and DL models was examined. ConvNet learning: The study [153] employed a semi-CNN to extract characteristics from the spectrogram data. These features were entered into an SVM with different contexts for the purpose of AER. Study [154] developed an improved version of the "AlexNet DCNN" that was pre-trained on the ImageNet dataset. They incorporated a "discriminant temporal pyramid matching (DTPM)" technique to generate a comprehensive feature representation at the utterance level. "Deep ConvNet models" are employed for the purpose of extracting emotion features from unprocessed audio input in the context of multi-task AER.

ML-based AER includes effective feature representation training and classification methods for the purpose of predicting emotions. Robustness is achieved by combining a range of audio parameters, such as prosodic, spectral, and voice-quality characteristics. OpenSMILE is recognized as a significant toolkit utilized for the extraction of audio features. Efficiency variations in AER can be attributed to the utilization of various model-feature configurations.

Gaussian Mixture Models (GMM) and their various modifications are considered to be advanced forms of constant Hidden Markov Models (HMM) that are ideally suited for AER based on global features. Authors [155] introduced a GMM methodology that integrates prosodic, audio quality, and MFCC [156] characteristics. SVM employs kernel algorithms to map emotion vectors into a higher-dimensional space, so producing an ideal

classification through the use of a maximum intervals hyperplane.

Zhao et al. [157] developed a spatial CNN combined with an attention-based BLSTM for extracting deep-spectrum features. These features were then merged and input into a DNNs to predict emotions, making audio classification or representation features more available under the DNNs, as shown in Table 2.7..

Table 2.7. Audio Emotional Recognition Methods Using DNN

Authors	Year	Method	Accuracy Result
Burnett et al. [158]	2014	Optimized multi-channel DNNs with 2D graphical representation of acoustic speech features	Significant improvement in emotion recognition accuracy 92%
Han et al. [159]	2014	DNNs classification for final utterance-level features	Effective feature extraction and classification for SER, which shows an accuracy of 87%
Mirsamadi et al. [160]	2017	Recurrent neural networks with local attention	Achieved a 20% relative improvement in Speech Emotion Recognition (SER), resulting in a new accuracy rate of 96%.
Ortega et al. [161]	2019	Multimodal fusion DNNs for audio, video, and text	The CCC values are 60.6% for arousal, 53.4% for valence, and 17.0% for liking.
Heracleous et al. [162]	2020	Feature extraction in DNN-based SER	Highest accuracy of 58.49% in noisy data conditions
Zhang et al. [163]	2021	Pre-trained DCNN with attention mechanism for emotion features	Improved recognition performance with attention layer, achieving a UAR of 87.86% on EMO-DB and 68.50% on IEMOCAP
Sun et al. [164]	2022	Multi-class speech emotion recognition using two-stage bottleneck features and multi-classifier joint decision (MCJD) algorithm	Average accuracy of 87.08% on Berlin Database
Pan et al. [156]	2023	Combination of 1D CNN, LSTM, and DNNs with data augmentation	SThe DNNs method achieved a superior recognition rate of 96.21% on the IEMOCAP database.
Jutono G. et al. [7]	2023	Multi-feature extraction using DNNs	Accuracy of 93.61%, sensitivity of 73.80%, and specificity of 96.34%

DNNs have proven their efficacy in audio emotion recognition and sentiment analysis across various studies, as demonstrated in Table 2.7.. Burnett et al. (2014) optimized a multi-channel DNNs with a 2D graphical representation of acoustic speech features, significantly improving emotion recognition accuracy to 92%. Han et al. (2014) applied DNNs classification for final utterance-level features, resulting in an 87% accuracy rate, showcasing the model's capability in processing speech features. Mirsamadi et al. (2017) utilized recurrent neural networks with local attention, achieving a 20% improvement in SER with an accuracy of 96%. Ortega et al. (2019) integrated audio, video, and text data using a multimodal fusion DNNs, achieving CCC values of 60.6% for

arousal, 53.4% for valence, and 17.0% for liking, indicating the potential of multimodal approaches. Heracleous et al. (2020) focused on DNN-based SER in noisy conditions, achieving an accuracy of 58.49%. Zhang et al. (2021) improved recognition performance with a pre-trained DCNN and attention mechanism, achieving UARs of 87.86% on EMO-DB and 68.50% on IEMOCAP. Sun et al. (2022) employed a multi-class speech emotion recognition method using two-stage bottleneck features and an MCJD algorithm, achieving an average accuracy 87.08% on the Berlin Database. Pan et al. (2023) combined 1D CNN, LSTM, and DNNs with data augmentation, resulting in a 96.21% recognition rate on IEMOCAP. Jutono G. et al. (2023) achieved an accuracy of 93.61%, sensitivity of 73.80%, and specificity of 96.34% using multi-feature extraction with DNNs. These studies highlight the substantial advancements and high accuracy rates that DNNs can achieve in emotion recognition, emphasizing their transformative impact in the field.

2.1.5. Visual Analysis

Experts employ modern technology to analyze individual feelings in visual data photographs and video clips. Emotions like happiness and surprise can be determined by analyzing a person's facial expressions, body language, and other surroundings using computer vision and ML algorithms. Applications in user experience architecture, research on markets, and psychological examinations are enhanced by emotional analysis' ability to analyze visual data and reveal latent reactions. : Face Emotional Analysis (FEA) refers to technological approaches that use algorithms that identify faces to read and determine human emotions by analyzing facial expressions. Through the examination of notable features such as face facial muscles and actions, the objective is to detect a range of emotions. Table 2.8. presents an overview of FER techniques, encompassing significant studies, feature representations, classification algorithms, a database, and effectiveness. Also, we can see the Visual emotion recognition become more attractive with the CNN model as shown in Table 2.9.

The datasets related to FEA, the JAFFE dataset comprises a total of 213 photos depicting seven distinct facial movements. These expressions were captured through the collaboration of 10 Japanese models. The "extended Cohn-Kanade (CK+)" dataset was constructed from CK+. The "Oulu-CASIA NIR-VIS dataset, also known as Oulu-CASIA" [177], was published in 2011. It comprises a collection of 2,880 image sequences acquired using two distinct imaging systems under three different illumination circumstances. Due to the limited availability of 3D emotional talking face data, the researchers depended on employing facial blend shapes for supervision. The approach allows for creating realistic 3D faces from 2D emotional data. As an outcome, they have developed a comprehensive dataset named 3D-ETF, which can be used to train the network [178]. Many methods of FEA based on ML and DL, which are based FEA infrastructure, predominantly originate from recognized pre-trained ConvNets such as VGG, GoogLeNet, and ResNet. Therefore, the approach of DL-based FEA can be categorized into two main components:

Table 2.8. Face Emotional Recognition Methods

ML-based FER				
Authors	Year	CASE II (%)	BU-4DFE (%)	Features
Wang	2023	267.37	—	MDMO
Paronett	2023	—	—	FLP
Mohite	2023	45.75/166.80	—	LBP-MOP
Jabberi	2023	—	90.12	2D+3D maps
Banerjee	2023	—	—	IFSL
Liu	2023	—	95.13	DSF
DL-based FER				
Authors	Year	CASE II (%)	BU-4DFE (%)	Features
Huang	2023	—	83.32	FLM-CNN
Yang	2023	42.70	—	GCN-CNN
Tang	2023	84.10	—	STRCN
Razak and Sahran	2023	86.80	—	ICE-GAN

"ConvNet-RNN learning for FEA and adversarial learning for FEA," which takes into consideration the variations in network designs.

Adversarial learning: Generative Adversarial Networks (GANs) perform a crucial role in generating synthetic face expression photographs, encompassing a wide range of stances and viewpoints. Study [179] introduced a GAN with an Autoencoder FEA architecture to conduct FEA that was invariant of position, as mentioned in Figure 2.4.. The proposed model aims to generate facial images showcasing a variety of expressions, regardless of the specific pose. To minimize the impact of semantic disturbances on prediction accuracy, the authors of [180] introduced a Semantic Neighbourhood Aware (SNA) network. This network features an asymmetric FEA with added noise to improve performance. The

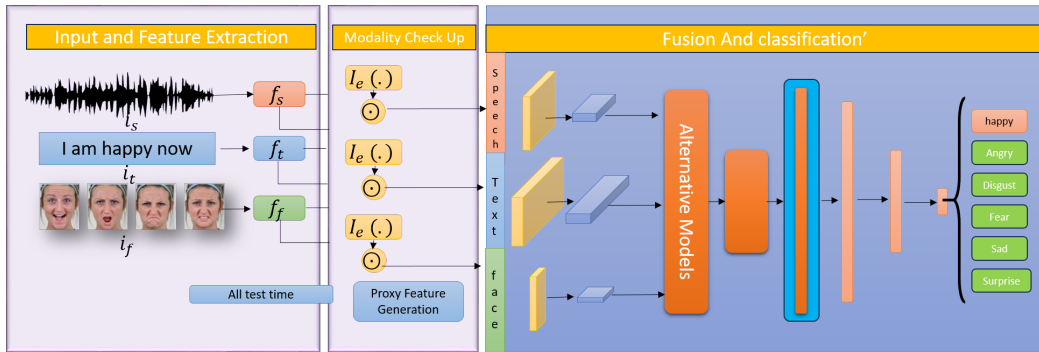


Figure 2.4. (FEA) learning [179].

researchers and authors of [181] developed a framework named "Identity-Aware and Capsule-Enhanced GAN (ICE-GAN) for recognizing and synthesizing facial mi-cro-expressions." The proposed framework combined an "AE-based generator with a

Table 2.9. Face Emotional Recognition Methods Using CNN

Authors	Year	Method	Accuracy Result
Pranav et al. [165]	2020	Deep Convolutional Neural Network (DCNN)	Good accuracy in classifying of 78.04%.
Adil [166]	2021	CNN model trained with grayscale images from the FER 2013 dataset	80% accurate for four emotions (happy, sad, angry, fear) and 72% accurate for five emotions (happy, sad, angry, neutral, fear)
Banerjee [167]	2021	Efficient deep learning technique for facial emotion recognition	Accuracy of 95.65% for emotion recognition
Sana et al. [168]	2022	CNN algorithm for facial emotion recognition	Improved emotion recognition accuracy 73.02%
Akram et al. [169]	2023	Pre-trained CNN (AlexNet, VGG-16) with SVM and Ensemble classifiers	Improved accuracy by 65.5%
Qu et al. [170]	2023	CNN using PyTorch	Significant enhancement in emotion recognition performance 60.20%
Tataji et al. [171]	2023	Cross Connected CNN (CC-CNN) using feature level fusion	High accuracy in facial expression recognition
Mohana et al. [144]	2023	Hybrid CNN-LSTM Network	achieves an overall accuracy of 94.61% in recognizing facial expressions.
Singh et al. [172]	2023	Hybrid CNN and ConvLSTM	The combination of the two methods results in an accuracy of 91.28%
Smitha et al. [173]	2023	Ensemble CNN methods for video emotion recognition	Achieved 92.07% recognition accuracy
Saravanan et al. [174]	2023	CNN model with six convolutional layers, two max pooling layers, and two fully connected layers	Final accuracy of 60%
Soultana et al. [175]	2023	Enhanced CNN-Based Model	Improved accuracy of 98.41%, precision of 98%, F1-score of 94%, and recall of 96%
Balk et al. [176]	2023	CNN-LSTM based image captioning	BLEU Score of 39%

Capsule-Enhanced Discriminator (CED)."

ConvNet-RNN: Video temporal correlates have a significant role in capturing dynamic face sequences. RNNs, such as LSTM models, efficiently capture temporal features in data. On the other hand, CNNs improve at learning spatial features from expression-state frameworks. ConvNet-RNN utilizes transmitted fusion or ensemble techniques in the context of FEA. The end-to-end FMER architecture was presented by the authors of [182], who involved the cascade of spatial features into LSTM. The study [172] introduced a novel deep evolutionary spatial-temporal network. This system is composed of two main components: a "part-based hierarchical bidirectional RNNs (PHRNN) and a multi-signal convolutional neural network (MSCNN)." The PHRNN is designed to analyze temporal facial movement data, while the MSCNN focuses on capturing physical appearance.

Geometry: In the study [183], authors employed 52 : Facial Landmark Points (FLP) to determine the geometric locations and degrees of facial features. These FLP were utilized for automatic FEA in the sequencing of facial images. A fuzzy logic model was used to analyze important aspects of AAM (Active Appearance Model) and identify facial expressions accurately. The recognition process was informed by previous information obtained from the "facial action coding system (FACS)." The "local prominent directional pattern descriptor (LPDP)" was developed to solve the instability problem from descriptions in the geometries of various distorted local frameworks.

In the research [169], the authors analyze several ML methods for classification by extracting learned features from a pre-trained CNN. Assessments were performed on two popular CNN architectures: AlexNet and VGG-16, using the MLF-W-FER dataset of masked facial expressions. Using the AlexNet and VGG-16 constructions, their research showed that the SVM and Ensemble classifiers performed better than the SoftMax classifier.

The significance of physiological emotion detection using ECG methods based on DL and ML cannot be overstated. A study by [184] introduced an innovative structure called *EmotionalGAN* to enhance the generalization capability of emotion recognition models. This improvement was achieved by integrating enhanced ECG data generated by "EmotionalGAN" into the emotion recognition model. Using GAN-based models in emotion recognition tests resulted in a notable gain of approximately 5% in median accuracy, underscoring their superiority compared to using only the initial data.

The study [185] employed a methodology to identify "human emotional stress by extracting three time-domain variables, namely heart rate, mean R peak amplitude, and mean R-R intervals." An alternative approach involves converting time-domain ECG features into corresponding features in different domains. In the study conducted by researchers [186], various signal processing techniques such as "Fast Fourier Transform (FFT), Discrete Wavelet Transform (DWT), and Hilbert Huang Transform (HHT) were employed to convert ECG signals into frequency-domain characteristics." Subsequently, Principal Component Analysis (PCA) and Tabu search algorithms were utilized to

identify significant features within the low-frequency, high-frequency, and combined (low and high) frequency ranges.

In contrast to other peripheral neurophysiological signals, EEG has the unique ability to directly assess alterations in brain activity, hence providing information into the internal characteristics of mental and emotional states. In addition, EEG with an increased temporal resolution enables the real-time monitoring of an individual's emotional state. Recently, several approaches for emotion identification have been developed based on EEG signals.

The study [187] developed a regularized graph neural network (RGNN) that measures the biological topology between various brain regions, enabling the examination of local and global relationships between various EEG channels. Authors of [188] introduced the "regional-asymmetric CNN (RACNN), which incorporated temporal, regional, and asymmetric extractors" of features to leverage spatial data from nearby and symmetric channels. Three feature ex-tractors are enhanced by incorporating an asymmetric differential layer to gather the discriminative information. This was achieved by taking into account the asymmetrical attribute of emotional reactions. The Recurrent Attention CNN (RACNN) demonstrates notable outcomes, with average accuracies of 96.88% and 96.28%.

The efficacy of ML-based emotion recognition using EEG data is contingent upon a suitable layout of feature extraction and methods for categorization. The SVM or its derivatives are frequently employed as machine learning-based classifiers for emotion classification using EEG data [189]. Researchers of [190] conducted a study in which they integrated the mutual information-based EEG selection of features method with SVM to enhance recognition accuracy.

2.1.6. Graph Neural Networks an Overview

GNNs have emerged as a prominent area of study in recent years due to their capacity to manage large datasets effectively. Unlike traditional neural networks that operate on structured data, GNNs are specifically designed for unstructured data represented by graphs. They can directly perform node-level, edge-level, and graph-level predictions on graph data.

Among the various types of GNNs, GCNs have received considerable attention owing to their similarity to traditional CNNs [191, 192, 193]. GCNs find extensive applications in computer vision, natural language processing, social network analysis, and even in disciplines like physics, chemistry, and biology [194].

In the domain of computer vision, GCNs have been applied to several tasks including action recognition [195, 196], face clustering [197], object detection [198], visual reasoning [199], and image generation using scene graphs [200, 201]. GCNs can be broadly classified into spatial and spectral, which will be elaborated upon in the following sections.

The development of our algorithm is driven by several critical advancements and

Table 2.10. Review of GNN Methods and Applications (2018 - 2020)

Author	Methods	Brief Information	Gaps	Future Directions
Zhitao Ying et al. (2018)	GNN	Overview of methodologies	Practical application challenges, scalability issues	Improve scalability, deployment
Amir Zadeh et al. (2018)	Multimodal analysis	CMU-MOSEI dataset challenges	Handling diverse and noisy data from real-world settings	Enhance preprocessing, robust analysis
Didan Deng et al. (2018)	Multimodal analysis	Integrates visual, audio, text	Handling long-term dependencies, synchronization of modalities	Improve dependencies, synchronization
Dong Zhang et al. (2019)	GCNs	Context and speaker dependencies	Scalability and real-time application	Enhance scalability, real-time
Devamanyu Hazarika et al. (2019)	Dialogue-based GCNs	Contextual conversation modeling	Handling long-term dependencies and varied dialogue structures	Improve dependencies, dialogue structures
Xueyan Wang et al. (2019)	GNNs for image understanding	Integrates visual cues	Integration of complex visual features and scalability	Enhance integration, real-time applications
Zheng Lian et al. (2020)	GNNs, Self-Attention	Contextual conversation modeling	Handling large numbers of speakers	Optimize structure, multimodal data
Ziniu Hu et al. (2020)	GNNs	Generative pre-training techniques	Generalization across graph types and tasks	Explore transfer learning, adaptation
Shun Zhang et al. (2020)	GNNs and IoT	GNN applications in IoT	Practical deployment challenges, energy efficiency	Address deployment, energy efficiency
Chen Ma et al. (2020)	Survey on GNNs in recommender systems	GNNs in recommendation systems	Scalability, diverse recommendation scenarios	Enhance scalability, diverse scenarios
Samitha Liyanage et al. (2020)	Multimodal fusion	Integrates data modalities	Synchronization of multimodal data, computational complexity	Improve synchronization, computational efficiency
Bhaskar Kumar Mishra et al. (2020)	Heterogeneous GRNN	Emotion recognition integration	Handling large-scale data, real-time processing	Enhance scalability, real-time applications
P. Venka. et al. (2020)	Multimodal ER	Complex interaction recognition	Handling diverse and noisy data sources	Enhance preprocessing, additional modalities
Zhuo Chen et al. (2020)	Intra- and inter-modal	Modal dependencies exploration	Integration of long-term dependencies	Improve dependencies, real-time applications
Shuliang Gao et al. (2020)	Multiscale GNNs	Multiscale learning framework	Handling large-scale and incomplete datasets	Enhance scalability, incomplete data
Bhaskar Kumar Mishra et al. (2020)	Deep GCNs	Deep graph convolution	Synchronization and integration of multimodal data	Improve synchronization, additional modalities
Zhen Han et al. (2020)	Pair-based view GNNs	Knowledge graph reasoning	Scalability, complexity with large knowledge graphs	Improve scalability, efficient algorithms
Esma Mansouri-Benssassi et al. (2020)	SNNs, GCNs	Multisensory emotion recognition	Integration of temporal and spatial data in real-time	Enhance real-time processing, integration

Table 2.11. Review of GNN Methods and Applications (2021 - 2024)

Author	Methods	Brief Information	Gaps	Future Directions
Adam Gibson et al. (2021)	Various GNN architectures	Comprehensive tutorial on GNNs, covering concepts, architectures, and practical applications.	Practical challenges in specific domains	Address domain-specific challenges, optimize for large-scale datasets
Wei Wang et al. (2021)	GNNs for malware detection	Detects Android malware using GNNs, leveraging structural information of app function call graphs.	Handling obfuscated or disguised malware	Improve robustness against obfuscation, explore real-time detection
Young-Ju Choi et al. (2021)	Residual-based GCN	Proposes a residual-based GCN for emotion recognition in conversations within Smart IoT, leveraging residual connections.	Scalability, real-time implementation in IoT systems	Improve scalability, explore real-time applications in diverse IoT environments
Liyan Zhang et al. (2022)	Text pretraining, heterogeneous GNN	Combines BERT for text pretraining and HGAN for hotel review classification, integrating user preferences.	Handling colloquial review texts, scalability	Address colloquial text challenges, improve scalability
H. F. T. and Das (2024)	Contextualized GNNs	Integrates trimodal emotion recognition, significantly improving.	Integration and synchronization of multimodal data	Enhance data integration techniques, explore additional modalities

challenges highlighted in recent literature, as shown in the above 2.10., and 2.11.. [202] extensive review (2018) on GNNs underscores their versatility and potential in managing complex data structures across various domains, establishing a robust foundation for further innovations in this field. Additionally, [203] work on DialogueGCN (2019) demonstrates the efficacy of GNNs in modeling context and speaker dependencies in conversations, achieving state-of-the-art performance in emotion recognition tasks. This underlines the foundational strengths of existing GNNs frameworks, providing a solid basis for the development of our algorithm. Moreover, the challenges of multimodal data integration are significant, as evident in studies like [204], which introduce a graph network-based multimodal fusion technique to enhance machine learning performance. The authors highlight the synchronization and integration of multimodal data as major hurdles, emphasizing the need for advanced algorithms like CA-WGNN that can effectively weigh and integrate various data attributes. The Synch-Graph approach by [205] utilizes neural synchrony and graph convolutional networks for multisensory emotion recognition, illustrating the benefits of capturing and leveraging complex interdependencies among different data modalities, further highlighting the complexity of the problem addressed by CA-WGNN. Additionally, the introduction of [206], which employs a two-stream heterogeneous graph recurrent neural network for emotion recognition, validates the necessity of sophisticated methods to handle large-scale and real-time data processing. These insights collectively underscore the compelling need for

an advanced algorithm like CA-WGNN. This algorithm builds upon the foundational strengths of existing GNNs frameworks and addresses the integration and synchronization challenges inherent in multimodal data. Consequently, CA-WGNN has the potential to significantly advance the field of sentiment analysis and emotion recognition, emphasizing the importance and impact of this work.

The thesis used GNN and multimodal emotion recognition, not randomly; some papers encouraged us to make the AC-WGNN algorithm. Despite the substantial advancements in emotion recognition using Graph Neural Networks (GNNs), several gaps and challenges still need to be addressed. Many existing methods need help efficiently handling multi-modal data, the complexity of managing interactions between different modalities, and the computational overhead associated with sophisticated fusion techniques. For instance, while [207] and [208] made significant strides in emotion detection and multi-modal fusion, their models face challenges with complex data interactions and high computational costs. Similarly, [209] and [210] identified issues with stacking multiple GNN layers and correcting erroneous predictions in dialogues. Still, these solutions require extensive annotated datasets and complex training processes.

Our groundbreaking article, *"TER-CA-WGNN: Trimodel Emotion Recognition Using Cumulative Attribute-Weighted Graph Neural Network,"* is poised to revolutionize the field. We aim to address these persistent gaps by introducing a novel approach that integrates textual, audio, and visual data through a Cumulative Attribute-Weighted Graph Neural Network (CA-WGNN). Using cumulative attribute weighting in our GNN model is a unique design choice that enhances the accuracy and efficiency of emotion recognition. By assigning weights to attributes cumulatively, our approach captures the importance of different features across modalities, providing a more nuanced and dynamic data integration. This method effectively addresses the limitations of static window sizes and complex hierarchical structures identified in previous studies by [211] and [85]. Moreover, the cumulative attribute weighting simplifies the fusion process, reducing computational overhead and making the model more practical for real-world applications.

Our approach dynamically adjusts to varying data instances by focusing on cumulative attributes, multi-modal extraction features, and embedding, enhancing the model's adaptability and performance without needing large annotated datasets and extensive training times. This algorithm not only improves the scalability of our model but also ensures that it can affect diverse and dynamic real-world scenarios, such as customer service interactions and therapeutic settings. Thus, our research fills the identified gaps by providing a more efficient, accurate, and scalable solution for multi-modal emotion recognition, building upon the foundational work of previous studies while addressing their limitations. Rest assured, our model is not just a theoretical concept but a practical tool ready to be implemented in real-world settings.

Table 2.12. Review of GNN Methods and Applications (2019 - 2022)

Author	Year	Problems Addressed	Accuracy Results	Gap Identified
H. Chen et al. [207]	2019	Emotion detection in conversations with party-specific state tracking.	Significant margin improvement on IEMOCAP and AVEC	Complex inter-party dynamics
Y. Ren et al. [208]	2019	Fusion of multiple modalities using graph networks.	State-of-the-art on IEMOCAP dataset	Computational overhead fusion
S. Poria et al. [209]	2020	Context-sensitive and speaker-sensitive emotion recognition.	1%~2% improvement over state-of-the-art	Limitations in GNN stacking
P. Shenoy et al. [210]	2020	Correcting erroneous predictions in dialogues.	Significant improvement, metrics not specified	Complex training process
A. Ghosal et al. [204]	2020	Combining multiple modalities for emotion recognition using graphs.	3-5% improvement on benchmark datasets	Large complex graphs
Z. Zhang et al. [212]	2020	Integration of multiple modalities for emotion recognition.	Outperforms state-of-the-art on IEMOCAP	Managing modality interactions
Z. Zhang et al. [211]	2021	Sentiment analysis in heterogeneous data reviews.	83.21% F1 Score on IEMOCAP	Dynamic window selection
Y. Hu et al. [213]	2021	Integration of multiple modalities in dialogues.	F1 scores up to 84.50%	Computational fusion overhead
X. Ma et al. [85]	2021	Hierarchical graphs for emotion recognition across modalities.	Improved accuracy over baseline models	Managing hierarchical graphs
W. Liu et al. [214]	2022	Understanding multimodal conversational context.	Outperforms methods on IEMOCAP and MELD	Contextual interaction modeling

2.2. Comprehensive Analysis of Trimodal

Data acquisition process could enhance the precision of emotion detection compared to using a single modality or a multi-modal approach with emotion recognition and SA. Several studies have evaluated the effectiveness of trimodal fusion in AI. These and other studies suggest that trimodal fusion can be a practical approach for AC, particularly in naturalistic and conversational settings. However, it is essential to note that the performance of trimodal fusion systems can vary depending on the specific implementation and the characteristics of the data being analyzed. Further research is needed to understand better the strengths and limitations of trimodal fusion in AI. The concept of MSA refers to the study of emotion in multi-modal scenes. Emotional information in videos is analyzed using trimodal information. Text modal information is extracted from the video using the speaker's speech content. [96], [38]. The present study critically analyzes the methodologies for procuring and initializing data for a future in-depth examination. It highlights the essentiality of precisely acquiring data from a broad range of sources, verifying data authenticity, and managing predicaments such as missing data, background disturbances, and inconsistencies. Data preprocessing methods, including data purification, standardization, and alteration, are extensively analyzed, emphasizing their integral functions in fortifying the credibility and reliability of the final assessment. The discussion further elucidates the significance of a meticulously engineered data preprocessing channel within the framework of data-informed decision-making and

predictive modeling.

This investigation is in sync with the current trajectories in AC, predominantly investigated via ML or DL grounded methodologies. These methods aspire to decode and comprehend affective indicators across a myriad of affective modalities. A noticeable upward trend towards exploiting DL in affect recognition is observable throughout all the reviewed literature in the tables, especially those that assess methodologies founded on AI principles. A substantial fraction of the literature on Multi-modal affective analysis emphasizes the indispensability of Multi-modal fusion methods [215]. However, it is crucial to acknowledge that Multi-modal affective analysis can extend its scope by integrating various modality combinations. As such, using many methodologies for categorization and prediction in both emotion recognition and sentiment analyses becomes feasible by introducing an innovative technique that validates the assessment in this domain.

Speech models differ in emotional expression due to the frequency spectrum of each interval. Facial expressions have been analyzed in each frame of the video. In this modality, facial expressions are primarily studied. The results of classification are ranked according to their probability. A smile can express sadness, and vice versa, while people may speak angry words if they are highly fearful [216]. Using MER, features from different modalities are combined for classification, and the probability distributions of sentiment categories are estimated using this information. A growing number of social media users means more multi-modal information (text, visual, and audio) is available to understand emotions. To keep up with the rapid progress of AI, an interface needs to gather data from multiple sources and determine attitudes based on that data. For building emotionally intelligent machines, ER is critical for detecting emotions from enriched input signals like texture, speech, and visual or multimedia signals. This article explores how these three modalities interact to influence emotional polarity.

Human affective datasets fall into two categories: non-spontaneous and spontaneous. Originally, non-spontaneous datasets mainly stemmed from actor performances, viewed as reliable due to their skill in expressing emotions. For instance, the Berlin Database of Emotional Speech (Emo-DB) [217] contains utterances by ten actors, showing emotions like happiness and fear. Yet, these emotions might be over-emphasized. To bridge this gap, spontaneous datasets emerged, like the Belfast Induced Natural Emotion (Belfast) [218]. Derived from forty participants at Queen’s University, UK, it features video clips showcasing five emotional states. This multi-dataset method values concurrent data collection.

Creating a trimodal dataset AC system requires setting up sensors or other devices to collect data from audio, visual, and physiological modalities. Using a microphone to record spoken words or different sounds would be an example of how the audio modality might be captured [39]. Regarding the visual aspect, a webcam or another camera could capture static or moving images. Sensors can measure heart rate, skin conductance, and other physiological signals from the physiological modality. It lists datasets associated with the

review in Table 2.13. to make the explanation more understandable.

Table 2.13. Overview of Datasets in Affective Computing

Dataset	Year	Source	Videos	Speakers	Language	Purpose
IEMOCAP [219]	2008	Actors	90	10	English	Emotion recognition in scripted and improvised performances
ICT-MMMO [220]	2013	YouTube, ExpoTV	20	370	English	Sentiment analysis in movie reviews
MOUD [221]	2013	YouTube	40	80	Spanish	Multimodal sentiment analysis in product reviews
MOSI [222]	2016	YouTube	93	89	English	Sentiment intensity in spoken opinions
NNIME [223]	2017	Recordings	44	6701	Chinese	Emotional expression in Chinese dialogues
CMU-MOSEI [224]	2018	YouTube	3228	1000	English	Multimodal sentiment and emotion analysis
MELD [225]	2019	TV Series	13000	Multiple	English	Emotions and sentiments in TV series dialogues
MEmoR [226]	2020	Video Clips	5502	70	English	Emotion detection focusing on non-verbal cues
CH-SIMS [70]	2020	TV Series	2281	474	Chinese	Sentiment analysis in video segments
CMU-MOSEAS [227]	2021	YouTube	4000	1645	Spanish, English, German	Sentiment and emotion analysis in diverse languages
MuSe-CaR [228]	2021	YouTube	291	70	English	Multimodal sentiment analysis in car reviews

The Multi-modal Corpus of Sentiment Intensity CMU-MOSI dataset collects 2199 opinion video clips. Each opinion video is annotated with sentiment in the range $[-3,3]$. The dataset is rigorously annotated with labels for subjectivity, sentiment intensity, perfume and per opinion annotated visual features, and per-milliseconds annotated audio segments [37], [229], [230], [67]. The Speech Analysis and Interpretation Laboratory developed Interactive Emotional Dyadic Motion Capture "Interactive Emotional Dyadic Motion Capture" IEMOCAP [231]. Actors perform emotive scripts and improvised scenarios designed to evoke five specific emotions. The emotions and movements of actors are detailed by marking their faces and gestures. The utterance level, discrete categorical annotations, and continuous attributes are used to label. It makes a random extraction from the training set with a ratio of 0.2 for validation. ICT-MMMO [149] includes 343 movies reviewed on YouTube. Each video contains the subject standing before the camera and expressing his viewpoint on a specific film. A solid positive sentiment, a weakly positive sentiment, a neutral sentiment, a strongly negative sentiment, and a weakly negative view are all labeled.

The Neural Network-based Integrated Multi-modal Emotion [223] framework, launched in 2017, amalgamates diverse virtual, textual, and auditory data into a dataset that showcases a spectrum of emotions. Comprising 6,701 unique samples from 44 Drama major participants at the National Taiwan University of Arts, this dataset emphasizes

monolingual Chinese dialogues and provides continuous emotional dimensions, including valence and arousal. These participants, aged 19-30, were grouped into various gender dyads and engaged in natural conversations in a controlled setting. Recordings utilized Bluetooth wireless microphones and a SONY HDR-P J790V camcorder, capturing both audio and high-definition video. Emotion annotations used the Feeltrace software, recording at a 1-second interval. Furthermore, physiological data was captured using ECGs with a specialized Texas Instruments chip, and sessions were meticulously transcribed. The dataset is accessible at <http://nnime.ee.nthu.edu.tw/>.

This approach incorporates audio features typical of emotion recognition using speech and visual elements to encode valuable valence information. SA has been studied extensively in the past decade, specifically automated review analysis. MOUD contains 80 product reviews in Spanish. The dataset was gathered from YouTube using keywords that would lead to product reviews and recommendations for the dataset. Instead of targeting just one product type, we used various product names to give the dataset some generality [232].

The MELD has been enhanced and extended to encompass audio, visual, and dialogue instances. It was created by improving and expanding the existing Emotion Lines dataset. MELD contains over 1400 dialogues and 13000 utterances from the Friends TV series. Multiple speakers facilitated the dialogues. There is also a sentiment annotation (positive, negative, and neutral) for each phrase and seven emotion classes for each utterance in conversation [233].

The CMU-MOSI dataset was created by Amir and contained text, video, and audio. It is the world’s largest multi-modal dataset. The video labels cover 250 hot topics and come from 1000 different speakers. In each video transcription, the audio and phonemes grades are aligned. The video was assessed using a crowdsourcing process on three Amazon Machinery Turkey platforms. The Ekman criterion categorized emotions: happiness, sadness, anger, fear, disgust, and surprise. A random selection of topics and monologue videos is used for each utterance. It is transcribed and appropriately punctuated [84]. The Chinese single and MSA dataset (CH-SIMS) contains annotations for 2,281 refined video segments that are both multi-modal and independent. Researchers can use separate unimodal annotations to study modal interactions or perform unimodal sentiment analysis. In this framework, late fusion serves as a baseline for learning multi-tasks. Several experiments showed that the methods could understand more distinct unimodal representations using CH-SIMS [70]. Harvesting opinions from the Web database L. Morency et al. [67] A dataset of online social videos has been created to address this new research direction. YouTube videos are divided into 13 positives, 22 neutrals, and 12 negatives. They encompass the diverse aspects of SA, such as diversity, multi-modality, and ambient noise.

The dataset MSA in Car Reviews (MuSe-CaR). MSA in the Wild is a multi-modal (text, audio, and video) dataset collected to understand MSA. For example, sentiment has been

linked to a topic or entity during product reviews (e.g., automobile reviews). Between the mid-20s and late 50s are the estimated ages of professionals, semi-professionals (influencers), and casual reviewers. A small percentage of our learners are non-native English speakers fluent in English despite being non-natives [77]. MER in videos is a dataset dedicated to Multi-modal emotion detection in videos and is distinct in its focus on video-based emotional analysis. It's an innovation that annotates detailed emotions for speakers and those without dialogue. MEmoR comprises 5,502 video clips and 8,536 data samples sourced from the renowned TV series "The Big Bang Theory." It features annotations specific to individuals, capturing 14 emotions as identified in Plutchik's emotion wheel. In these clips, individuals without dialogue lack audio-textual cues and often blend in visually, while those speaking might have limited visual details. MEmoR offers a realistic and demanding environment for testing. The proposed MEmoR dataset includes both short-term scenarios and external data, catering to a range of analytical approaches.

To produce the eNTERFACE database, 42 subjects of 14 nationalities were asked to complete six stories designed to evoke a particular emotion (Ekman's six basic emotions). The experimenter explained the six situations (stories) to the participants. A tale that aroused anger: You live in a foreign country. Two human experts judged each level to determine if it clearly expressed an emotion. It uses visual feeling, textual emotion, and audio emotion datasets. This town only has one bank, which is open till 4 PM today. Ultimately, it can test data across all three trimodal methods using the eNTERFACE dataset [234].

There are several factors to consider when collecting and organizing tabled trimodal datasets. The scope of the trimodal data will depend on the specific goals and requirements of the AC system. Also, it is essential to carefully consider the content of the data when collecting and organizing the datasets (e.g., the content of the data, Data Quality, Data storage, or Data labeling), as this will help to ensure that the data is relevant and appropriate for the intended use. An example of how data might be collected would be as follows:

- Text data: Text data can be anything from transcriptions of spoken language to social media posts to survey responses.
- Voice data: Such as recordings of speech or other sounds, voice data may include spoken words or sounds.
- Visual data: Images or video recordings of facial expressions or other visual information can be included as visual data.

Handling trimodal datasets for AC systems requires comprehensive knowledge of the data's breadth and substance, affirming its pertinence and appropriateness. These datasets, including text, voice, and visual information, require meticulous organization,

keeping data quality, storage, and labeling in mind. By following these guidelines, researchers and developers can improve the datasets’ practicality, consequently boosting the performance and precision of the corresponding AC systems.

It is essential to preprocess trimodal datasets as part of the trimodal fusion process. Preprocessing typically involves many tasks that clean, transform, or otherwise prepare the data for analysis. Some common data preprocessing tasks for trimodal datasets include (Removing noise, handling missing or corrupted data, normalizing or scaling the data, and transforming or extracting specific features or characteristics from the data). Preprocessing the trimodal datasets can help improve the accuracy and effectiveness of the ACsystem by ensuring that the data is consistent and usable. It is essential to consider the necessary preprocessing steps carefully, as this will depend on the data’s nature and the analysis’s requirements. Data processing improves quality and extracts information from data sources and has become a standard technology in healthcare, agriculture, computer science, and engineering [229].

Feature extraction from trimodal datasets is an essential step in the trimodal fusion process because of its importance, 0, and it is among the first steps of traditional ML algorithms [230]. Feature extraction involves identifying and extracting the data’s most relevant and informative characteristics, which can be used as input to an intelligent method model for emotion recognition. Different feature extraction techniques can be used, depending on the data’s nature and the analysis’s goals. It is common to use Z-score standardization techniques such as voice normalization and voice intensity thresholding before extracting acoustic characteristics from audio sources. OpenSMILE, an open-source software toolkit, is commonly used to extract and analyze acoustic features, is commonly used [235]. Some standard methods include filtering (to extract specific features or characteristics), transformation (converting data between different representations), and dimensionality reduction (to make it easier to analyze and process). It is essential to carefully consider which feature extraction techniques are most appropriate for the specific dataset and the goals of the analysis. By extracting the most relevant and informative features, you can improve the accuracy and effectiveness of the AC system.

In conclusion, affective computing, including emotion recognition or sentiment analysis, has made substantial progress over the past decades, particularly with advanced machine learning algorithms and multimodal data integration. Early research laid the foundation by establishing emotional databases and exploring the interplay between emotions and computational systems. Modern approaches, leveraging sophisticated techniques like GNNs and deep learning, have significantly enhanced the accuracy and robustness of emotion recognition systems. By integrating text, audio, and visual data, these systems can more effectively capture human emotions’ nuanced and multifaceted nature than

unimodal systems. However, challenges persist, particularly in data preprocessing, feature extraction, and handling complex, multimodal datasets. Addressing these challenges is crucial for advancing the field and developing more intuitive and empathetic human-computer interactions. Future research should continue to refine multimodal fusion strategies and explore innovative techniques to improve the precision of emotion recognition. This evolution in affective computing holds immense potential, inspiring a future of more responsive and emotionally aware technological systems. These advancements will enhance their utility across various applications such as healthcare, education, and human-computer interaction, fostering optimism and excitement for the future.



3. MULTIMODAL EMOTION RECOGNITION

Multimodal systems have shown greater efficiency in sentiment analysis and emotion recognition than unimodal approaches. To develop an intelligent system, it is essential to review recent successful methods. A key preprocessing step for multimodal fusion involves affect recognition, which lays the groundwork for integrating complex trimodal data. This chapter outlines the proposed approach's processing steps in detail. These steps include acquiring, collecting, and cleaning data and comprehensively explaining graph visualization techniques. Given its importance in corpus-based studies, we use a comparative analysis approach as our methodological framework. Various trimodal emotion datasets enable the analysis of multimodal data, capturing a wide range of human expressions, both acted and natural.

ER and SA aims to understand emotions within diverse and intricate environments. Establishing a robust multimodal framework becomes more intuitive with a foundational understanding of unimodal affect recognition [27]. This study integrates emotional information conveyed through text, audio, and visual cues in videos. Combining these features significantly enhances performance, surpassing the effectiveness of any single modality [236]. Textual information is extracted from the spoken content in videos, while speech models are evaluated based on their emotional conveyance, influenced by the frequency spectrum of each interval. Additionally, facial expressions are analyzed meticulously on a frame-by-frame basis, primarily focusing on this modality.

Figure 3.1. illustrates the integration of trimodal processes in affective computing. The diagram is divided into four primary areas. The first area focuses on data acquisition and preprocessing. It emphasizes data-gathering methods, critical analysis, authenticity, preprocessing steps, and data reliability. The second area highlights current trends and methodologies. The third area delves into modality-specific insights. The fourth area discusses the application and importance of these processes. The second area explores the latest trends and techniques, focusing on ML and ML in affect recognition and various multimodal fusion methods. The third section provides insights from specific modalities such as text from speech, audio speech models, and visual facial expressions, highlighting their contributions to emotion recognition. Finally, the fourth area discusses the application and significance of this technology, focusing on using social media as a data source, interfaces for analysis, the importance of ER, and the utilization of human affective datasets. This comprehensive structure demonstrates how different aspects of affective computing are integrated for efficient and effective emotion recognition and analysis.

3.1. Overview of Affective Computing

Analyzing emotions in content that includes text, sound, and images is a complex task that involves understanding emotional cues from words, tone of voice, and facial expressions. These cues, which help identify and measure primary emotions, require a keen

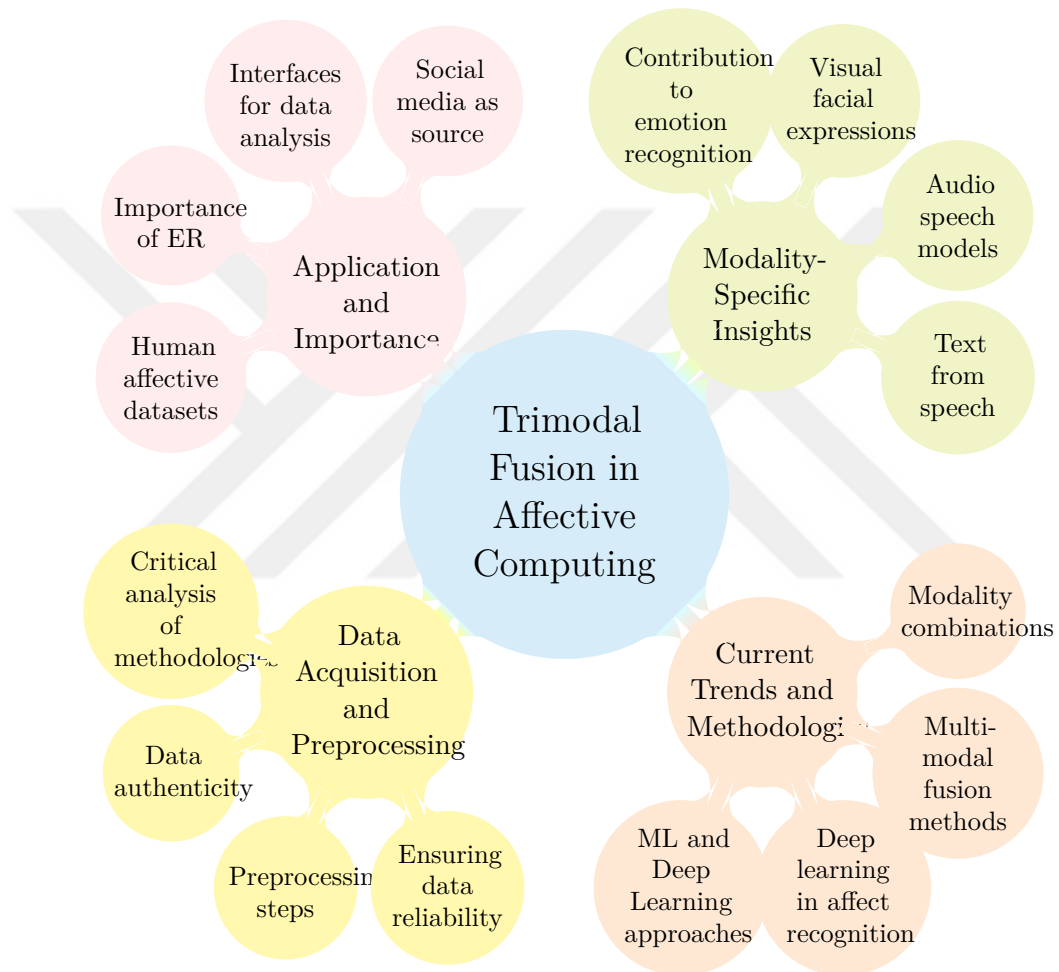


Figure 3.1. Trimodal Fusion in Affective Computing.

eye and ear. Paying attention to changes in voice, facial expressions, and the context of the text is critical for accurately identifying a person's emotional state. For instance, if a woman says "yes" to a marriage proposal with feeling, current methods that examine text, audio, and visual modes might not accurately predict the outcome. Combining these methods for a comprehensive emotional analysis is a more practical approach, but it's not without its challenges.

Humans express emotions in various ways, and understanding these expressions is at the heart of affective computing. The concept of 'Emotive Source,' or 'Speaker Embedding,' is a key element in this field. It can be categorized into three groups: multi-behavior techniques, which include facial expressions and body language; multi-physiological methods, such as heart rate and skin conductance; and behavior-physiological variants, which combine both behavior and physiological data. 'Emotive Source' refers to the individual from whom emotional data such as facial expressions, vocal patterns, or textual responses are collected for emotion recognition studies. This concept is not just important; it's crucial in multimodal emotion recognition, combining multiple data types to assess emotional states accurately. For instance, a person's facial expression and heart rate can be used to determine their emotional state during a stressful situation.

The term 'Emotive Source' highlights the individual as the primary source of emotional signals analyzed to understand affective states. It includes data from visual expressions, audio cues, and textual interactions, contributing to a comprehensive understanding of emotions. 'Emotive Source' specifies the origin of emotional data, similar to how 'data source' or 'input source' are used in technical contexts. This designation not only clarifies discussions about data flow in emotion recognition systems but also aids in structuring research methodologies and reports, particularly in studies involving complex data integration from multiple sensory inputs. Understanding the role of 'Emotive Source' in data flow is key to advancing our understanding of affective computing.

3.1.1. Trimodal Emotion Recognition

In text data, feature extraction involves identifying and extracting specific words or phrases or more abstract features such as sentiment or topic. Various techniques can be used for this purpose, depending on the nature of the data and the analysis goals. Standard methods include Bag of Words, n-grams, Term Frequency-Inverse Document Frequency (TF-IDF), and Word embeddings [237]. Extracting the most relevant and informative features from text data can enhance the accuracy and effectiveness of an AC system that uses text data as input.

Various AI methods are utilized for processing natural language, particularly in text analysis and sentiment analysis NLP. When dealing with spoken language, words are transcribed exactly as uttered by the speakers. To derive meaning from these words, a tool called Word2Vec is employed. Word2Vec converts each word into a numerical format, enabling further processing and analysis by intelligent machines [235]. Future directions in

NLP involve advancing the use of intelligent machines in text analysis and exploring new possibilities in sentiment analysis. Text analysis is used to develop annotated datasets that enable automatic emotion recognition. An annotated text entry is first transformed into a feature vector, which is then used to train ML algorithms and validated against annotated reference text. Initially, the text was classified into subjective and objective categories using supervised sentiment analysis [238].

In emotion recognition, a multimodal approach enables the simultaneous extraction and conversion of acoustic signals and images into text. Automatic speech recognition efficiently converts spoken language into readable text and is essential in applications such as speech-to-text systems where audio is the primary input. This technology is primarily based on DNNs. While DNNs are powerful, they require a lot of labeled training data to perform optimally. Detecting human emotions from text is a complex and challenging task [239]. With voice data, feature extraction involves identifying and extracting specific acoustic or prosodic features of the speech signal, such as pitch, loudness, or speaking rate. Various techniques can be used for this purpose, depending on the nature of the data and the analysis goals. Examples include Mel Frequency Cepstral Coefficients, Prosodic features, Acoustic features, and Word embeddings. Extracting the most relevant and informative features from voice data can enhance the accuracy and effectiveness of an AC system that uses voice data as input. Audio mining involves searching for and analyzing spoken words or phrases in audio files. Large amounts of recorded speech are usually stored in databases that are then searched for specific keywords, obtaining valuable information from many recordings. The frequency of occurrence of keywords in spoken words determines keywords of interest.

"Speech data" refers to recordings of human languages spoken over various media, such as telephone, webcasts, conversations, music, meetings, voice messages, lectures, television, and radio. Even though individual audio concepts provide strong predictors, a multimodal fusion produces the best results, indicating that each contains complementary information [240].

DNNs are crucial for audio emotion recognition due to their superior accuracy and robust feature extraction capabilities. For instance, Kumar et al. (2022) achieved an impressive 95.99% accuracy using DNNs on the IIT Roorkee Multimodal dataset, illustrating their efficacy in accurately identifying emotions from audio data. Ahmed et al. (2023) demonstrated the power of DNNs combined with 1D-CNNs, and LSTM, models, achieving 90.47% accuracy on the : Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) dataset, showcasing the ability of DNNs to capture intricate audio features and temporal dynamics. Xu et al. (2021) showed the flexibility of DNNs by combining them with RNNs and pooling techniques, achieving moderate accuracy (58.80%) on the IEMOCAP dataset. Gao et al. (2023) used Adversarial Autoencoders on IEMOCAP, achieving 58.38% accuracy, highlighting the adaptability of DNNs for advanced methods like adversarial training. Mustaqeem et al. (2023) demonstrated the

versatility of DNNs with Fully Connected CNNs on the : Berlin Database of Emotional Speech (Emo-DB) dataset, achieving 84.30% accuracy. Comparatively, Kaur and Singh (2023) showed that while DNNs using SVM achieved 43.80% on : Linguistic Data Consortium (LDC) and 79.10% on Emo-DB, they provide competitive results against traditional methods. Bhosale (2023) indicated that while traditional methods like SVM achieved 68.00% accuracy on Emo-DB, DNNs often deliver better or comparable results, with continuous performance improvements. Hence, DNNs high accuracy, robust feature extraction, versatility across datasets, and capability to integrate advanced techniques underscore their importance in audio emotion recognition see Table 3.1..

Table 3.1. Comprehensive Review of Emotion Recognition Using DNNs

Authors	Dataset	Year	Language	Fusion Type	Method	EL	Accuracy (%)
P. Kumar et al. [86]	IIT Roorkee Multimodal	2022	Eng.	HY	DNNs	6C	95.99%
Ahmed et al. [39]	CREMA-D	2023	Eng.	FL	1D-CNN-LSTM-GRU	6C	90.47%
Xu et al. [146]	IEMOCAP	2021	Eng.	Pooling	DNN/RNN	6C	58.80%
Gao et al. [150]	IEMOCAP	2023	Eng.	DL	Adversarial AE	6C	58.38%
Mustaqeem et al.	Emo-DB	2023	Eng.	DL	FC CNN	6C	84.30%
Kaur and Singh [148]	LDC, EmoDB	2023	Eng.	ML	SVM	6C	43.80/79.10%
Bhosale [149]	Emo-DB	2023	Eng.	-	SVM	6C	68.00%

The Internet is the largest multimodal data repository, including images and other data types. In addition to text, images are also important types of multimodal data. If these images could be transformed into high-level constructs that could be analyzed, their analysis would add significant value to the original data. Image mining involves extracting implicit knowledge, patterns, or information not explicitly stored in images. Methods used include image processing, pattern recognition, image retrieval, machine learning, image mining, and data mining [96].

An essential part of trimodal fusion is analyzing visual data. Visual data provide information on human emotions, including facial expressions, body language, and other nonverbal cues. By inputting visual data into affective computing systems, their accuracy and effectiveness can be improved by extracting the most relevant and informative features. Depending on the data and the analysis goals, several techniques can be used for feature extraction from visual data. Among them are scale-invariant feature transform, speeded-up robust features, Harris corner detector, local binary patterns, and facial feature detection. An AC system that uses visual data as input can be improved by extracting the most relevant and informative features [241].

The table 3.2. summarizes various CNNs models used by different authors for emotion

recognition and sentiment analysis. Each row lists the authors and the specific CNNs model they employed in their research. CNNs are crucial for image emotion recognition and sentiment analysis due to their ability to extract relevant features from images automatically. They are also effective in processing video data by segmenting the video into individual frames, extracting features from them, and then using these embeddings for classification tasks. For instance, models like CA-WGNN can utilize these embeddings for improved emotion classification, as shown in the table.

Table 3.2. CNN Models Used for Emotion Recognition and Sentiment Analysis

Authors	Dataset	Year	Model	Performance
Poria et al. [69]	IEMOCAP	2016	CNN + RNN + openSMILE	Superior performance
Hossain M. [242]	eINTERFACE, BAUM-2	2019	2D CNN (Audio), 3D CNN (Visual)	87.02, 56.12
Mustaqeem et al. [243]	Various	2019	CNN-LSTM	84.30
Ghaleb [244]	CRE-MA-D, CREMF-D	2019	MERMC	91.5, 66.5
Wu et al. [66]	Various	2021	3D-CNN + Attention Fusion	86.2
Pan et al. [1]	Various	2021	CNN + SVM	85.97 (mean precision)
Jothimani et al. [245]	Various	2023	2D/3D CNNs + ELM + SVM	Improved emotion estimation
Mittal et al. [75]	Various	2021	CNN + Early-Fusion	Enhanced prediction accuracy
Schoneveld L. [246]	RECOLA	2021	Visual Audio	CCC for arousal and valence
Pan [1]	BAUM-1s	2023	Audio-Visual CNN, Linear-SVM	85.97 (Precision)
Yang [177]	Various	2023	GCN-CNN	42.70
Tang [247]	Various	2023	STRCN	84.10
Mohana et al. [144]	Various	2023	Hybrid CNN-LSTM	High accuracy on hybrid datasets
Huang [248]	Various	2023	FLM-CNN	83.32
Razak and Sahran [249]	Various	2023	ICE-GAN	86.80

3.2. Artificial Intelligence

Artificial Intelligence AI has experienced significant advancements in recent decades. The 1980s marked a shift towards Machine Learning ML, where computers began to learn from data without explicit programming. This momentum continued into the 1990s with the development of algorithms such as Support Vector Machines (SVMs). The early 2000s saw the foundational work for Deep Learning DL, with Geoffrey Hinton and colleagues formally introducing the term in 2006 [250].

The 2010s brought significant breakthroughs in DL, particularly in computer vision

and natural language processing. Concurrently, Graph Neural Networks GNNs emerged in 2005 as a robust framework for processing graph-structured data [251]. GNNs have gained traction due to their efficacy in complex network analysis and their integration with ML and DL techniques.

This historical progression underscores the interlinked and evolving nature of AI, ML, DL, and GNNs, highlighting their collective impact on various fields, from healthcare to finance. These advancements illustrate the transformative potential of AI technologies in solving complex real-world problems. Learning-based approaches in feature transformation have revolutionized practical applications with their accuracy and effectiveness. A feature transformation function can be learned rather than manually designed, which has proven to be highly effective and produces accurate results. Nevertheless, there are many ways to design a trainable feature transformation function, and not all will achieve linear separation of classes within a feature space. Consequently, the design of trainable feature transformation functions needs to be more generalized.

3.2.1. Neural Networks

At the heart of deep learning resides the artificial neuron, drawing inspiration from the neurons in the brain. This remarkable concept, foundational to the field of neural networks, simulates the behavior of its biological counterpart, known for its ability to process and transmit information within the human nervous system. An artificial neuron, often simply called a "neuron," is a mathematical function modeled after the biological neurons observed in nature. It receives input, processes it according to its rules, and then passes on an output, much like how a biological neuron receives signals through its dendrites, processes them in the cell body, and sends them through its axon.

The naming of the artificial neuron as such is deeply rooted in its mimicking of the biological neuron's structure and function. Its design, highlighting a simplified yet effective abstraction, is influenced by the structure of a biological neuron, illustrated in Figure 3.2. the Neural Network. This parallel in design is symbolic and functional, embodying the blend of simplicity and complexity that enables biological and artificial networks to perform tasks ranging from simple calculations to complex decision-making processes. Neurons create links with each other as dendrites connect to axon terminals, forming synapses that allow signal transmission between neurons based on the inputs received from connected neurons. This process illustrates the connection between two neurons [252]. Current graph construction techniques frequently start with an unsupervised learning approach, which does not require labeled data initially. As the process advances, labeled samples become crucial. They serve as prior knowledge, helping to refine the graph structure for more targeted learning tasks. Specifically, research by Dhillon et al. [253] focuses on this phase. It investigates how labeled samples can enhance node similarity assessments within the graph. This enhancement is vital for improving the accuracy of the learning model in subsequent tasks.

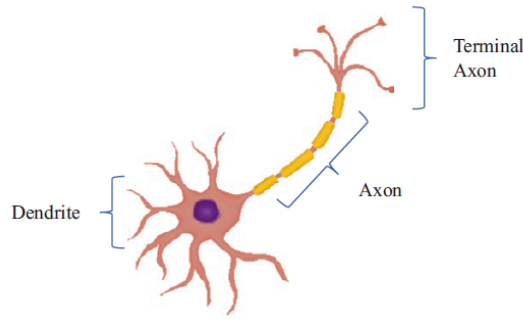


Figure 3.2. Neural Network

Rohban et al. [254] propose a supervised method, suggesting optimal neighborhood graphs are subsets of NN graphs, providing adequate manifold sampling. Following groundwork by Ozaki et al. [255], Berton et al. [256] developed the graph based on the Informativeness of Labeled Instances (GBILI), which utilizes label information for better classification accuracy, albeit with high time complexity. Lilian et al. [257] improved this by introducing the Robust Graph that Considers Labeled Instances (RGCLI) method, enhancing graph robustness through optimization. Recently, Zhuang et al. [258] introduced a semi-supervised learning approach incorporating labeled data into a Low-rank Representation (LRR), significantly benefiting label inference by adding supervised information.

3.2.2. Graph Neural Networks

GNNs are a specialized variant of DNNs engineered to process data with graph-like structures [259]. They represent a significant advancement in machine learning and artificial intelligence, driven by the pervasive use of interconnected data across various domains such as social networks, biology, recommendation systems, and cybersecurity. GNNs excel at modeling and understanding complex relationships, surpassing the capabilities of traditional machine learning models in addressing real-world challenges. By analyzing detailed structural data of graphs, GNNs effectively capture crucial dependencies, connections, and contextual relationships, facilitating informed decision-making. This diagram visually encapsulates the hierarchical relationship among the key concepts within the field of artificial intelligence AI, illustrating how each subsequent category is a specialized subset of the preceding. At the broadest level, AI represents the overarching domain concerned with creating systems capable of performing tasks that typically require human intelligence. Nested within AI, ML emerges as a subset, focusing on algorithms and statistical models that enable computers to perform specific tasks without explicit instructions, instead relying on patterns and inference.

The efficacy of machine learning techniques is profoundly influenced by both the algorithms' design and the data representation quality. Suboptimal data representations,

lacking critical information or containing erroneous or redundant data, can detrimentally affect algorithm performance across various applications. Representation learning aims to distill essential information from data efficiently without unnecessary duplication. This process, commonly called feature engineering, traditionally depends on human expertise and prior knowledge of the data and its specific applications.

Deploying machine learning and other artificial intelligence algorithms often requires substantial human effort in designing preprocessing pipelines and data transformations. Feature engineering utilizes human creativity and knowledge to select and structure discriminative information that enhances the effectiveness of machine learning tasks. For example, political scientists may create lists of keywords that enable text classifiers on social media to detect discussions about societal events. Similarly, in speech transcription, essential features might be derived from raw sound waves using methods like Fourier transformations.

Despite its extensive adoption, feature engineering presents drawbacks, such as needing significant input from domain experts and the risk of incomplete or biased feature extraction. The domain experts' knowledge frequently constrains the value and discriminative power of the features. In areas with limited human expertise, like early cancer detection, deciding which features to extract poses a significant challenge. As a result, a principal aim within the machine learning and artificial intelligence communities has been to minimize dependency on feature engineering, thereby accelerating the development and enhancing the efficacy of new applications.

Additionally, GNNs cover a broad spectrum, necessitating detailed study to clarify all relevant aspects. In our thesis, we specifically explore the application of GNNs in emotion classification, utilizing a novel algorithm integrated within a contemporary framework tailored for this task.

Further specialized within ML is NN, a paradigm inspired by the biological neural networks constituting animal brains, which underscores systems capable of learning and making intelligent decisions. The most specialized subset depicted is DL, a class of neural networks characterized by deep architectures and the ability to learn from a vast amount of unstructured data. Within DL, specific architectures like CNNs, RNNs, and GNNs are highlighted, each tailored for different types of data and tasks CNNs for visual data, RNNs for sequential data, and GNNs for data best represented as graphs. This structure illustrates the containment and specialization within AI and the progression from broad concepts to specific technologies and methodologies [260]. GNNs have become a potent tool for analyzing data with intricate graph structures, such as brain networks, knowledge graphs, consumer purchasing patterns, and biological pathways. Despite their success in handling various graph-related tasks, GNNs encounter notable obstacles. Their complexity often makes them difficult to interpret, and they struggle with effectively incorporating information of different levels of importance in a scientifically sound manner.

Graph theory elucidates the intricate relationships between objects through nodes and

their interconnections, known as edges. This field has garnered significant attention due to its capacity to model complex systems prevalent across various disciplines, particularly with the advent of machine learning technologies. In the field of social sciences, it facilitates the analysis of social networks, whereas in the natural sciences, it is instrumental in studying behavior systems, proteins, and knowledge graphs. Unlike traditional, linear data structures, graphs embody non-Euclidean characteristics, presenting both challenges and opportunities in the context of machine learning [261]. Graph Neural Networks GNNs have emerged as a potent tool for analyzing graph data, demonstrating exceptional proficiency in node identification, link prediction, and clustering. The efficacy of GNNs in navigating the intricate domain of graph analysis has spurred considerable interest in their potential. Figure 3.3. delineates the fundamental components of a GNNs, emphasizing the nodes and their linkages. Graphs are versatile

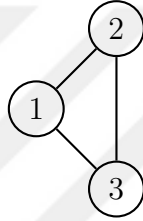


Figure 3.3. A simplistic representation of a GNN, showcasing nodes.

structures that model systems such as communication networks, supply chains, and organizational hierarchies. They consist of nodes (vertices) and edges that connect these nodes. Edges can be directed (one-way) or undirected (mutual). Unlike linear (Euclidean) spaces, graphs are non-Euclidean, making distance measurements between nodes non-linear. This non-linearity complicates using standard neural networks designed for Euclidean spaces in analyzing graph data. Here's a simple visual illustrating the difference between directed and undirected edges. As Figure 3.4. illustrates, directed edges (with arrows $A \rightarrow B$) showcase a one-way link from one node to another, while undirected edges (without arrows $C-D$) represent a two-way, reciprocal link between nodes. This distinction is crucial for understanding graph dynamics and applying neural network models to graph data [261]. As mentioned earlier, CNNs cannot handle non-Euclidean tasks, whereas

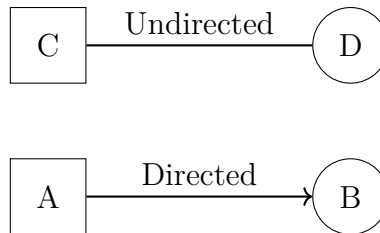


Figure 3.4. Comparison of directed and undirected edges.

GCNs can manage both Euclidean and non-Euclidean tasks.

But what does this mean?

To understand, we first need to grasp Euclidean space. This familiar geometry, established by the ancient Greek mathematician Euclid, describes a flat space where familiar rules apply: parallel lines never meet, and the sum of angles in a triangle is always 180 degrees. This intuitive geometry underpins classical physics and our everyday understanding of space. Non-Euclidean geometry, on the other hand, does not follow Euclid’s postulates. There are two main types: hyperbolic and spherical. In these geometries, rules change—parallel lines may diverge or converge, and the sum of angles in a triangle can be less or more than 180 degrees. Essentially, non-Euclidean spaces are curved.

When we talk about graphs in the context of graph neural networks GNNs, "non-Euclidean" becomes relevant because graphs don’t fit the conventional Euclidean framework. Graphs consist of nodes (vertices) and edges (connections) representing complex relationships, from social networks to molecular structures. Graph data is irregular, unlike images, which fit into a regular, grid-like structure suited for CNNs. Each node can have varying connections, and distances between nodes are not uniform or meaningful in an Euclidean sense. This irregularity and complexity make graph data inherently non-Euclidean.

Therefore, this fundamental difference makes applying CNNs to graph data challenging. Graphs need different approaches to define locality, connectivity, and operations like convolution. This is where GNNs excel, as they are designed for the non-Euclidean nature of graph data. They model relationships and patterns within graphs using techniques adapted from, but not identical to, those used in CNNs for Euclidean data.

3.2.3. Evolution of Research in General Graph Neural Networks

The introduction of Graph Neural Networks GNNs by Gori in 2005 [262] and further elaborated by Franco Scarselli et al. in 2009 [251] marks a significant advancement in neural network research, particularly for processing graph-structured data. Unlike traditional methods that require transforming graphs into vector formats, GNNs preserve the inherent topological features of graphs, enhancing their analysis across various fields such as chemistry, software engineering, and image processing. These networks support diverse graph types, including directed, undirected, labeled, and cyclic graphs, and utilize a unique information diffusion mechanism to update node states, eliminating the need for pre-processing. Empirical studies affirm GNNs’ effectiveness in both node-centric and graph-centric tasks, showcasing their robustness, generalization capability, and pivotal role in extending neural network applications to graph-based data, thus establishing a new paradigm in machine learning and neural network research communities. GNNs are designed to capture the attributes and connections of nodes within a graph, learning a state embedding $h_v \in \mathbb{R}^s$ representing the neighborhood information. Various outputs, including the distribution of expected node labels, can be predicted using this embedding.

The state embedding h_v can be used to predict node label distribution, as Zhou J. et al. demonstrated in [263]. GCNs was introduced in 2017 by Thomas Kipf and Max Welling [264], which uses a first-order approximation of a localized spectral filter on graphs to adapt convolutional neural networks CNNs to work with graph-structured data. To handle the complexity of graph data, GCNs extend CNNs. GNNs are a set of methods to apply deep neural networks to graph-structured data. Since the graph structure is not a regular grid, classic deep neural networks cannot be easily generalized to graph-structured data. To facilitate node-focused tasks, GNNs aim to learn features for each node. For graph-focused tasks, the learning objective typically involves learning features representing the entire graph, where learning node features is usually an intermediate step. The process of learning node features typically utilizes both the input node features and the graph structure. More specifically, this process can be summarized as follows:

$$F^{of} = h(A, F^{if}) \quad (3.1)$$

Here, $A \in \mathbb{R}^{N \times N}$ represents the adjacency matrix of the graph with N nodes, and $F^{(if)} \in \mathbb{R}^{N \times d_{if}}$ and $F^{(of)} \in \mathbb{R}^{N \times d_{of}}$ are the input and output feature matrices, respectively, where d_{if} and d_{of} are their respective dimensions. The superscripts "if" and "of" indicate the input and output of the filtering, respectively. Accordingly, the operator $h(\cdot, \cdot)$ is referred to as a graph filter. Figure 3.5. illustrates a typical graph filtering process, where it's noted that

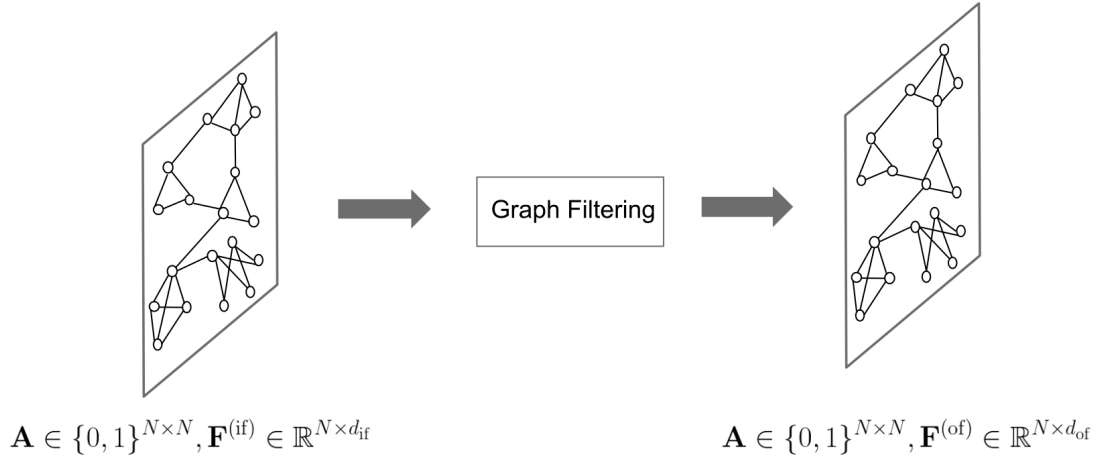


Figure 3.5. The process of graph filtering.

the filtering operation does not alter the graph structure, only enhancing the node features. The graph filtering process is sufficient for node-focused tasks, and multiple graph filtering operations are often stacked sequentially to create the final node features. However, for graph-focused tasks, additional processes are required to construct the features of the entire graph from the node features. Like classical GNNs, pooling operations summarize node features to create graph-level features. Classical GNNs are applied to data on regular grids,

but the graph structure is irregular, necessitating specialized pooling operations in GNNs. Intuitively, pooling operations on graphs should use graph structure information to guide the pooling process. In fact, pooling operations often take a graph as input and produce a coarser graph with fewer nodes. Therefore, the key to pooling operations is maintaining the graph structure (or adjacency matrix) and creating node features for the coarsened graph. Generally, as shown in Figure

$$A^{(op)}, F^{(op)} = \text{pool}(A^{(ip)}, F^{(ip)}), \quad (3.2)$$

Where $A^{(ip)} \in \mathbb{R}^{N_{ip} \times N_{ip}}$, $F^{(ip)} \in \mathbb{R}^{N_{ip} \times d_{ip}}$ and $A^{(op)} \in \mathbb{R}^{N_{op} \times N_{op}}$, $F^{(op)} \in \mathbb{R}^{N_{op} \times d_{op}}$ are the adjacency and feature matrices before and after the pooling process, respectively. Similarly, the superscripts "ip" and "op" are used to denote the input and output of the pooling, respectively. Note that N_{op} represents the number of nodes in the coarsened graph and $N_{op} < N_{ip}$.

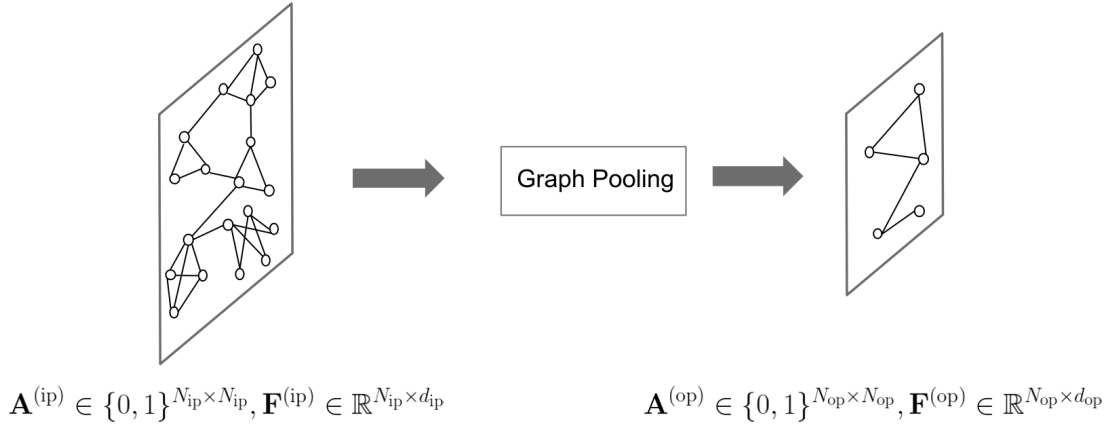


Figure 3.6. The process of graph pooling.

A typical GNNs model architecture consists of graph filtering and/or graph pooling operations. For node-focused tasks, GNNs use only graph filtering operations. They typically consist of multiple consecutive graph filtering layers, where the output of one layer serves as the input for the next layer. For graph-focused tasks, GNNs require both graph filtering and graph pooling operations. Pooling layers often divide the graph filtering layers into blocks. This section briefly introduces the general architectures for GNNs, then details representative graph filtering and graph pooling operations.

We provide a comprehensive overview of the GNNs adapted for both node-level and graph-level tasks. Firstly, we introduce some notations used in the subsequent sections. We denote a graph as $G = \{V, E\}$. The adjacency matrix of a graph with N nodes is represented as A . Associated features are represented as $F \in \mathbb{R}^{N \times d}$, where each row of F

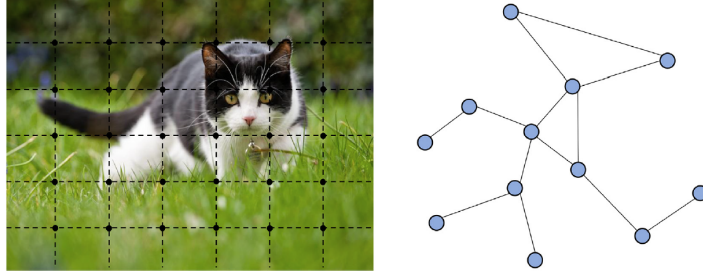


Figure 3.7. A general GNN architecture for node-focused tasks [263].

corresponds to a node, and d is the dimension of the features.

3.3. Approach of General Graph Neural Networks Frameworks

GNNs are inspired by CNNs, as highlighted by LeCun et al. [265]. CNNs excel at extracting and composing multi-scale localized spatial features, significantly powering advancements across various machine learning fields and driving the deep learning revolution. Delving deeper into the core aspects of CNNs, such as local connections, shared weights, and multi-layer structures, we see their relevance in addressing graph-related problems. Specifically, graphs inherently exhibit local connectivity; shared weights minimize computational expenses compared to traditional spectral graph theory [266]. Multi-layer architectures are crucial for capturing hierarchical patterns and features of different scales. However, CNNs are traditionally limited to processing regular Euclidean data, such as images (2D grids) and text (1D sequences), which can also be conceptualized as graph instances. This limitation challenges adapting CNN methodologies to graph structures, particularly in defining localized convolutional filters and pooling operators for the transition from Euclidean to non-Euclidean domains, as illustrated in Figure 3.7.. This difficulty impedes the straightforward application of CNN techniques to graph data [263]. This thesis will investigate GCNs, which generalize convolution operations to graph-structured data. Building upon the significant achievements of CNNs in deep learning, GCNs adapt these principles to the graph domain. Notably, advancements in this field offer novel applications, particularly in the domain of AC with a focus on multimodal emotion recognition. Our model leverages advanced techniques for emotion recognition. In our study, we utilized two distinct datasets to conduct our experiments: The CMU-MOSI dataset represents a seminal advancement in multimodal sentiment analysis by integrating visual, auditory, and textual data to facilitate comprehensive emotion and sentiment recognition. It is a critical resource for extracting nuanced emotional cues, leveraging video and audio features to provide a multidimensional analysis framework. The dataset's extensive and diverse emotion labeling is instrumental in developing sophisticated models for precise emotion detection. The distinctive feature of CMU-MOSI

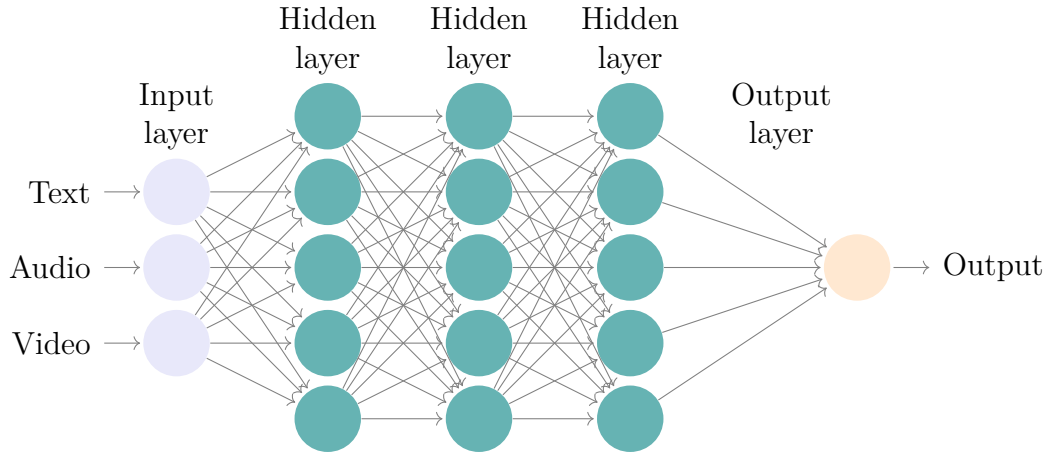


Figure 3.8. Multimodal Emotion Recognition using Trimodal

is its ability to offer a detailed portrayal of human sentiment, rendering it essential for applications to decipher complex emotional expressions in real-world interpersonal communication. Integrating multiple data modalities establishes a solid basis for exploring and modeling intricate sentiment dynamics.

3.3.1. Emotive Source

Humans express emotions in many ways. "Emotive Source," or "Speaker Embedding," can be divided into three groups: multi-behavior techniques, multi-physiological methods, and behavior-physiological variants. The "Emotive Source" refers to the individual from whom emotional data, like facial expressions, vocal patterns, or textual responses, are gathered for emotion recognition studies. This concept is essential in multimodal emotion recognition, which combines various data types to evaluate emotional states accurately. Our methodology introduces "Emotive Source" as a technique for creating distinct profiles from voice recordings through LSTM networks, focusing on text analysis. This is complemented by DNNs for audio feature extraction and CNNs for video analysis, integral to our innovative approach involving the Cumulative Attribute-Weighted Graph Neural Network CA-WGNN for advanced emotion recognition. Feature extraction spans text, audio, and video modalities. Textual analysis via LSTM networks uncovers linguistic markers of emotion, offering insights into the emotional depth of speech. DNNs are employed for audio, targeting speech's unique intonations and rhythms, which carry key emotional signals. For video, CNNs analyze non-verbal cues like facial expressions and body language, crucial for interpreting emotions. These combined techniques underscore our comprehensive strategy for multimodal sentiment analysis, highlighting the importance of Emotive Source in understanding emotion from various data types. The construction of an LSTM network commences with discerning non-essential textual data elements using a decision-making process governed by the

sigmoid function. This process evaluates the utility of data based on the output from the previous LSTM unit, g_{s-1} , and the current input W_s , effectively filtering information through the forget gate (u_i). The forget gate applies the sigmoid function to produce a vector, where each element's value, ranging from 0 to 1, dictates the extent to which elements in the cell state h_{i-1} should be retained or discarded as shown in Equation 4. 1: The function σ represents the sigmoid activation, and the components Y_u , V_i , and b_u are critical for the forget gate's functionality. The network then evaluates the necessity of updating the cell state with new input V_i , utilizing both sigmoid and tanh functions for this purpose. This dual-layer approach facilitates the determination of information relevance and its integration into the cell state, as delineated by Equations 4. 2 to 4. 4:

Here, C_{i-1} and C_i signify the previous and current cell states. The output generation mechanism involves conditioning the cell state for output through a sigmoid layer (O_i), which selects critical output elements. This output, h_i , is then modulated by the cell state C_i after passing through a tanh function, ensuring output values fall within the -1 to 1.

3.3.2. Exploring Deep Learning Architectures

Deep learning, a subset of machine learning, capitalizes on modern computers' enhanced computational power and memory to perform complex mathematical operations on large datasets. This facilitates the construction of DNNs with multiple layers.

Key Advantages of Deep Learning:

- *Simplicity*: Deep learning streamlines model architecture by repetitively utilizing basic network layers, thus forming large and efficient networks.
- *Scalability*: These models excel with large datasets and outperform traditional methods such as kernel machines, which become computationally inefficient as data grows.
- *Domain Transfer*: Models trained on one task can generalize to perform other related tasks, proving beneficial in scenarios with limited data.

Deep learning sets new benchmarks in various applications, notably in image tasks such as detection, segmentation, classification, and recognition. CNNs demonstrate significant efficacy in image recognition, a key area in computer vision [267]. A number of AI technologies, including CNNs and LSTM networks, represent cornerstone technologies for a wide array of applications, from image recognition to sequence data processing. The effectiveness of these architectures in handling their respective tasks is underscored by extensive research, including the work by Yildirim et al.[268], which highlights the capabilities of neural networks in complex applications. CNNs are renowned for their proficiency in processing spatial data, leveraging convolutional layers to detect patterns and features in images, a principle extensively reviewed by Khorrami et al. [269]. This architecture's ability to hierarchically extract features makes it ideal for tasks such as

image and video recognition, contributing significantly to advancements in computer vision. On the other hand, LSTM, an evolution of RNNs, are engineered to address the challenges of processing sequential data, particularly the problem of capturing long-term dependencies. Traditional RNNs struggle with the gradient vanishing problem, which LSTM mitigate through their innovative 'memory cell' design. This feature allows LSTM to selectively retain and update information over long sequences, enhancing their performance in complex time series analysis, natural language processing, and more, as discussed in the works of Chao et al. [270].

Text Processing via LSTM: Long Short-Term Memory Networks LSTM, a specialized form of RNNs, are particularly suited for text-processing applications. Their design effectively captures long-term dependencies within sequential data, a critical aspect of natural language processing NLP. LSTM are instrumental in tasks that rely on the sequential nature of text, such as sentiment analysis, language modeling, and text generation. By preserving information over long sequences, LSTM enables more coherent and contextually relevant text analysis, enhancing the performance of chatbots, predictive text input, and automated content generation systems.

Audio Processing through DNNs: Due to their versatile architecture, DNNs offers various applications. DNNs have shown remarkable proficiency in speech recognition, audio classification, and feature extraction tasks in the audio domain. By learning complex representations of audio signals, DNNs can uncover intricate patterns in sound data, facilitating a deeper understanding of audio content. This makes them invaluable for developing advanced audio processing systems, including voice-activated assistants, music recommendation services, and sound anomaly detection in security systems.

Video Processing with CNNs: CNNs excel in handling spatial data, making them an ideal choice for video processing tasks. Their architecture, built upon convolutional layers, is adept at extracting hierarchical features from images and video frames. This capability is crucial for performing object recognition, detection, and segmentation within video content. CNNs' effectiveness in discerning patterns and features at various levels of abstraction enables sophisticated video analysis applications, ranging from automated surveillance to enhanced video content interaction.

3.3.3. Dimensionality Reduction and Graph Embeddings:

Dimensionality reduction techniques, such as embeddings, are crucial for simplifying complex, high-dimensional data into manageable, low-dimensional representations. These transformations enable effective visualization and advance machine learning tasks by retaining vital information. In fields like natural language processing and image recognition, embeddings improve algorithms' interpretability and operational efficiency

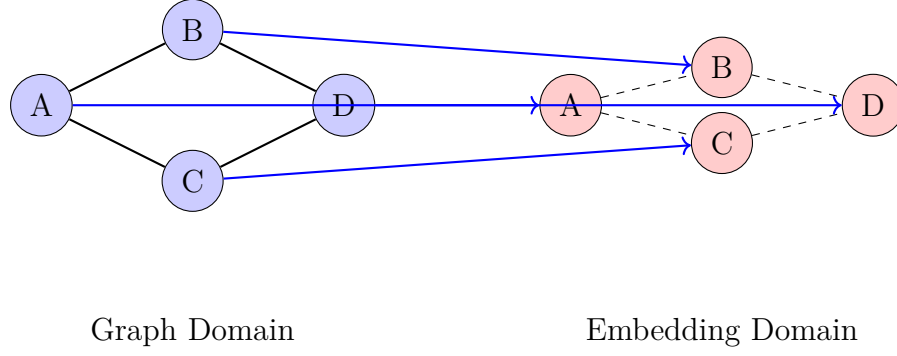


Figure 3.9. Graph Embedding.

[271]. GNNs adapt traditional neural network capabilities to process graph-structured data efficiently. GNNs embeddings aim to generate concise representations of nodes, edges, or whole subgraphs. These representations capture structural and attribute-based details, significantly enhancing tasks such as node classification, link prediction, and graph classification [272]. The motivation for this study stems from the advancements in graph embedding techniques, as demonstrated by the works of Cai et al. [240], Cui et al. [273], among others. These methods represent graph details such as nodes, edges, and subgraphs as compact vectors, offering a more flexible approach than traditional graph analysis methods that depend on manually crafted features, which can be rigid and expensive. Inspired by the breakthroughs in word embeddings by Mikolov et al. [274], newer methods such as DeepWalk, node2vec, LINE, and TADW have emerged. However, challenges persist, as highlighted by Hamilton et al. [275], including issues with parameter sharing among nodes that lead to computational inefficiencies and scaling problems, as well as difficulties in adapting to dynamic or newly introduced graphs. Graph embedding, also known as node embedding, efficiently translates each node in a graph into a compact, low-dimensional vector. This process maintains the essential characteristics of the original graph. In the graph domain, nodes, connected by edges, showcase their structural links. Each node is depicted as a vector in the embedding domain, encompassing its features and relational context. Graph embedding aims to map nodes from the graph to the embedding domain while preserving structural and relational integrity. This method is depicted visually in Figure 3.9..

3.4. Fusion Techniques

Trimodal fusion is a technique employed in AC to integrate and analyze multiple data sources to recognize and interpret human emotions. The term "trimodal" indicates using three distinct data types: text, voice, and visual. Text data encompasses written or typed language, including social media posts, emails, and spoken language transcripts. Voice data pertains to the acoustic properties of speech, such as pitch, volume, and

intonation. Visual data includes information captured through cameras or other imaging devices, covering facial expressions, gestures, and body language. By combining these data types, trimodal fusion aims to understand an individual's emotions and behavior comprehensively. This integration can be achieved through various techniques, including ML algorithms and natural language processing methods. Significant research is ongoing to effectively integrate these modalities in ER studies.

In the context of fusion, selecting relevant information from different modalities and filtering out noise is critical due to both their independence and correlation, presenting challenges in identifying useful information. Previous fusion approaches have focused on learning multimodal representations by concatenating features through various fusion mechanisms, such as late fusion, early fusion, and hybrid fusion. In most reviews of MER, multimodal fusion strategies are classified into early fusion (feature-level fusion), decision-level fusion, and hybrid fusion, which combines different modalities in multimodal affective analysis [96], [44], [6], [276]. To identify the most suitable fusion technique, it is frequently necessary to test various methods and evaluate their performance.

Integration of many frameworks of Emotive Source and the fusion approach are two fundamental components of trimodal emotions. Trimodal combinations can be categorized into three distinct groups: "multi-behavior techniques, multi-physiological methods, and behavior-physiological variants." Two fundamental fusion procedures are employed in multimodal emotional analysis, including "feature-level fusion and decision-level fusion." Combining feature vectors by factorized bilinear pooling is a widely employed technique for feature-level fusion. Decision-level fusion commonly employs majority or median voting. Linear weighted computing, also known as linear weighted fusion, is a technique that may be used in fusions.

3.4.1. Feature Layer Fusion

Feature-level fusion (or early aggregation) is the process of concatenating the features extracted from signals from different modalities and then feeding them into an emotion classifier [235]. It combines the features from other modalities at a specific Multi-modal information fusion application to the human level, for example, by adding or averaging their elements. [75]. Each input channel's characteristics are combined into one vector in the feature-level fusion before any classification operations are performed. It is possible to use this approach in a variety of ways, such as the possibility to use this approach in a variety of ways. Fusion of features that don't happen at the same time. Integrating highly disparate input features presents a challenge when combining modalities at the feature level. Interface multiple inputs simultaneously while researching a modality's classification system is not straightforward [277]. For example, if the aim is to combine features extracted from voice, text, and visual data using feature-level fusion, a new set of features incorporating all three elements can be obtained by taking the

element-wise sum of each modality. The features can also be combined using other operations, such as multiplication or concatenation [278].

Figure 3.10. illustrates the concept of feature-layer-fusion in a neural network model. The

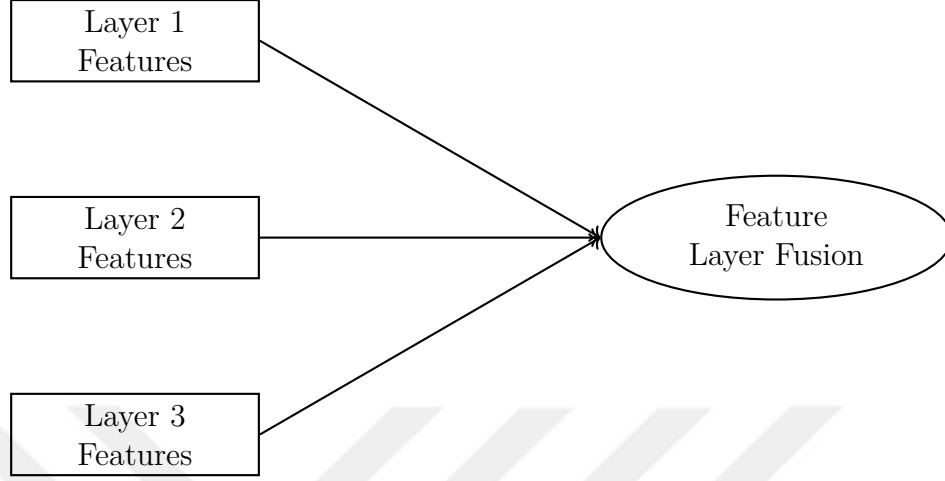


Figure 3.10. Feature Layer Fusion

diagram is composed of three distinct layers labeled "Layer 1 Features," "Layer 2 Features," and "Layer 3 Features." Each layer represents a different stage in the neural network, specifically designed for extracting certain features from the input data. The arrows in the diagram depict the flow of these extracted features. They move toward a central node, labeled "Feature Layer Fusion," indicating integrating features from multiple layers." This central node represents the fusion process, combining features from all layers to create a comprehensive representation. The fusion process is crucial for integrating multi-level feature information, enhancing the model's ability to learn and recognize complex patterns within the data. This approach leverages the strengths of individual layers, providing a more robust and holistic feature set for subsequent processing tasks in the neural network.

3.4.2. Decision-level Fusion

A decision-level fusion involves modeling and categorizing each modality separately [235]. At the end of the process, the unimodal results are combined by selecting appropriate metrics, such as statistical weighting, sums, and products, and the majority of simple operators are available. Several studies have determined that decision-level fusion is the best method for combining data because different classifier errors tend to be uncorrelated, and the methodology is feature-independent [35]. In this approach, the prediction process involves individual classifiers operating on specific modalities, and their outputs are subsequently merged. By combining the projections from diverse classifiers, the accuracy of the final prediction can be enhanced when additional modalities offer

complementary information. Nevertheless, if the modalities exhibit redundancy or if the classifiers require improved calibration, the effectiveness of this fusion at the decision level may be diminished [278]. Figure 3.11. illustrates a decision-level fusion (or late fusion)

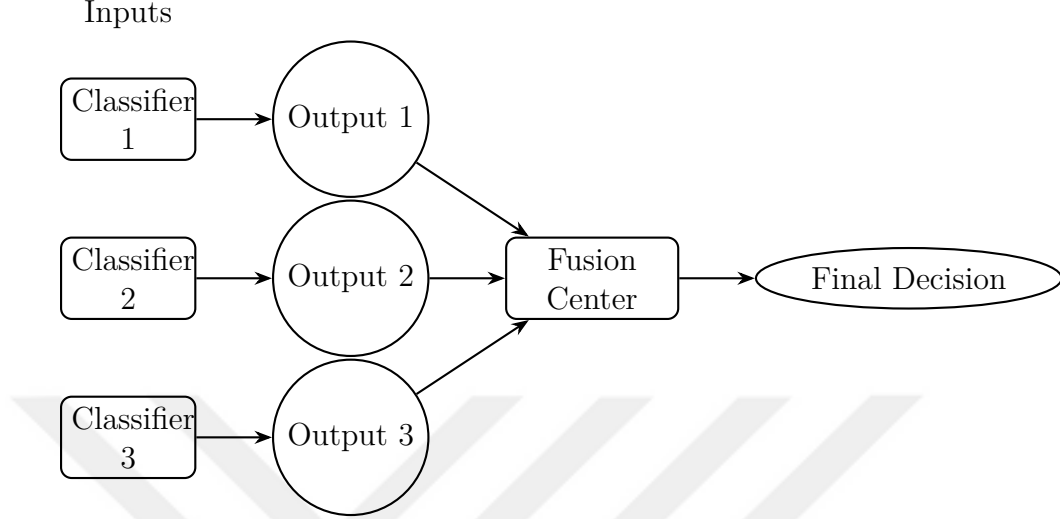


Figure 3.11. Decision-level fusion framework with three classifiers.

framework where the outputs of multiple classifiers are combined to make a final decision. This setup uses three classifiers, each processing the input data independently and generating its respective outputs (Output 1, Output 2, and Output 3). These outputs are then sent to a fusion center, aggregating the information from all classifiers to produce a unified final decision. This approach leverages the strengths of individual classifiers by combining their decisions, potentially enhancing the overall accuracy and robustness of the system.

3.4.3. Mixed Layer Fusion

Features and decision levels are combined in a hybrid multi-modal fusion strategy, combining the benefits of early and late fusion. Consequently, researchers are motivated to employ the hybrid fusion strategy to address numerous multi-modal analysis challenges. The Visual Spoken Textual Additive Net (VISTA Net), an ER system, was developed by Kumar and his team [86]. The KAAP method quantifies the influence of each modality and its respective attributes in predicting a certain emotion class. It amplifies performance by amalgamating information from three different domains: visual, spoken, and textual. This approach combines text, speech, and image features to recognize emotions, resulting in improved overall results. VISTA Net’s hybrid fusion technique makes it more effective and powerful.

A hybrid fusion of intermediate and decision-level information determines the

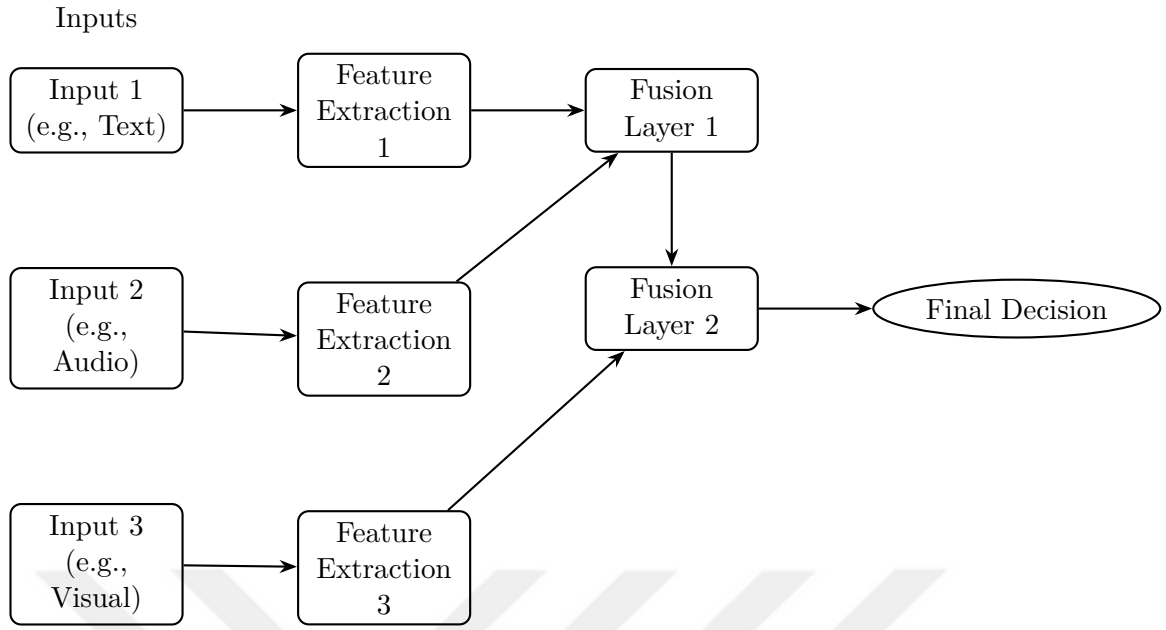


Figure 3.12. Mixed layer fusion framework integrating different modalities at multiple layers.

weights of these modalities. Instead of unimodal or bimodal, all three modalities improve system performance. Using a proposed new interpretability technique, K- Average Additive Explanation, it identifies each modality's contribution and important features toward predicting a particular emotion class. This dataset comprises generic images, corresponding speech utterances, text transcripts, and discrete class labels, such as happiness, sadness, hatred, and anger, which are all part of the IIT-R MMEmoRec dataset. The first step is to train SER, TER, and IER models for bimodal ER. Hybrid fusion strategies have the advantage of incorporating multiple sources of information, which can lead to more accurate and robust models. However, it is essential to carefully consider the characteristics of the data and the specific task at hand to determine the most appropriate fusion strategy. Trying multiple methods and evaluating their performance may be necessary to decide on the most effective approach.

Figure 3.12. depicts a mixed-layer fusion framework, which integrates different modalities at multiple stages within the processing pipeline. The diagram shows three types of inputs: text, audio, and visual data, which are processed independently through their respective feature extraction layers (Feature Extraction 1, Feature Extraction 2, and Feature Extraction 3). The extracted features from each modality are then combined at intermediate fusion layers (Fusion Layer 1 and Fusion Layer 2), allowing for the integration of multimodal information at various levels of abstraction. The fusion layers amalgamate the features and feed into the final decision-making layer, producing the final decision output. This approach leverages the complementary nature of different modalities, enhancing the system's ability to make accurate and robust decisions by incorporating diverse data sources throughout the processing stages.

3.5. CA-WGNN Algorithm

This section aims to demonstrate the core concepts behind CA-WGNN by explaining the terms cumulative, attributes, weights, and GNNs. Additionally, a sample code is provided to demonstrate the practical application of these concepts. The table below (Table 3.3.) provides detailed descriptions of these core concepts. Understanding these terms is crucial for comprehending how CA-WGNN operates and effectively implementing the model.

Table 3.3. Core Concepts in CA-WGNN

Term	Description
Cumulative	In the context of CA-WGNN, the term cumulative refers to the aggregation or summation of attribute weights over the nodes and edges within a graph. This cumulative process helps capture the collective significance of attributes, thereby enhancing the model's predictive capabilities.
Attributes	Attributes are the features or characteristics associated with the nodes (and sometimes edges) in the graph. In the provided code example, these attributes correspond to the input features of the dataset, with each feature having 20 dimensions, as specified by <code>n_features=20</code> .
Weights	Weights are numerical values that represent the importance or contribution of each attribute (or feature) to the model's predictions. These weights are learned during the training process and are utilized to adjust the influence of input features on the output.
GNNs	GNNs are a class of neural networks designed to operate on graph-structured data. They leverage nodes (entities) and edges (relationships) to capture intricate dependencies and interactions within the data. CA-WGNN is a specialized type of GNNs that employs cumulative attribute weights to improve its learning performance.

3.5.1. Implementation

The following Python code illustrates the implementation of a neural network model, incorporating the concepts discussed above. This model includes various layers such as input, reshape, convolutional, attention, flatten, dense, and output layers. Although this example does not explicitly implement a GNNs, it includes essential neural network components like attention layers, which are integral to advanced models like CA-WGNN.

Finally, the model is evaluated using the `evaluation_metrics` function, and the results are printed. This document has explained the core concepts behind CA-WGNN, providing a detailed understanding of cumulative attributes, weights, and GNNs. The Python code exemplifies these concepts in practice, showcasing a neural network model that handles synthetic classification data.

3.5.2. Dataset

Dataset CMU-MOSI The CMU-MOSI dataset stands at the forefront of Sentiment Recognition, highlighting the capabilities of multimodal sentiment analysis [67]. This pioneering dataset is a rich source of emotional insights, capturing a wide spectrum of emotions through spoken words by adeptly combining visual and auditory data. This approach offers a holistic perspective on human sentiment expression. CMU-MOSI’s distinctive feature is its comprehensive labeling of various emotions, making it an invaluable resource for developing and refining emotion identification models, especially for real-world communication applications. Its strength lies in providing a three-dimensional view of sentiment analysis by integrating text, audio, and video data.

3.5.3. Text Emotive Source

In emotion analysis, our research prioritizes precision and depth by employing advanced computational methods to decode the nuanced spectrum of human emotions. We adopt a multimodal approach that integrates text, audio, and video analyses to achieve precise emotion identification. For textual data, we utilize Long Short-Term Memory LSTM networks to capture the intricacies of emotional language. Concurrently, we analyze audio features with DNNs to discern patterns in intonation and cadence, which are essential for interpreting emotional states. Additionally, CNNs are employed to examine video features, focusing on facial expressions and body language, crucial for emotion detection.

Our audio embedding phase leverages Mel-Frequency Cepstral Coefficients (MFCC), Frequency Domain Linear Prediction (FDLP), and wav2vec-based self-supervised speech representations. This approach enhances the extraction of speaker-specific information while minimizing extraneous noise. Each feature undergoes processing through distinct branches, converging at a shared segment-level layer to strengthen speaker discrimination. This architecture boosts noise resilience and improves speaker verification performance in challenging acoustic environments.

Initially, raw audio features are illustrated through MFCC, displaying the diversity of coefficients extracted from audio samples. Subsequently, visualizations of post-DNN embeddings demonstrate the network’s capability to reduce dimensionality, resulting in 3D embeddings that effectively encapsulate the original MFCC for accurate speaker verification. These visualizations underscore the effectiveness of our multimodal approach in emotion analysis.

Our data processing framework employs a central dataset analyzed through three specialized neural network models tailored for each data type. LSTM networks process text data, emphasizing its sequential and temporal characteristics.

3.5.4. Long Short-Term Memory Architecture

The Python code below demonstrates the construction of an LSTM network using Keras, designed to process sequential data and recognize temporal patterns in emotion data. LSTM networks are a specialized Recurrent Neural Network (RNN) capable of learning long-term dependencies. The key components of the LSTM architecture used in the code are as follows: the LSTM layer learns long-term dependencies and handles sequential data.

The parameter `return_sequences=True` ensures that the output is a sequence, which is essential for stacking multiple LSTM layers. The first LSTM layer outputs sequences to be fed into the subsequent LSTM layer. In contrast, the second LSTM layer does not return sequences, thereby preparing the data for the Dense layer. The Dense layer, used for the final output, can be applied to regression or classification tasks. The Dense layer has a single neuron in this model, making it suitable for regression or binary classification tasks.

3.5.5. Audio Emotive Source

DNNs have become widely used in industry and research due to their high success rates in recognition tasks. DNNs are especially good at handling problems where multiple labels must be classified. This study focuses on distinguishing between two different phases of gait: the swing phase and the stance phase. The dataset used for training includes five parameters related to gait. A typical DNNs has an input layer, an output layer, and at least two hidden layers. Important parts of the network include the activation function, learning rate, gradient descent optimizer, and training stages. The optimizer improves the gradient descent process. One important feature of DNNs is using the softmax function, shown in Equation (3. 3). This function normalizes the neural network's outputs to a range between 0 and 1, making it very useful for tasks where multiple classes need to be predicted. In these tasks, the model assigns its prediction to the output node with the highest probability.

$$r_h = \frac{d^h}{\sum_{u=1}^b d^u} \quad (3. 3)$$

In Equation (3. 3), r_h represents the normalized output for the h -th class. The term d^h is the exponential of the raw output (logits) for class h , and the denominator is the sum of exponentials of the raw outputs for all classes u from 1 to b . This normalization ensures that the outputs sum to 1, making them interpretable as probabilities.

The following Python code demonstrates how to implement a DNNs for extracting speaker embeddings from audio files. The process involves loading and processing audio data, defining and training a DNNs, and generating speaker embeddings.

The extract speaker embedding function loads an audio file and extracts MFCC, commonly used features in audio processing. The mean of these features is calculated to obtain a fixed-size speaker embedding.

The create speaker embedding model function defines a DNNs with multiple dense layers. The softmax activation function is used in the final layer to normalize the outputs for multi-class classification.

The 'generate speaker embeddings' function iterates over audio files in a specified directory, extracts features using the 'extract speaker embedding' function, and compiles these features into an array of embeddings. The function first initializes an empty list to store the embeddings. It then filters out files that do not meet the specified naming criteria or are not in WAV format. If an error occurs during the extraction process, it catches the exception and prints an error message, ensuring the process continues without interruption.

The example usage section specifies an audio directory, and a sample audio file is processed to determine the input shape. The DNNs model is then created, compiled, and used to generate speaker embeddings for all audio files in the directory. The embeddings are split into feature and target arrays for further processing.

3.5.6. Methods for Analyzing Emotions in Audio Data

Recent advancements in DNNs have proven highly effective in recognition tasks across both research and industrial spheres, especially in multi-label classification challenges. This study zeroes in on employing DNNs to differentiate between the gait swing and stance phases, utilizing a dataset that includes five gait-related parameters. A conventional DNNs setup involves an input layer, multiple hidden layers at least two and an output layer. The essential elements of DNNs architecture, including the activation function, learning rate, gradient descent optimizer, and various training phases, are all designed to enhance the gradient descent method, especially when applied to audio, as shown in Figure 3.13. [279]. In this detailed explanation of the DNNs architecture, we highlight the practical benefits of deep learning techniques for speech emotion recognition.

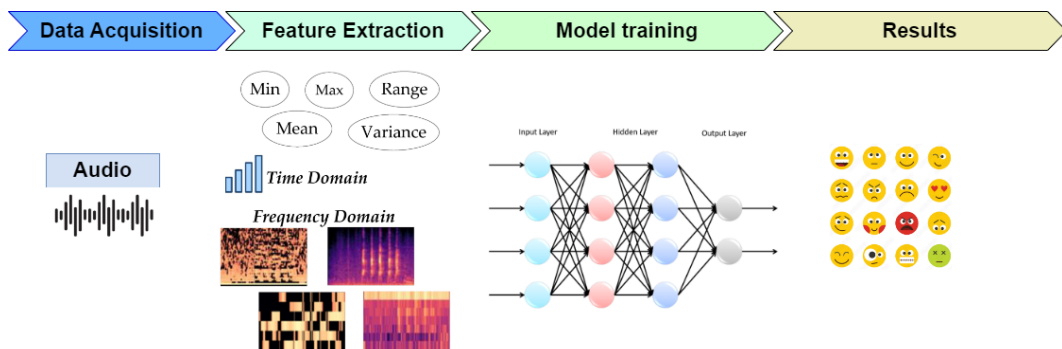


Figure 3.13. Deep Neural Network architecture.

A pivotal feature of DNNs is the softmax function application, which normalizes outputs to a range between 0 and 1, facilitating its utility in multi-classification scenarios where the model predicts the class corresponding to the highest output node probability,

as illustrated in Equation

$$r_h = \frac{d^h}{\sum_{u=1}^b d^u} \quad (3.4)$$

Method execution time exhibits a linear positive relationship with the encoding dimension, itself influenced by the number of units within the DNNs. Reducing deployment units post-layer split may affect execution time and encoding dimension, particularly in relation to offloading challenges.

3.5.7. Deep neural Network Architectures

The following Python code demonstrates the construction of a DNN using Keras. This DNN architecture is designed for processing high-dimensional data to recognize patterns and make predictions. DNNs consist of multiple layers of fully connected (Dense) neurons, which process input data through a series of transformations. The following layers are key components of the DNNs architecture used in the code:

Dense Layer: These layers are fully connected layers. After the input layer, the first layer specifies the number of neurons and the activation function (e.g., ReLU). Subsequent layers further process the features. In the provided model, the first Dense layer has 128 neurons with a ReLU activation function, the second Dense layer has 64 neurons with a ReLU activation function, and the final Dense layer has 10 neurons with a softmax activation function for multi-class classification.

3.5.8. Video Emotive Source using Convolutional Neural Networks

Densely connected layers follow convolutional and pooling layers despite variations in specific architectural designs. Traditional neural networks are reminiscent of this design. This deep structure is built by layering input, convolutional, pooling, and flattening elements. This hierarchical data integration through these layers culminates in a dense representation, ultimately leading to an output layer that signifies the final processing outcome.

GCNs or ConvNet is a deep learning model designed to learn directly from data. CNNs excels at identifying patterns in images, making them ideal for recognizing objects, classes, and categories. They also effectively classify audio, time series, and signal data. When handling large volumes of training data, CNNs often use a mini-batch approach. This section describes how CNNs process video data from the CMU-MOSI dataset. The model is compiled using the Adam optimizer and Mean Squared Error (MSE) as the loss function.

Video Frame Extraction and Embedding Generation:

This part describes extracting frames from a video and generating speaker embeddings for each frame using the CNNs model. The video is read frame by frame and

saved as an image file. These images are then preprocessed and fed into the CNNs model to obtain the speaker embeddings, which are saved as numpy files.

Loading and Evaluating Embeddings: After generating the embeddings, this section loads and evaluates them using various classification metrics. The embeddings are loaded from the saved numpy files, and metrics such as accuracy, precision, recall, and F1 score are calculated to assess the model's performance. A classification model is trained and evaluated using the generated embeddings. The model's performance is measured using accuracy, precision, recall, and F1 score. The evaluation function provides a detailed classification report that includes the model's performance for each class.

Inspired by neuronal structures in the cerebral cortex, DL architectures have demonstrated efficacy across various machine-learning applications. These architectures operate as feed-forward networks, enabling unidirectional information propagation from input to output layers. The design and functionality of these systems draw heavily from biological processes, particularly emulating the hierarchical organization of the visual cortex composed of simple and complex cell layers. Although CNN architectures can vary, they generally follow a common pattern of convolutional and pooling layers, capped by densely connected layers. This approach is similar to traditional neural networks, building depth by layering input, convolutional, pooling, and flattening components. This hierarchical data integration through these layers culminates in a dense representation, ultimately leading to an output layer that signifies the final processing outcome.

3.5.9. Application of Convolutional Neural Networks

Recent advancements in the application of Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs), and Deep Neural Networks (DNNs) have shown significant improvements in the analysis of visual and sequential data, particularly in the domain of emotion recognition from facial expressions and other temporal data. This document provides an academic explanation and Python code implementation of these networks for processing visual and sequential data from emotive sources.

3.5.10. Convolutional Neural Network Architectures

The following Python code demonstrates the construction of a CNN using Keras, a deep-learning framework. This CNN architecture is designed to process visual data and recognize emotions. CNNs are a class of deep learning models designed explicitly for processing grid-like data such as images. The following layers are critical components of the CNN architecture used in the code:

Conv2D Layer: The Conv2D layer applies convolutional filters to the input image, creating feature maps that capture various spatial hierarchies. The parameters (3, 3)

denote the filters' size. This layer is essential for extracting features from the input data.

MaxPooling2D Layer: The MaxPooling2D layer reduces the spatial dimensions of the feature maps, which helps control overfitting and reduce the computational cost. This layer performs down-sampling by taking the maximum value over an input window (typically 2x2).

Flatten Layer: The Flatten layer converts the 2D matrix of features into a 1D vector. This transformation is necessary for feeding the data into the fully connected (Dense) layers.

Dense Layer: The Dense layer, also known as a fully connected layer, connects every node in the layer to every node in the previous layer. This layer is primarily used for classification tasks after the convolutional and pooling layers have extracted features. In the provided model, the first Dense layer has 128 neurons with a ReLU activation function, and the final Dense layer has 10 neurons with a softmax activation function for multi-class classification.

3.5.11. Cumulative Attribute-Weighted Graph Neural Network

After the feature extraction phase, we introduce the CA-WGNN, which effectively integrates data from text, audio, and video sources to serve as the core of our trimodal emotion identification system. Using TensorFlow and Keras, this Python code demonstrates how to build and train CA-WGNN model. Below is a detailed explanation of each section of the code:

- **Importing Libraries:**

- TensorFlow and Keras are imported to build and train the neural network.
- Various functions from `sklearn.metrics` are imported to evaluate the model.
- `to_categorical` from Keras converts class labels to a binary class matrix.
- `train_test_split` from `sklearn.model_selection` is used to split the dataset.
- `make_classification` from `sklearn.datasets` is used to create a synthetic dataset.
- NumPy is imported for numerical operations.

- **Evaluation Metrics Function:**

- This function, `evaluation_metrics`, evaluates the trained model using accuracy, precision, recall, and F1 score.
- It predicts the class probabilities for the test data, converts them to class labels, and computes various performance metrics.

- It also generates a classification report showing the model’s performance on each class.
- **Synthetic Data Creation:**
 - A synthetic dataset with 1000 samples and 20 features is created using `make_classification`.
 - The dataset is split into training and test sets with an 80-20 split using `train_test_split`.
 - Class labels are converted to categorical format using `to_categorical`.
- **Building the CA-WGNN Model:**
 - The model architecture is defined using Keras layers:
 - * **Input** layer with a shape of 20.
 - * **Reshape** layer to reshape the input for the convolutional layer.
 - * **Conv1D** layer with 64 filters and a kernel size 3.
 - * The **Attention** layer is used to enhance the features learned by the convolutional layer.
 - * **Flatten** layer to flatten the output from the attention layer.
 - * **Dense** layer with 32 units and ReLU activation.
 - * **Output** layer with 3 units and softmax activation for multi-class classification.
- **Compiling and Training the Model:**
 - The model is compiled with the Adam optimizer, categorical cross-entropy loss, and accuracy metric.
 - The model is trained on the training data for 100 epochs with a batch size of 32 and 20% of the data used for validation.

This step finalizes the model setup and initiates training to learn from the data.

3.5.12. Model Evaluation Metrics

Evaluating a classification algorithm’s predictive performance is fundamental in machine learning. The binary classification confusion matrix is a pivotal tool for assessing a model’s effectiveness. Table 3.4. demonstrates the relationship between the model’s predictions and the actual outcomes. The matrix includes True Positives (TP) and True Negatives (TN), representing correct predictions for the positive and negative classes, respectively. False Positives (FP) occur when the model incorrectly identifies a negative instance as positive, whereas False Negatives (FN) happen when it fails to recognize a

positive instance. This matrix illustrates the model’s accuracy and serves as the foundation for calculating key performance metrics such as precision, recall, and the F1 score. [280].

Table 3.4. Binary classification confusion matrix. Image adapted from.

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

The performance of classifiers is evaluated using key metrics: Accuracy, Macro Precision, Macro Recall, and the F1 score. These metrics are defined as follows:

- **Accuracy:** This metric quantifies the proportion of correctly predicted instances in the dataset and is calculated using the formula 4. 19: TP (True Positives) and TN (True Negatives) are the numbers of correctly identified positive and negative instances, respectively, and Total represents the sum of all instances.
- **Precision:** Defined as the ratio of correctly predicted positive instances to the total predicted as positive, Precision indicates the model’s ability to avoid false positives as 3. 5:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3. 5)$$

FP (False Positives) denotes instances incorrectly classified as positive.

- **Recall:** Recall, or sensitivity, measures the proportion of actual positives that are correctly identified and is given by 4. 21: FN (False Negatives) are actual positives that were incorrectly classified as negative.
- **F1 Score:** The F1 score is the harmonic mean of Precision and Recall, providing a balance between them, especially important in the presence of class imbalances as Model evaluation metrics quantitatively assess machine learning model performance, which is crucial for evaluating prediction or classification accuracy. These metrics vary by model type (regression, classification) and analysis goals, with accuracy, precision, recall, and the F1 score being key for understanding various performance aspects.

3.5.13. Improved Accuracy and User Experience

Tailoring machine learning models to specific data modalities significantly enhances emotion recognition accuracy. This improvement bolsters the system’s fundamental

capabilities and elevates the user experience by enabling more personalized and impactful interactions. As a result, the application becomes more engaging and efficacious, underscoring the pivotal role of precision and recall metrics in evaluating model performance.

Table 3.5. provides a comparative analysis of various models for emotion recognition, highlighting their contributions and performance metrics. The table includes models such as DialogueRNN, AF-CAN, BiERU, DialogueGCN, RGCN, MMGCN, DSAGCN, GraphMFT, and TER-CA-WGNN. Each model employs techniques and architectures to capture and process emotional cues from dialogues and conversations. The performance metrics include accuracy (Acc), weighted average F1-score (wa-F1), and F1-score, which are crucial for evaluating the effectiveness of these models in real-world applications.

Table 3.5. Comparison of Various Models for Emotion Recognition

Model	Contribution	Acc	wa-F1
DialogueRNN	Uses three GRUs to model speaker information, contextual information from the preceding dialogue, and emotional information [207].	63.40	62.75
AF-CAN	Utilizes a context-aware recurrent neural network to simulate interactions and dependencies between speakers [281].	64.60	63.70
DialogueGCN	Employs a relational GCN to describe the dependencies between speakers in the graph [203].	65.54	65.04
RGCN	Generates a complex contextual feature for each individual utterance using an internal feature extractor based on ResNet [282].	65.08	–
MMGCN	Designs a spectral domain graph convolutional network to encode multimodal contextual information [283].	66.22	–
DSAGCN	Proposes a conversation emotion analysis model that combines dependency parsing with GCN [284].	63.50	61.70
GraphMFT	Constructs three graphs (V-A graph, V-T graph, and A-T graph) [204].	67.90	68.07
TER-CA-WGNN	Introduces a trimodal emotion recognition approach using a Cumulative Attribute-Weighted Graph Neural Network (CA-WGNN) to integrate textual, audio, and visual data [24].	93.00	93.00

3.5.14. Enhanced Performance and Scalability

Optimizing algorithms for distinct tasks improves data processing speeds and accuracy, elevating system performance. Employing specialized algorithms, such as convolutional neural networks CNNs for video analysis and RNNs or transformers for text processing, demonstrates the potential to accelerate data handling. Additionally, these algorithms facilitate efficient system scaling, managing increasing data volumes without sacrificing performance. This efficiency highlights the significance of accuracy and the F1 score in evaluating a system’s scalability and operational efficiency.

3.5.15. Robustness, Reliability, and Real-Time Processing

Enhancing algorithmic design to handle data variability and integrating error-checking enhances system robustness and reliability. The ability for real-time processing is essential for dynamic user interactions, highlighting the importance of algorithmic efficiency in system responsiveness. Evaluating system robustness and reliability through precision and recall metrics ensures a consistent and dependable user experience.

3.5.16. Adaptive Learning and User Interaction

Incorporating adaptive learning into algorithms allows the system to evolve in response to user interaction patterns, steadily enhancing effectiveness and personalization. This ongoing adaptation, underpinned by thorough model evaluation, guarantees the system's alignment with user needs and preferences, maintaining relevance and responsiveness over time.

3.5.17. Navigating Challenges

Integrating complex algorithms into multimodal emotion recognition systems introduces challenges such as complexity management, resource demands, and privacy and security assurances. Addressing these issues necessitates a balanced approach, with evaluation metrics providing crucial insights for refining algorithm choice and implementation. These strategic algorithm incorporations significantly influence system performance, accuracy, and user engagement. Model evaluation metrics are pivotal in this context, guiding the optimization process to improve system functionality and user interaction. By judiciously selecting and applying algorithms, mindful of the inherent challenges, developers can enhance their applications' effectiveness and adaptability, meeting the dynamic needs of users in multimodal environments.

3.6. Affective Computing in Society

AC increasingly influences our daily lives, significantly enhancing personalization in various sectors, including customer service, healthcare, and education. This technology specializes in real-time emotion recognition and response, aiming to refine user experiences with heightened accuracy and relevance. Beyond general interaction, it plays a crucial role in mental health contexts, providing essential support by comprehensively understanding and addressing individual emotional requirements. Moreover, AC advances communication by employing sophisticated natural language processing and ML algorithms. These technological improvements pave the way for more intuitive, authentic, and tailor-made user experiences in the future, marking a transformative approach to human-computer interaction.

3.6.1. Health

The integration of IoT and AIs revolutionizing healthcare through Smart Health, leveraging wearables and advanced networks like 5G for real-time health monitoring [276]. This system actively responds to demands, utilizing massive data (DaaS) and superior detail fusion (InaaS) from multiple resources [285]. Notably, changes in speech, facial expressions, and movements indicate depression [71]. The urgency for real-time emotional health surveillance is escalating, necessitating effective mental health computing. Researchers strive for precise emotion recognition to formulate appropriate treatments [84], enhancing patient monitoring in healthcare systems [41]. Within the Information Fusion field, AI merges multi-source data for health intelligence, underpinned by collective intelligence principles. Emotion recognition is vital, even for military medical staff under stress [286]. A novel hybrid approach addresses COVID-19, potentially improving diagnosis by pinpointing DNA biomarkers, facilitating prompt, accurate tests, and improved management [287]. This convergence of technology and informatics underscores the transformative potential of integrated healthcare solutions.

3.6.2. Business

Business requirements analysis is pivotal in modern firms' strategies to navigate changes in the AC e-market, spurred by social media's influence on competitive dynamics. Assessing consumer sentiment helps e-businesses understand customer opinions on their offerings. SA plays a key role in managing online brand reputation by detecting negative feedback, thus allowing immediate counteractions [54]. Identifying valued products or services enables companies to tailor marketing strategies and boost the effectiveness of targeted campaigns.

3.6.3. Security

AC involves processing sensitive emotional and biometric data, necessitating stringent security measures like secure coding, vulnerability scanning, and encryption to prevent unauthorized access or alteration. Upholding data privacy is essential, mandating informed consent for data collection and advocating for anonymization or de-identification. Secure data storage and transfer protocols are vital, alongside individuals' rights to manage their data [288], [289]. This field extends to analyzing social media content, enabling the detection of politically charged videos through emotional content assessment. Advanced techniques may predict negative polarization in posts using visual and audio cues, potentially identifying detrimental content before online dissemination.

3.6.4. Education

Emotional research in education is pivotal, influencing student experiences and educational outcomes [290]. This field involves meticulous research design, defining variables, and objectives for valid data collection, particularly using classroom action research methods. A notable intervention involved employing audio-visual media in classrooms, significantly enhancing student performance.

3.7. Challenges and Solutions

AC enhances technology interactions by recognizing and responding to emotions for better behavior development. In particular, MAC leverages audio, visual, and text modalities for improved emotional management. Yet, the move to TAC, which integrates even more modalities, poses issues. Its complexities arise from intricate algorithms, privacy concerns, and users' comfort with data sharing. One notable hurdle in this realm is SA; Challenges such as domain specificity, sarcasm interpretation, and multi-modal dynamics are prevalent. There's a push to adapt models across various domains and integrate facial and vocal indicators to tackle these. The lack of rich multi-modal emotion datasets is also a setback. Addressing this requires extensively annotated datasets that capture standard modalities and embrace physiological signals and body language. Such comprehensive datasets can propel DL in AC. Furthermore, crowdsourcing is seen as a potential way to label public social web data to make up for the lack of private sentiment data annotations. A key strategy in understanding and addressing these complexities is simplification, exemplified in tools like Figure 3.6.. By distilling intricate topics, we can gain a more profound grasp of the subject's intricacies.

MAC promises enhanced emotional management, behavior development, and improved technology interactions. Personalizing customer service with tailored responses is a significant advantage. Emotion's optimal modality for expression varies, and hypotheses in emotion classification show variations in audio, facial, and textual cues. Emotion recognition and SA using trimodal data are complex tasks that require all three modalities for effective completion.

Table 3.6. Overview of Unresolved Challenges in TER and Potential Solutions

Category	Challenges	Predictions for Solutions
Datasets	Difficulty obtaining and aligning data from all modalities. Privacy and security concerns during collection and review. SA nuances including domain dependency and sarcasm detection.	Automated multi-modal collection with timestamped alignment. Enhanced encryption and differential privacy. Adapting models and harnessing facial/vocal cues with crowdsourcing.
Feature Extraction	Emotion expression is influenced by culture and gender. Variation in dyadic dialogue content. Expression based on the speaker's situation.	Context-aware models with cross-cultural training. Contextual analysis and adaptive models considering prior situations.
Methods and Algorithms	Algorithmic complexity and opacity. Reliable emotion detection and classification.	Explainable AI models with continuous learning.
Fusion	Seamless fusion necessity in computing.	Advanced fusion techniques including cross-modal attention mechanisms.
Validation and Results	Interpretation variations among individuals over time.	Multi-rater consensus with continuous feedback loops.
Other Challenges	Ethical concerns like manipulation and discrimination. Designing emotion-centric interfaces and interactions. Users' reluctance to provide data.	Ethical frameworks with transparent policies. Emotion-aware interfaces using user-centric design. Opt-in systems highlight clear user benefits.

3.7.1. Obstacles and Alternate Approaches

Adapting models to suit diverse domains and incorporating both facial and vocal cues can effectively mitigate these concerns. While annotating sentiment in private data may be constrained, leveraging crowdsourcing can facilitate the labelling process for

publicly available data sourced from the social web. As part of TAC, the central objective revolves around creating systems and devices capable of perceiving, comprehending, processing, and simulating human emotions. Human communication typically manifests across multiple modalities, necessitating concurrently considering textual, auditory, and visual elements to glean emotional and semantic information accurately. This underscores the significance of seamless integration in practical computing applications.

The prospect of AC also brings forth several challenges and regulatory hurdles. These obstacles encompass the requirement for data originating from multiple modalities for trimodal fusion, the nuanced expression of emotions across diverse cultures and genders, and the influence of the chosen method for sentiment extraction, especially in dyadic dialogue contexts. Moreover, the increased data collection in AC raises privacy concerns, while ethical considerations concerning emotion manipulation and discrimination must also be addressed. Achieving precise emotion recognition and integration into everyday technology necessitates the development of user interfaces tailored to accommodate the intricacies of human emotion, making these challenges integral aspects of the field's advancement.

3.7.2. Privacy

The evolution of social media and AI has driven numerous applications to focus on discerning user emotions. This often requires users to send data to a server, a process susceptible to hacking and re-identification risks. Interlopers can effortlessly access sensitive details from intercepted data. A refined approach to safeguarding privacy involves transmitting data representations created on devices to the server. Subsequently, the server processes the data representation through a presupposition mechanism for additional tasks. However, sensitive demographic details can still be inadvertently disclosed through certain technologies [25].

Information such as ethnicity, age, and gender play a crucial role in employment, law enforcement, and credit assessments. Thus, a fundamental privacy objective is to exclude demographic information from Multi-modal data representations while sustaining the efficacy of emotion recognition tasks. To the best of our understanding, the inaugural study addressing privacy-augmented emotion recognition in Multi-modal data was introduced by [291]. We anticipate an increase in studies concentrating on privacy matters in the future.

3.7.3. Open Directions and Challenges Remaining

Despite substantial advancements in the sphere of MSA and emotion detection, a multitude of unresolved complexities and burgeoning avenues for research prevail. The scrutiny of inter-modality associations, alongside fusion approaches, necessitates further investigation to garner cohesive interpretations from variegated data streams. A critical

impediment to the progression of sturdy algorithms is the dearth of an all-encompassing, annotated Multi-modal dataset. In addition, there is an urgency to delve into context-sensitive analysis to augment the precision and dependability of sentiment identification. The manifestation of sentiments across cultural and linguistic variations presents additional hurdles, necessitating customized algorithms that cater to diverse demographics.

Finally, the ethical implications inherent in emotion recognition, specifically those revolving around privacy and the possibility of misuse, mandate the establishment of rigorous standards and legislation. The emergence of progressively sophisticated DL methodologies and the growing accessibility of Multi-modal data transform these challenges into stimulating prospects for future inquiry. To safeguard the continued pertinence of TAC, it is essential to place emphasis on these unresolved conundrums.

The increased use of advanced DL techniques, such as enhanced DL architectures and training strategies, has been underscored by various studies, suggesting the need for further advancements to manage the multi-modality of data better [43]. Notably, both bidirectional LSTM networks and attention-based mechanisms have demonstrated potential for sentiment and emotion classification [61], even as convolutional multiple kernel learning techniques have found utility in MER [69]. The fusion of multiple information sources is another significant area of exploration with diverse applications in sectors like healthcare, e-commerce, and the military [215], [285]. Moreover, semi-supervised approaches are gaining attention, albeit requiring further optimization for different Multi-modal scenarios [83], [72]. The potential of these technologies, particularly in healthcare, disease diagnostics, and military contexts, is widely recognized, though practical implementation and effectiveness studies in real-world settings remain open research directions [286], [287]. SA, specifically in e-commerce, has made strides, but there is a need for more accurate prediction methods applicable across languages and cultural contexts [54].

Emotional behaviors within conversations have also been examined, prompting future work to focus on real-time emotion detection in conversations using Multi-modal fusion [76], [74]. Research has yielded promising results in advanced fusion techniques, but work is needed to enhance these methods for different data types [78], [85]. Context-dependent SA, including the influence of nonverbal behaviors on word representations, is another area necessitating further exploration [74]. Novel models for emotion recognition, such as those developed through the MuSe 2021 challenge, indicate the need for more innovative and robust models that can handle Multi-modal data effectively [77]. Finally, self-attentive feature-level fusion and memory networks have shown promise for MSA, warranting further research to optimize performance and explore applicability in different contexts [278].

3.7.4. The Current State and Prospects

our study explores advanced research methods, validating findings through timely analysis, and discusses a trimodal affect analysis framework, emphasizing continual progress in the field.

3.7.5. Emerging Technologies and Applications

AC has gained traction, particularly in emotion recognition and SA, within multi-modal and unimodal contexts due to increased social media content and improved human-computer interfaces. The focus is on TAC, involving the intricate process of gathering and synchronizing data across text, voice, and visual cues. Maintaining consistency and quality in data labeling while respecting privacy and ethical standards is paramount for dependable datasets.

These datasets, spanning 2008 to 2021, are primarily extracted from platforms like YouTube, TV series, ExpoTV, and theatrical sources, offering a rich diversity in languages such as English, Spanish, Chinese, French, Portuguese, and German. This linguistic variety expands the analytical horizon of AC globally. Concentrating on MER and MSA, these resources are instrumental in building AI systems with enhanced emotional and cultural acumen, thereby revolutionizing human-computer interactions. The datasets' accessibility on public domains significantly boosts their availability for scholarly and developmental endeavors. They are crucial in cultivating AI models proficient in nuanced emotional responses and understanding, as detailed in Table

Table 3.7. The most frequently utilized trimodal datasets among researchers.

Data Set	Created By	Description
IEMOCAP	Southern California University	Contains audiovisual and 3D motion data from 10 actors with emotional expressions, annotated into categorical and dimensional labels.
MOUD	Carnegie Mellon University	Consists of 80 product review videos in Spanish, annotated with audio, visual features, and sentiments.
MELD	Enhancing EmotionLines	Over 1400 dialogues and 13000 utterances from Friends TV series, labeled with emotions and sentiments.
CMU-MOSEI	Carnegie Mellon University	Contains 23,846 utterances from YouTube videos, gender balanced, annotated for sentiments and emotions.
EmoReact	Carnegie Mellon University	Features 1102 audio-visual clips of children, annotated for basic, complex emotions, and valence.

3.7.6. Ethical Considerations and Cultural Variances

Ethical paradigms, fundamentally shaped by sociocultural distinctions, exhibit considerable diversity across nations and societal factions, underscoring the relativistic nature of ethics in the absence of universally established norms. This dichotomy, termed relativism versus universalism, manifests palpably in a globalized milieu, given the intersecting socio-cultural norms. It further accentuates the importance of culturally contingent ethical considerations. For instance, culturally disparate communication paradigms can emphasize varying elements such as punctuality or flexibility, thereby creating distinct expectations in interpersonal exchanges.

Similarly, the business sector reflects cultural influences in their ethics, as practices including gift-giving, nepotism, bribery, and corruption elicit varying reactions across cultures. Navigating these complexities necessitates respectful engagement and understanding of cultural differences, commitment towards promoting human rights, fairness, and ethical standards, and developing competencies in cross-cultural understanding and ethical literacy. The landscape of research ethics, especially about human-subject research, further embodies these cross-cultural variations. This necessitates rigorous scrutiny to ensure the proposed research activities' benignity and compliance

with the locale-specific regulations before initiation. The Charter of Fundamental Rights of the European Union, accentuating informed consent, epitomizes such legal frameworks. Adherence to jurisdiction and field-specific ethical standards, securing informed consent, and obtaining relevant research ethics committees' approval emerge as non-negotiable imperatives in ethical research conduct [292].

3.7.7. Real-World Instances of Affective Computing

In recent days, there has been a rapid increase in emphasis among research teams on the practical implementation of affective computing in real-world scenarios. "Sentic-Net, under the leadership of Erik Cambria from Nanyang Technological University (NTU)," incorporates the outcomes derived from emotional computing and sentiment assessment into several aspects of everyday life [293], encompassing HCI, financial domains, and the monitoring of social media platforms.

The utilization of textual sentiment analysis is observed in recommender systems, where it effectively utilizes a range of feedback messages. Utilization of visual emotion analysis technology has been found to have significant implications in various domains [294], including educational instruction, decision support systems, HCI, and assistance for those with audio spectrum disorder. Integrating multimodal affective analysis techniques enhances the system's resilience and durability, creating opportunities for a wide range of practical applications in real-world scenarios.

3.8. Classificatory and Predictive Dimensions

As part of AC, classification and prediction are closely related steps. Type assigns a class label to each sample based on the features extracted from the input data. At the same time, prediction uses the patterns and relationships learned from the training data to predict future data. A class label might be a different emotion and sentiment for the case of affective classification and prediction. Type involves assigning a classification label based on the features extracted from the data to each input sample. In the prediction step, patterns and relationships learned from the training data are used to predict emotions expressed in future data. The specific approach will depend on the data characteristics and the task needed for classification and prediction in AC. Significant algorithms emerge from the data in reviewing the trimodal research on sentiment analysis and ER. Below, we outline several prevalent methods in AC that leverage diverse approaches and classifiers. Figure 3.14. and 3.15. illustrate these architectural frameworks for clarity. These approaches and algorithms form the focus of research and development efforts in multimodal ER, enabling the exploration of various methods to recognize and understand emotional expressions more accurately.

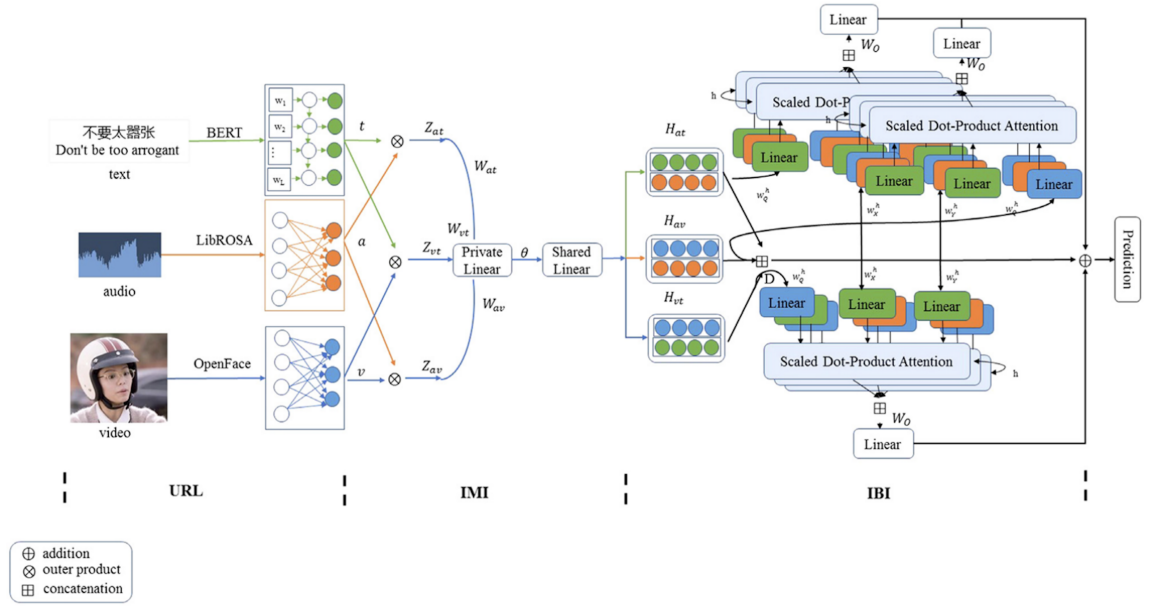


Figure 3.14. The network architecture of BIMHA for affective computing [66].

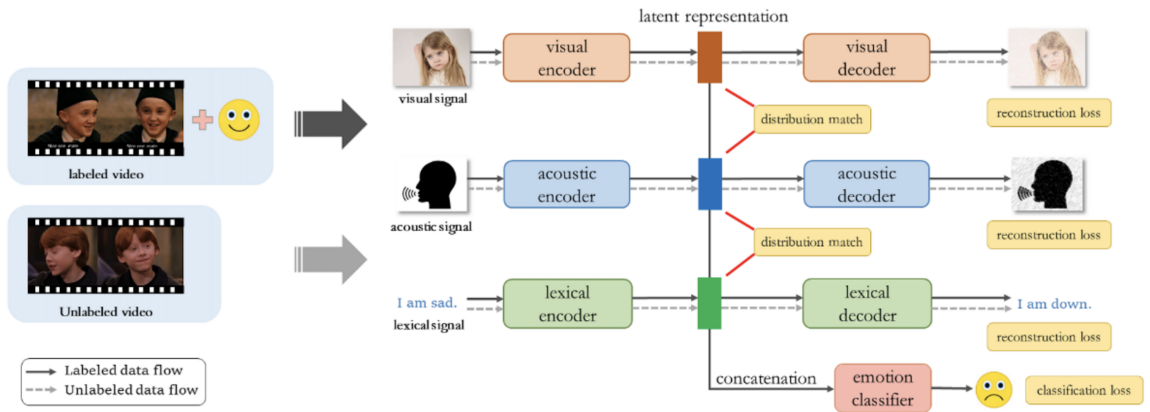


Figure 3.15. The semi-supervised multimodal emotion recognition framework [83].

3.8.1. Classification

In the classification step of AC using text, voice, and visual modalities, the goal is to assign a class label to each input sample based on the features extracted from the data. For example, in the case of emotion classification, the class labels might be different emotions, such as happiness, sadness, anger, and so on. Several different approaches can be used for variety, including:

- **Supervised learning:** A model can be trained on a labeled dataset with input features and class labels. Based on the pattern and relationships learned from the training dataset, it can predict the class label for new input data. The specific algorithm will depend on the characteristics of the data and the specific task at hand. Using techniques like multi-task and semi-supervised learning for ER. Emphasis on accuracy is paramount [295]; ML algorithms like decision trees, SVMs, and NN can be used for supervised learning.
- **Unsupervised learning:** In unsupervised learning, the model is not provided with class labels and must learn to identify patterns and relationships in the data independently. This can be useful for tasks such as clustering, where the goal is to group similar samples without prior knowledge of the class labels. Unsupervised learning algorithms do not require a labeled training dataset and can help explore and understand the underlying structure of a dataset. Some standard unsupervised learning algorithms include k-means, hierarchical, and density-based clustering.
- **Semi-supervised learning:** Semi-supervised learning is a hybrid approach that combines supervised and unsupervised learning. It is often used when limited labeled data is available and can be more effective than supervised or unsupervised learning alone in these cases. Some standard algorithms used for semi-supervised learning include self-training, co-training, and multi-view learning.
- **Semi-supervised:** Using both labeled and unlabeled data for training, semisupervised learning enhances model accuracy, especially when labeled data are scarce. In robotics and game playing, semisupervised reinforcement learning uses this approach in partially known environments to learn actions to maximize rewards.

3.8.2. Prediction

In predictive analytics, both historical and current data are utilized to forecast activities, behaviors, and trends. These predictive models assess input datasets to anticipate future events and outcomes. Intelligent algorithms analyze combined features to predict emotions. This analysis is a vital aspect of predictive analytics. The model's performance is largely influenced by the nature of the data and the specific prediction algorithm employed. When choosing a prediction algorithm, it's essential to consider the distinct characteristics of the data and the desired model performance. Moreover, methods

for interpretability have been established to identify key features within each modality that are instrumental in predicting specific emotional categories [86].

3.8.3. Recommendations and Development

In AC, the most insightful reviews merge AI with human psychology, emphasizing the growing significance of data extraction within this domain. The trimodal approach stands out for its ability to dissect and discern complex human emotions through diverse techniques. Nonetheless, it does demand substantial computational resources and raises potential privacy concerns. The advantages of this methodology are predominantly seen in improved human-computer interactions and decision-making processes. These improvements appear to be fueled by leaps in computational power, an abundance of data, and a continuous quest for more accurate emotion detection methods.

The trimodal research review for SA, presented herein, encompasses work from various authors and utilizes an array of datasets, fusion mechanisms, methods, languages, and accuracy levels. One of the remarkable peaks is Zadeh’s application of Word-Level Fusion via MFN+DFG on the CMU-MOSI dataset, which recorded an 85.2% accuracy rate. In a similar vein, T. Wu achieved a commendable 86.2% accuracy using Bimodal Fusion with BIMHA across several datasets, such as CMU-MOSI and IEMOCAP. Conversely, certain efforts did not meet these standards. For instance, Q. Portes’ utilization of LSTM, incorporating both hierarchical and non-hierarchical fusion on the MELD dataset, reached only a 57.50% accuracy. Similarly, W. Yu’s strategy of combining a Memory Fusion Network with DNNs and LSTM on the CH-SIMS dataset culminated in a 64.74% accuracy. Intriguingly, a recent contribution by S. Mai et al. [87] heralded a substantial leap in AI through the introduction of the HyCon model, which notably set new standards in SA benchmarks. This model, along with its subsequent iteration, HyCon-v2, harnessed advanced language encoders for contrastive learning. Impressively, they surpassed earlier models, especially those incorporating BERT. Yet, the study’s emphasis on specific architectures like BERT and XLNet stirred discussions regarding their breadth and adaptability in wider scenarios.

The specific outcomes of these studies are depicted in Table 3.8.. Sentiment Analysis (SA) continues to evolve rapidly, and marked strides in Multi-modal data analysis, fusion techniques, and model innovation are evident. Collectively, these studies are expanding our comprehension of human emotions, underscoring the profound capabilities of Artificial Intelligence AI in navigating this intricate territory. A pivotal aspect of this analysis is the computation of methodological similarity, predicated on accuracy metrics. This evaluation elucidates the relative performance and effectiveness of various fusion approaches and methodologies employed across different studies. The accuracy metric, serving as a quantitative indicator of performance, enables a comparative analysis that highlights the efficacy of different models in understanding and interpreting human emotions.

Table 3.8. succinctly presents the top five studies in sentiment analysis research,

characterized by their superior accuracy rates. These studies leverage a spectrum of fusion techniques, including Hybrid Fusion, Bimodal Fusion, and Word-Level Fusion, to analyze and interpret multimodal data. Such diversity in approaches underscores the multifaceted nature of sentiment analysis and the creativity in methodological application within the field. Notably, the highest accuracy achieved, as demonstrated by Aslam A.[64], underscores the significant potential of Hybrid Fusion techniques in enhancing model performance. Moreover, the studies encapsulated within this table traverse a variety of datasets, such as IEMOCAP, CMU-MOSI, CH-IMS, and CMU-MOSI, reflecting the wide applicability and robustness of the proposed models across different contexts and languages. This versatility is crucial for advancing sentiment analysis, as it indicates the models' adaptability to diverse linguistic and cultural nuances inherent in human emotion expression. This table serves not only as a showcase of the leading accuracy achievements within sentiment analysis research but also as a testament to the ongoing innovation and exploration that is taking place in the field as a whole.

The methodological diversity and the high accuracy rates achieved are indicative of substantial progress being made in understanding and interpreting complex human emotions through AI, marking a significant step forward in sentiment analysis. Moreover, the Trimodal Research Review for Emotion Recognition systematically explores various fusion techniques and methodologies within Automated Content Analysis AC, showcasing the field's heterogeneity and identifying promising avenues for future exploration.

The detailed in Table 3.9. underscores our review's contributions, spotlighting significant advancements in accuracy that set new benchmarks. These highlighted studies exemplify the methodological variety and innovative approaches that have markedly propelled the capabilities of emotion recognition systems forward.

Table 3.8. Top Five Accuracy in Trimodal Sentiment Analysis Research

Authors	Method	Accuracy(%)
Aslam A.[64]	AMSAER	93.14
Wu T.[66]	BIMHA	86.2
Liang P.[60]	LSTHM	85.4
Zadeh A.[67]	MFN+DFG	85.2
Peng J.[65]	AMSAER	83.59

Our review maps the current landscape and stimulates further research by elucidating the complex interplay between different fusion types and their application in emotion analysis. It invites the academic community to build upon these foundational successes, suggesting that the future of emotion recognition lies in the exploration of uncharted methodologies and the creative integration of multimodal data. This forward-looking perspective emphasizes the critical role of diversity in research approaches

and fusion strategies as catalysts for breakthroughs in understanding and interpreting human emotions through technology.

The exceptional accuracy demonstrated by P. Kumar et al. [86] not only underscores the efficacy of integrating diverse modalities through hybrid fusion techniques and DNNs but also signifies a paradigm shift in how emotion recognition technologies are developed and implemented. This leap in accuracy from using the IIT Roorkee Multimodal dataset showcases the untapped potential in the nuanced combination of textual, visual, and auditory data, opening new avenues for research and application in areas ranging from interactive media to therapeutic settings.

The diversity in fusion methods, ranging from decision-level to feature-level, further accentuates the rich tapestry of strategies available to researchers aiming to refine the sensitivity and specificity of emotion recognition systems. Such a variety enriches the toolkit for scientists and engineers. It provides a more granular understanding of how different data types can be synergistically combined to mimic human-like emotional intelligence. Moreover, the transition towards more complex, hybrid models reflects a broader trend toward embracing complexity and nuance over simplicity and generalization. This evolution is not just about achieving higher accuracy; it is about developing systems that can understand and interpret human emotion’s subtleties in a meaningful and contextually relevant way. The use of advanced machine learning models and innovative fusion strategies, as documented in Table 3.9., indicates the field’s move towards systems that can operate with a degree of empathy and intuition previously thought to be the exclusive domain of humans.

Table 3.9. Top five accuracy in trimodal emotion recognition research

Authors	Method	Accuracy (%)
P. Kumar [86]	DNNs	95.99
S. Dham [71]	GMM, SVM	87.01
Y. Wan [73]	RAVEN	86.7
S. Mai [87]	HyCon	83.8
T. Mittal [75]	M3ER	82.70

Consequently, the table encapsulates a set of numerical achievements and a significant chapter in the ongoing narrative of emotion recognition research. It highlights a period of transformation where the confluence of multimodal data, cutting-edge fusion strategies, and sophisticated machine learning models heralds the onset of breakthroughs. The documented findings testify to the field’s rapid progress and guide future explorations to enhance the accuracy and depth of emotional understanding. This ongoing journey, marked by relentless innovation and discovery, promises to diminish further the boundary between human emotional insight and artificial intelligence, fostering more responsive,

empathetic, and ultimately human-centric technologies. In this chapter, we have delved into the intricate processes and methodologies of multimodal emotion recognition, underscoring the superior efficiency of multimodal systems over unimodal approaches. The detailed steps for acquiring, collecting, and cleaning data, along with the innovative graph visualization techniques, lay the groundwork for a unique and robust multimodal fusion framework. This framework, a novel contribution of our research, is crucial for integrating complex trimodal data, thereby enabling more accurate and comprehensive emotion recognition. The comparative analysis approach we adopted in this study, using various trimodal emotion datasets, has demonstrated the ability to capture a wide range of human expressions, both acted and natural. This methodological rigor not only highlights the robustness of our approach but also sets a new benchmark for future research in this domain, potentially shaping the direction of future studies in multimodal emotion recognition.

Our integration of emotional information from text, audio, and visual cues significantly enhances performance, surpassing the effectiveness of any single modality. This advancement is particularly evident in the comprehensive analysis of spoken content, speech models, and meticulous frame-by-frame facial expression analysis. The innovative application of graph neural networks in this context has shown remarkable potential in capturing the intricate dependencies between different modalities, further boosting the accuracy and reliability of emotion recognition systems. The illustrative figure on trimodal fusion in affective computing provides a clear and structured overview of the data acquisition and preprocessing methods, current trends and methodologies, modality-specific insights, and this technology's practical applications and significance. This holistic view underscores the transformative impact of integrating different aspects of affective computing for efficient and effective emotion recognition and analysis.

Overall, the contributions of this chapter are pivotal in advancing the field of multimodal emotion recognition. The insights and methodologies presented here not only lay a solid groundwork for future research and development but also promise significant advancements in various applications. From enhancing human-computer interactions to improving mental health monitoring and personalized user experiences, our work addresses current challenges and opens new avenues for exploring the vast potential of multimodal affective computing in understanding and responding to human emotions with unprecedented accuracy and depth.

4. TER-CA-WGNN: TRIMODEL EMOTION RECOGNITION USING CUMULATIVE ATTRIBUTE-WEIGHTED GRAPH NEURAL NETWORK

Understanding the feelings of other people is an essential skill for day-to-day life. People's responses and behaviors are impacted as soon as their feelings are noticed by other people. If one believes another person is angry, they should approach them with care [296]. It is essential for a wide variety of sectors and businesses, such as AI, gaming, entertainment, HCI, surveillance, and robotics, to have the capability to understand and identify emotions in sensor data [297]. In this assignment, rather than trying to recognize emotions in real life, concentrate on identifying emotions, as shown in the assignment. Collecting data from which individuals express their emotions is vital to constructing efficient AI systems that can identify emotions [298]. There are many different types of modalities; some examples are body language, gestures, speech, written material, voice modulations, facial expressions, and walking patterns.

Recognizing human emotions' influence on our day-to-day lives is essential, which is HCI systems that can recognize these emotions are becoming more crucial [299]. People operate in a manner that is consistent with their emotional perceptions so that they can respond to environmental stimuli. The ability to comprehend human feelings and behaviors is a capability that is very important for intelligent systems, those used in the fields of medicine, robotics, and surveillance [300]. The first and most important phase in the process of emotion identification is the collection of a wide variety of data formats that represent human emotions. An inherent multi-modality characterizes the manifestation of human emotions to the phenomenon [301]. There are several ways in which a person's emotional state can be communicated, including their voice intonation, pace, facial expressions, word choice, and body language. Based on this, it is reasonable to assume that using a variety of modalities instead of depending on one can provide more favorable outcomes [302]. The assistant editor who was responsible for reviewing the examination of the article and giving final clearance for publishing was tough to get. Considering that humans are not particularly adept at immediately distinguishing distinct emotional categories in text, audio, or video, it is vital to have evidence that has been accurately labeled [303]. Researchers in emotion recognition have enabled public access to datasets for identifying emotions in both text and images [304]. However, there is a noticeable lack of datasets encompassing all three modalities, audio, text, and video, simultaneously, unlike those that are limited to single or dual modalities. Creating such comprehensive, trimodal datasets involving video, audio, and text is a complex and costly endeavor [305]. This work evaluates the efficient exploitation of datasets containing varied modal

information. The primary objective of this research is to advance the field of ER by introducing and validating a novel trimodal approach CA-WGNN.

Novelty and Contributions of the Study

This study introduces a groundbreaking approach to emotion recognition with the novel Cumulative Attribute-Weighted Graph Neural Network (CA-WGNN) model. Diverging from the conventional focus on primary emotions employs the Interactive Emotional Dyadic Motion Capture IEMOCAP and Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSI) datasets. This approach is a significant deviation from existing literature, marking our investigation as the first to apply the CA-WGNN model to these specific datasets, thereby addressing a notable gap in the field.

The innovative edge of our study lies in the CA-WGNN model's ability to capture nuanced emotional states, a facet often overlooked in prior research. By broadening the horizon of emotion recognition beyond traditional categories, we introduce a novel perspective to the domain. The following points encapsulate the core contributions of our study:

- Employing the CA-WGNN model to analyze textual, audio, and video data, using the CMU-MOSI and IEMOCAP databases as trimodal inputs, enhancing the accuracy and scope of emotion recognition (ER).
- Introducing a unique methodology for trimodal emotion identification through the CA-WGNN model, integrating auditory, visual, and textual data.
- Innovating by combining data from multiple modalities through a weighted graph structure within the CA-WGNN model is a significant advancement in multimodal affective computing.

4.1. Literature Review

The study [306] suggested the DL technique is used to propose an emotion identification system based on emotional Big Data. Big Data includes audio and video. The suggested system processes voice signals in the frequency domain to create Mel-spectrograms, which can be used as images. Next, the Mel-spectrogram is supplied to a CNNs. To process video signals, the CNNs is fed representative frames from a segment. The study [307] suggested how well an LSTM mechanism based on DL can identify emotions in text. The research used the Emotion classification dataset, which categorizes emotions into six distinct sets. The findings of the experiment show that in comparison to other learning approaches, LSTM-based text emotion categorization offers considerably greater accuracy. The research [92] proposed a speech-text multimodal emotion detection model to improve emotional recognition system performance. The CNNs and LSTM were combined in binary channels to learn acoustic emotion features, while an effective

Bidirectional- LSTM network was used to capture textual features. Evolution of the Cumulative Attribute Approach in Computer Vision Regression initiated by [308] in 2013, the Cumulative Attribute approach in computer vision regression has significantly evolved, addressing complex challenges in fields like age and crowd density estimation. This novel approach centers on mapping sparse and imbalanced low-level visual features to a 'cumulative attribute space,' where each dimension is semantically meaningful. It reflects the continuous nature of scalar outputs such as age or people count. By ensuring both discriminative clarity and cumulative conditioning among attributes, Cumulative Attribute effectively utilizes the correlation among adjacent scalar values, showcasing substantial improvements in accuracy and efficiency, especially in scenarios with limited and unevenly distributed training data.

The study [161] introduced a new kind of DNNs that can use text, video, and audio to recognize emotions. Aiming to learn the representation for each modality and the best-combined representation to get the best prediction, the proposed DNNs architecture's independent and shared layers work together. Building upon this foundation, the Cumulative Attribute approach in 2019 was extended to multifaceted tasks like head pose estimation and color constancy [309]. The essence of Cumulative Attribute in this expansion lies in its ability to innovatively map input features to cumulative attributes, capturing intricate correlations between various target values. This method overcomes the limitations of traditional regression models by addressing the interdependencies they often ignore, leading to a more holistic and accurate analysis of multivariate data. Originally effective in single-output problems, the extension of Cumulative Attribute into multi-output regression underlines its versatility and potential in broader applications. The implementation of Cumulative Attribute involves a two-stage process: the initial stage focuses on attribute learning mapping from the feature space to the AC area, followed by the second stage, which maps these attributes to the output space. This approach significantly enhances the capability of regression models in visual tasks by leveraging the target space and employing attribute functions corresponding to various target variables.

The article [46] focuses on evaluating various pre-trained language models' performance and environmental impact in text analysis. It highlights the significant trade-offs between model efficiency and ecological footprint in natural language processing, underlining the importance of environmental considerations in developing and applying these technologies in sentiment analysis. The study [310] suggested the methods also neglect merging features from many sources for use in machine learning. Research should focus on developing LSTM-based models capable of handling multi-modal feature fusion and fully accounting for utterance relations. The study [165] proposed a new method for face emotion identification based on DL. To start with, the acquired dataset is noise-removed using a combined trilateral filter. After that, the filtered images are enhanced by using contrast-limited adaptive histogram equalization (CLAHE). The research paper [311] examined HCI, SER is a hot subject of study. To assess the performance of various

Artificial Neural Networks models, the study trains one utilizing MFCC's feature extraction and tests it on certain audio datasets.

The paper [312] presented an approach that captures the interlocutor and contextual states between utterances using a RNNs. Before fusing, the relationship between the modalities and their relative significance is understood via pairwise attention. The research [313] examined focuses on the topic of bias in emotion detection systems and how it relates to the modalities used. It examines the effects of multimodal techniques on system fairness and bias. The study [314] suggested that depth and heat detection models are trained using state-of-the-art DL techniques. To analyze publically accessible data in combination for the purpose, this is used in the training process to provide the appropriate annotations for a learning process. The article [157] maintained potential uses in HCI, ER has lately garnered a lot of attention.

Many spoken and non-verbal languages, such as text, images, and audio, convey human emotions. So, rather than an issue for single-modal learning, emotion detection works well as a multi-modal one. The study [315] investigated image, text, and tag-based emotion identification in this study. Nevertheless, an additional, disregarded obstacle might arise from using many modalities: the missing modality dilemma. Because not all social media posts have text, images, and tags, tests are not uncommon to miss one or two modalities. The study [316] suggested Attention-based Multimodal Sentiment Analysis and ER to address these difficulties. This paradigm uses visual, auditory, and textual intra- and inter-modality discrimination. Focusing on important tasks helps categorize sentiment and emotion from visual, spoken, and audio inputs. Modality-specific algorithms automatically extract semantic phrases, image regions, and audio features.

The utilization of Graph operations GNNs has emerged as a pivotal element in affective computing, particularly in emotion recognition. Alsaadawi and Das [19] underscore this significance through their innovative approach. Their research highlights how GNNs adeptly capture the intricate inter-modal relationships within multimodal data, facilitating more accurate and nuanced emotion detection. This advancement represents a significant leap forward in understanding and interpreting complex emotional states using computational methods.

The study [317] examined the : Proactive Service Recommendation Driven by Multimodal Emotion Recognition (PSRMER), which is a method that the study suggests as a solution to this problem. Using a TER model based on Transformer and PSRMER, the user's emotion is actively recognized first. The investigation [318] suggested as more and more researchers focus on the possibility of strong emotional bonds between humans and computers, there is a growing demand for reliable and practical methods for detecting people's emotional reactions. The study presents an EEG recognition model that uses a bandpass filter to pre-process the input signal.

The study [319] suggested the process of recognizing and comprehending emotions via integrating input from several modalities, including text, image, and audio, is known as

TER. Nevertheless, a major obstacle to this effort is the lack of labeled data. To do this, this research suggests a unique strategy that combines label correction techniques with consensus decision-making inside a semi-supervised learning framework. The paper khalane2023evaluating suggested highlights the dangers of using DL models with opaque architectures for crucial tasks like emotion identification and explains the fundamental need for human-readable explanations of the model’s inner workings.

The article [320] proposed MSMDFN, or multi-stage multimodal dynamical fusion network, aims to address this issue. The MSMDFN serves to acquire the cross-modal correlation-based joint representation. The article [321] discussed emotion categorization using a deep sequential model and a representative characteristic called Complex MFCC. To account for small differences in the underlying phonemes, the experimental design does not rely on the speakers involved. The study [322] approached TER called Contextualized Graph Neural Networks (COGMEN) is presented in the research. This system makes use of both local and global information and context. GNNs are the building blocks of the suggested model, which attempts to capture all the interdependencies in a dialogue.

The study [323] examined the recent advances in emotion identification and instances of their use are also included in this overview. Additionally, this poll evaluates several emotion detection sensors and contrasts their pros and cons. Researchers can learn more about current emotion detection systems with the aid of the suggested survey, which will make it easier to choose appropriate sensors, algorithms, and datasets. The study [324] suggested multimodal emotion identification algorithms are receiving more attention from researchers; it is unclear how certain settings might benefit from the combination of visual and non-visual data for emotion detection. The research examines the interaction between two important contextual elements and multimodal emotion components collected from text, tone, and facial expressions. The study [325] suggested that video emotion recognition is plagued by multimodal feature fusion. Fusing mode-specific feature matrices through neural networks is becoming common in DL. Multimodal analysis requires correlations and effective Unimodal features, unlike Unimodal issues. The study [326] examined the need for human-understandable interpretations of black-box DL models highlighted in this paper, which focuses on the hazards of employing these algorithms for vital tasks like emotion identification.

The article [327] examined low light, non-positive faces, and facial occlusion as a few of the natural challenges that make dynamic expression identification difficult in the field. Approaches that depend on vision can fall short when it comes to capturing the intricacies of human emotions, and the framework can efficiently extract multimodal information and provide considerable improvement, allowing us to tackle this problem. This study [328] suggested that as the understanding of emotions grows and new technologies become available, the work builds a reliable emotion identification system that can be used in real-world settings. The article [46] three corpus containing e-commerce data belonging to different languages such as Turkish, Arabic, and English were created and the

performances of deep learning and machine learning methods were examined comparatively by performing sentiment analysis on them.

The study [329] offered a unique multimodal emotion identification framework that names multimodal emotion identification based on cascaded multichannel and hierarchical fusion (CMC-HF). This framework concurrently uses visual, voice, and text data as multimodal inputs.

4.2. Proposed Approach for Emotion Recognition

Exploring the intricate domain of emotion analysis, our research emphasizes precision and depth, utilizing sophisticated computational methods to unravel and interpret the complex spectrum of human emotions. We implement an extensive, multi-modal approach that merges text, audio, and video analyses for the accurate identification of emotions.

This methodology includes the extraction and processing of textual information via speaker embedding in LSTM, a key aspect in grasping the nuances of emotional language. Simultaneously, audio features are analyzed using DNNs, which discerns patterns in intonation and cadence, essential for understanding emotional states. Complementing this, video features are scrutinized using CNNs to interpret facial expressions and body language, key components in emotion detection. In the audio embedding phase, our approach utilizes MFCC [330], frequency domain linear prediction (FDLP), and wav2vec-based self-supervised speech representations. This integrated strategy enhances speaker-specific information extraction while minimizing extraneous data such as noise. Each feature is processed through independent branches, converging at a shared segment-level layer, thereby reinforcing speaker discrimination. This architecture is particularly adept at improving noise resilience and speaker verification performance in challenging acoustic environments.

The visualizations in Figure 4.1. provide a dual perspective. Initially, raw audio features are displayed through MFCC, illustrating the variety in coefficients extracted from audio samples. This is followed by post-DNN embedding visualizations, which demonstrate the network’s capability to reduce dimensionality. The result is a set of 3D embeddings that effectively encapsulate the original MFCC for accurate speaker verification, thus highlighting the efficiency of our multi-modal approach in emotion analysis.

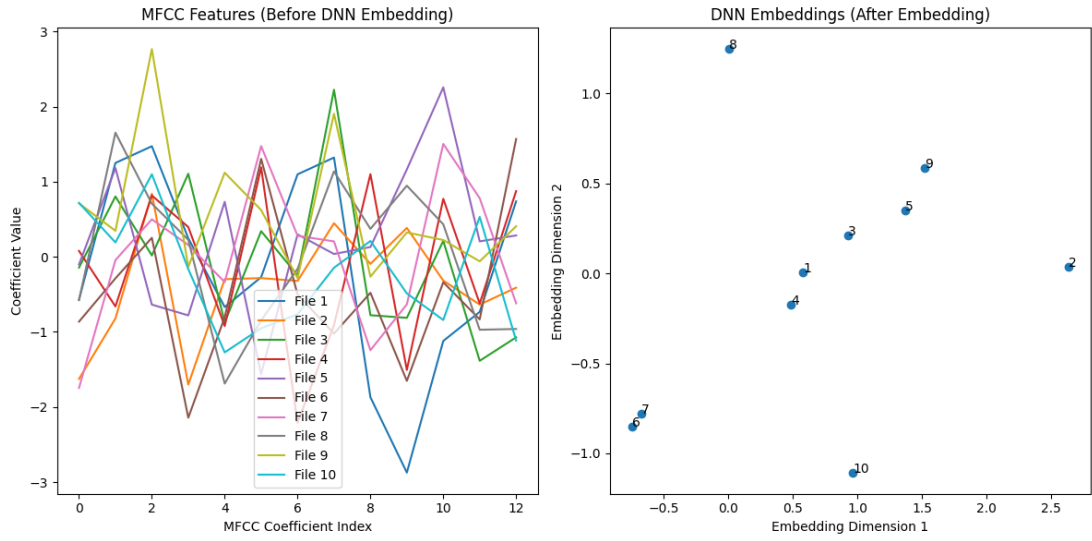


Figure 4.1. Graphic flow.

As illustrated in Figure 4.2., our data processing framework utilizes a central dataset that is analyzed through three different neural network models for each type of data. Text data is processed using LSTM, emphasizing the sequential or temporal aspects of text. DNNs are used for processing audio data, focusing on extracting complex features such as intonation and cadence, while CNNs are employed for video data, targeting the analysis of facial expressions and body language.

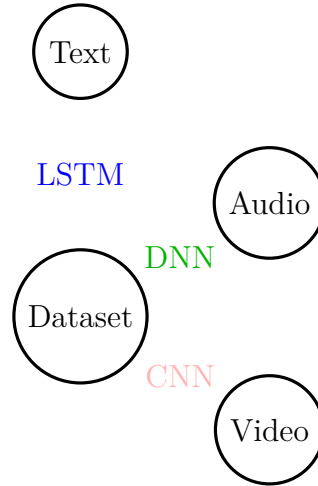


Figure 4.2. Multimodal data processing framework.

Following the feature extraction phase, we introduce the Cumulative Attribute-Weighted Graph Neural Network CA-WGNN. This forms the core of our trimodal emotion identification system, effectively integrating data from text, audio, and video sources. For a comprehensive understanding of our methodology, please refer to

Figure 4.3.. This figure outlines the flow of our process, providing visual insight into how each component synergistically contributes to emotion identification.

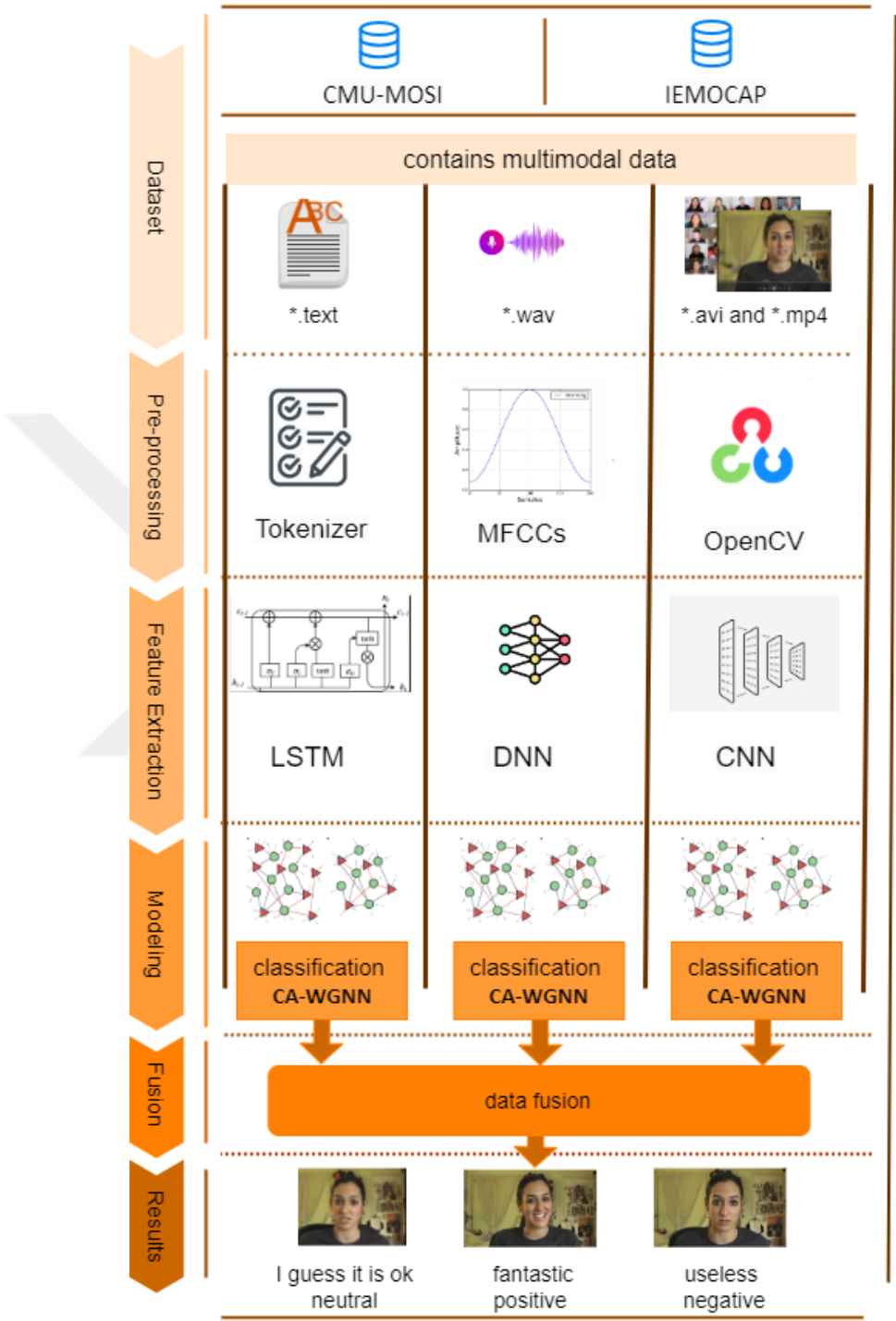


Figure 4.3. Proposed approach flow chart.

4.2.1. Dataset Description

Both academics and businesses are increasingly drawn to the study of subjectivity and emotion in online opinion videos. This emerging field aims to understand the nuanced sentiments expressed in these digital narratives. Unlike the more explored domain of text-based sentiment analysis, video, and multimedia sentiment analysis offers a unique challenge. It requires delving into the complex interplay of visual cues, vocal tones, and spoken words to discern underlying emotions and perspectives. The primary hurdle in advancing this field is the lack of a comprehensive dataset, along with established methodologies, baselines, and a thorough statistical analysis of the interactions between different modalities of data.

To address this gap, a significant dataset was compiled and is (multicomp.cs.cmu.edu accessed on 26 February 2024.) [67]. This dataset serves as a cornerstone for researchers and practitioners alike, enabling them to explore and analyze the rich tapestry of emotions and subjectivity present in online opinion videos. With this dataset, we can understand the emotional landscape of these videos and develop and refine multimodal sentiment analysis tools.

Dataset IEMOCPA

The IEMOCAP dataset [231], originating from the SAIL lab at USC, represents a comprehensive database that is a goldmine for researchers delving into the nuances of human emotions. Spanning twelve hours of rich data, it includes video, audio, facial motion capture, and text transcriptions, offering a holistic view of emotional expression. This diverse dataset transcends traditional boundaries by encompassing text, voice, images, and physical signals, thereby providing a multifaceted perspective on emotional communication. Annotated with both dimensional and categorical labels, the database focuses on dyadic encounters designed to naturally elicit a range of emotions.

This approach ensures the capture of genuine emotional responses in a controlled setting, making the data highly valuable for nuanced analysis. For scholars and practitioners in the field of multi-modal expressive human communication, the IEMOCAP dataset stands as an invaluable resource. It not only offers insights into the complex world of emotional expression but also provides a foundational platform for developing advanced Sentiment Recognition systems.

Dataset CMU-MOSI

The CMU-MOSI [67] dataset takes center stage in the realm of Sentiment Recognition, showcasing the power of multimodal sentiment analysis. This innovative dataset is a treasure trove of emotional insights, capturing a diverse range of emotions through spoken words. It does so by skillfully combining visual and auditory data, thus offering a holistic view of human sentiment expression. What sets CMU-MOSI apart is its comprehensive labeling of a

wide array of emotions. This feature makes it an invaluable asset for developing and refining emotion identification models, particularly for applications in real-world communication scenarios. The dataset's strength lies in its ability to provide a three-dimensional view of sentiment analysis by integrating text, audio, and video data.

4.2.2. Speaker Embedding

The journey of analyzing emotions in multimodal data begins with a meticulous feature extraction process, encompassing text, audio, and video modalities. For textual data, the method of speaker embedding in LSTM networks is employed. This technique is adept at capturing the subtle linguistic nuances that are often intertwined with emotional expressions, thereby providing a deep understanding of the spoken word's emotional undertones. When it comes to audio data, DNNs come into play. These networks are fine-tuned to extract audio features, specifically targeting the unique patterns found in speech intonation and cadence. This aspect of feature extraction is crucial, as the way we speak often carries a wealth of emotional information, sometimes even more telling than the words themselves.

For the video modality, the focus shifts to CNNs. These networks are particularly effective in parsing through visual data to identify key emotional indicators. They analyze facial expressions and body language, which are vital components of non-verbal communication. These visual cues are often the most direct and powerful conveyors of emotion, and their accurate interpretation is essential for a comprehensive understanding of multimodal sentiment analysis.

Text Speaker Embedding Using LSTM

An LSTM network is a marvel in the field of neural networks, distinguishing itself with its unique structure. Unlike traditional neural networks, LSTM are designed with interconnected nodes in their hidden layers. This architecture allows for the sequential flow of information, where the output of one layer seamlessly feeds into the input of the next. Such a configuration is particularly adept at handling time series data, which is a common characteristic of textual information in speech. The heart of the LSTM lies in its hidden layer, where time series data find their temporal abode. Figure 4.4. provides a visual representation of the LSTM's fundamental design [331], illustrating how it adeptly manages data over time.

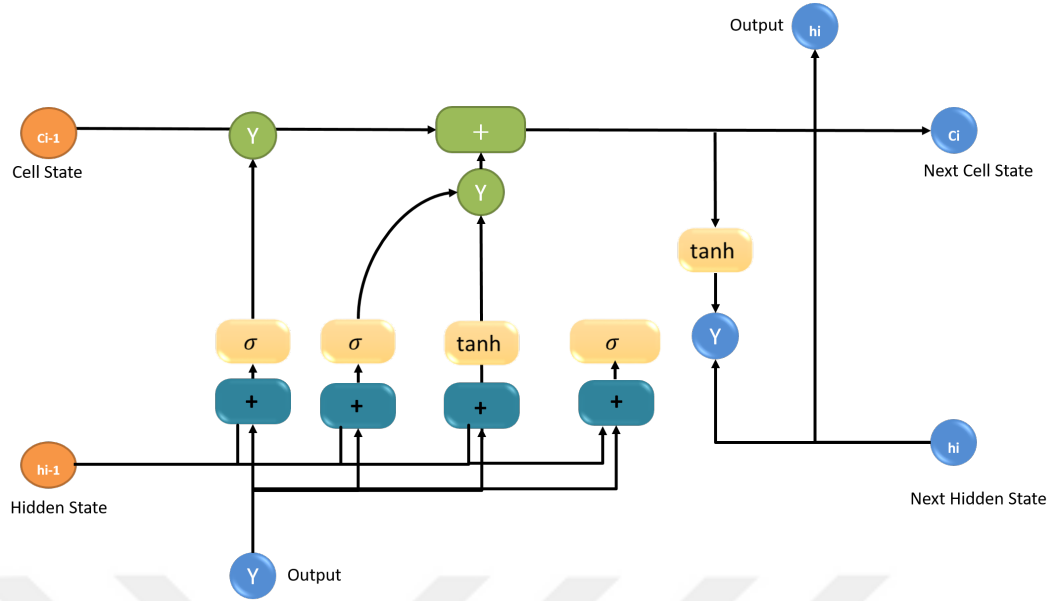


Figure 4.4. Structure of LSTM method.

A critical component of the LSTM is the 'cell', the basic unit of memory within the network. Each cell is responsible for transmitting not only the cell state but also the hidden state to the subsequent cell, creating a continuous flow of information. This flow of text data through the LSTM chain is unique. The data can traverse through the cell state with minimal or no changes, although some linear transformations may occur. The real magic of LSTM lies in its use of sigmoid gates. These gates act as regulators, updating cell states by selectively adding or deleting text data.

The operation of these gates involves a series of matrix operations, each with its independent weight selection, akin to a gate mechanism in traditional computing. One of the most significant advantages of LSTM is their ability to circumvent the problem of long-term dependencies. Traditional neural networks often struggle with retaining information over extended sequences, but LSTM overcome this hurdle. They employ gates to manage the network's memory effectively, ensuring that vital information is retained and irrelevant data is discarded, thus maintaining the integrity and relevance of the emotional cues encoded within the text.

The LSTM Network Mechanics

The initial phase in constructing an LSTM network involves identifying the textual data elements that are dispensable. This decision-making process employs the sigmoid function and hinges on the output of the preceding LSTM unit, g_{s-1} , at the $(s-1)th$ position and the current input W_s . The sigmoid function is pivotal in determining which segments of the previous output should be omitted. This gate, termed the forget gate (or u_i), generates a vector with each element ranging between 0 and 1, corresponding to

the respective elements in the cell state h_{i-1} Equation (4. 1):

$$u_s = \sigma(Y_u[h_{i-1}, V_i] + b_u) \quad (4. 1)$$

, σ denotes the sigmoid function. The matrices Y_u and V_i , along with the bias b_u , are integral to the forget gate's operation. Subsequently, the network decides on retaining and updating the text data within the new input V_i , influencing the modification and evolution of the cell state. This process involves both the sigmoid and tanh layers. The sigmoid layer initially determines the relevancy of new information for updating (0 or 1). The cell state is updated by multiplying these values (see Equation (4. 2)–4. 3 and (4. 4)):

$$q_i = \sigma(Y_q[h_{i-1}, V_i] + b_q) \quad (4. 2)$$

$$N_i = \tanh(Y_n[h_{i-1}, V_i] + b_n) \quad (4. 3)$$

$$C_i = C_{i-1} \cdot u_i + N_i \cdot q_i \quad (4. 4)$$

The variables C_{i-1} and C_i represent the previous and current cell states, respectively. The ultimate output kt is influenced by the cell states' output condition O_i Equation (4. 5), processed through a sigmoid layer that filters the essential components for output. The final cell state C_i undergoes a transformation through the tanh function, ensuring the output values range between -1 and 1 Equation (4. 6).

$$O_i = \sigma(Y_o[h_{i-1}, V_i] + b_o) \quad (4. 5)$$

$$h_i = O_i \cdot \tanh(c_i) \quad (4. 6)$$

Audio Speaker Embedding Using DNN

DNNs algorithms have been extensively utilized in both industrial applications and research in recent years, demonstrating a high success rate in recognition tasks. DNNs are particularly effective in multi-label classification problems. This study focuses on distinguishing between two distinct groups: the swing and stance phases of gait. The dataset for training encompasses five parameters related to gait. A typical DNNs architecture comprises an input layer, an output layer, and at least two hidden layers. Key components of the network include the activation function, learning rate, gradient descent optimizer, and the stages of training. The optimizer refines the process of gradient descent. One notable aspect of DNNs is their implementation of the softmax function, as shown in Equation (4. 7), which normalizes neural outputs to the $(0, 1)$ interval, making it highly useful for multi-classification tasks. In these tasks, the model's prediction target is assigned to the output node with the highest probability:

$$r_h = \frac{d^h}{\sum_{u=1}^b d^u} \quad (4. 7)$$

The execution time of the method is linearly positively correlated with the encoding dimension, which in turn is influenced by the number of deployment units in a DNNs. Following a layer split, the deployment units in a DNNs are reduced, leading to a potential increase in execution time and a decrease in the encoding dimension, especially when considering offloading issues.

Video Speaker Embedding Using Deep Learning architectures

Inspired by neuronal structures in the cerebral cortex, DL architectures have demonstrated efficacy across a variety of machine learning applications. These architectures operate as feed-forward networks, enabling unidirectional information propagation from input to output layers. The design and functionality of these systems draw heavily from biological processes, particularly emulating the hierarchical organization of the visual cortex composed of simple and complex cell layers. Despite variations in specific architectural designs, a common framework exists, comprising a sequence of convolutional and pooling layers, followed by densely connected layers. This design is reminiscent of traditional neural networks. The deep structure of these networks is constructed by layering these elements, often including input, convolutional, pooling, and flattening layers, as depicted in Figure 4.5. [331]. This hierarchical integration of data through these layers culminates in a dense representation, ultimately leading to an output layer that signifies the final processing outcome.

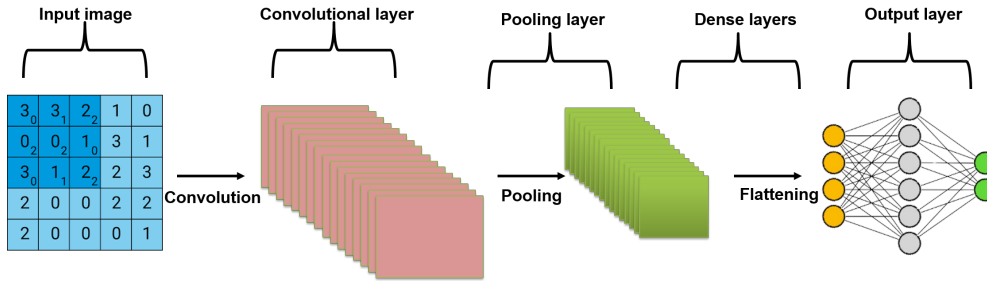


Figure 4.5. Architecture of CNN method.

A CNNs or ConvNet is a deep learning architecture designed to learn directly from data. CNNs are specifically well-suited for identifying patterns in images to recognize objects, classes, and categories. Moreover, they prove to be highly effective in classifying audio, time series, and signal data.

In dealing with large volumes of training data, CNNs often employs a mini-batch approach. Consider a set of activation features $X = [x_1, \dots, x_M]$ from M training images, where each feature vector x_i has a dimensionality of K , and $[x_1, \dots, x_M] \in \{-1, 1\}^M$ stores the ground truth labels. A robust classifier employs a boosting technique to make predictions. The classifier $H(\cdot)$ is a weighted sum of weak classifiers $h(\cdot)$, formulated as

Equation (4. 8):

$$H(x_1) = \sum_{j=1}^K \alpha_j h(x_{ij}, \lambda_j); \quad (4. 8)$$

$$h(x_{ij}, \lambda_j) = \frac{f(x_{ij}, \lambda_j)}{\sqrt{f(x_{ij}, \lambda_j)^2 + \eta^2}}; \quad (4. 9)$$

where $x_{ij} \in x_j$ is the j th activation feature of the i th image. Each feature $h(x_{ij}, \lambda_j)$ is associated with a potential weak classifier, producing outputs in the range $(-1, 1)$ Equation (4. 9).

The function $\frac{f(\cdot)}{\sqrt{f(\cdot)^2 + \eta^2}}$ simulates a sign method, optimizing gradient descent through derivative computation. In this framework, $h(x_{ij}, \lambda_j) \in K$ represents a decision stump, a one-level decision tree. The parameter η , adjustable by the distribution of $f(\cdot)$ as σ/c , controls the function's ramp. Here, σ is the standard deviation of $f(\cdot)$, and c is a constant. In this work, η is empirically set to σ/v . The weight of the i th weak classifier $\alpha_j \geq 0$ follows the constraint Equation (4. 10):

$$\sum_{j=1}^K \alpha_j = 1; \quad (4. 10)$$

When $\alpha_j = 0$, the corresponding neuron remains inactive in the neural network. The balance between the losses of strong and weak classifiers is given by Equation (4. 11):

$$\epsilon^B = \beta \epsilon_{strong}^B + (1 - \beta) \epsilon_{weak}; \quad (4. 11)$$

where $\beta \in [0, 1]$. The losses for strong and weak classifiers are defined as Equations (4. 12) and (4. 13):

$$\epsilon_{strong}^B = \frac{1}{N} \sum_{i=1}^N (H(x_i) - x_i)^2; \quad (4. 12)$$

$$\epsilon_{weak} = \frac{1}{MN} \sum_{i=1}^N \sum_{1 \leq j \leq K, \alpha_j=0} [h(x_{ij}, \lambda_j) - y_i]^2; \quad (4. 13)$$

where $\alpha_j > 0$ ensures that dormant neurons are excluded from the loss calculation.

4.2.3. Classification Using CA-WGNN

The Cumulative Attribute-Weighted Graph Neural Network (CA-WGNN) model, which has been used in geometric deep learning, uses a weighted graph structure to map interrelationships between modalities. This groundbreaking approach not only fosters a

comprehensive understanding of emotions by tapping into the synergistic potential of diverse data sources but also strives to classify emotions holistically. In the realm of emotion classification, it is paramount to adopt a holistic approach that integrates insights from multiple modalities. This method is especially effective in capturing the nuanced spectrum of human emotions. Specifically, the classification model distinguishes between three primary categories of emotions:

1. **Positive Emotions:** These include feelings such as joy, excitement, and a general sense of positivity. These emotions are characterized by their uplifting and affirmative nature.
2. **Negative Emotions:** This category encompasses emotions like sadness, fear, and disgust, which are typically associated with adverse experiences or perceptions.
3. **Neutral Emotion:** Representing a state that is neither explicitly positive nor negative, this category captures the subtlety of emotions that do not fall distinctly into the other two classifications.

The sophisticated structure of CA-WGNN, which is connectivity-driven, excels in learning from input graphs and deciphering complex system interactions through iterative data analysis across connections. This enables the model to generalize effectively to unfamiliar graphs, making predictions at both the network and node levels.

Thanks to its advanced capabilities, CA-WGNN is instrumental in a variety of applications, showcasing its versatility and effectiveness. In terms of implementation, the pseudo-code for CA-WGNN inference, as illustrated in Algorithm 1, delineates the step-by-step process of this model's operation. Moreover, the model's detailed structure, as shown in Figure 4.6., further explains its functional dynamics and connectivity-driven approach, highlighting its role as a pivotal tool in the realm of geometric deep learning.

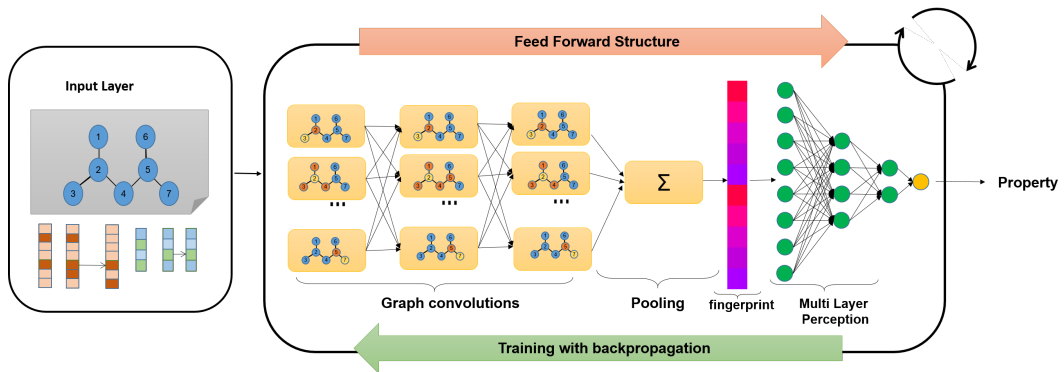


Figure 4.6. Structure of CA-WGNN.

Algorithm 1 Algorithm of CA-WGNN for emotion recognition

```
1: procedure GNN OPERATOR
2:    $F \leftarrow$  Number of layers in the GNN
3:    $C \leftarrow$  Set of nodes in graph  $S$  (assumed static)
4:    $A \leftarrow$  Set of edges in graph  $S$  (assumed static)
5:   Initialize Nodes and Edges:
6:   for  $c \in C$  do
7:      $z_c^0 \leftarrow [y_c, 0, \dots, 0]$ 
8:   end for
9:   for  $a \in A$  do
10:     $s_a^0 \leftarrow [h_c, 0, \dots, 0]$ 
11:  end for
12:  GNN Layered processing:
13:  for  $f = 1$  to  $F$  do
14:    Edge processing: (Order of edge and node processing may be interchanged or even interspersed.)
15:    for  $a \in A$  do (Order of aggregation and combination may be interchanged if aggregation is linear.)
16:       $p_a^{(f)} = \rho_A^{(f)}(s_a^{(f-1)}, z_w^{(f-1)} : w \in M(a))$ 
17:       $s_a^{(f)} = \phi_A^{(f)}(p_a^{(f)})$ 
18:    end for
19:    Node processing: (Order of edge and node processing may be interchanged or even interspersed.)
20:    for  $c \in C$  do (The order of aggregation and combination may be interchanged if aggregation is linear.)
21:       $e_c^{(f)} = \rho_C^{(f)}(z_c^{(f-1)}, z_w^{(f-1)} : w \in M(c))$ 
22:       $z_c^{(f)} = \phi_C^{(f)}(e_c^{(f)})$ 
23:    end for
24:  end for
25:  Readout:
26:   $\hat{x} = \phi_S(\rho_S(\{z_c^F, s_a^F : c, a \in S\}))$ 
27: end procedure
```

The CA-WGNN model starts with the graph's edges, vertices, and feature vectors as inputs (step 9). It executes in layers, each updating the feature vectors of all connected vertices and edges concurrently (steps 11, 14, 15). Combination functions, represented by neural networks, are used to update the combined edges and vertices (steps 13, 17). The readout process is carried out by an appropriate function, often a neural network (step 18).

The following equations explain the aggregation Equation (4. 14) and combination processes Equation (4. 15) at any given layer $f \in [1, F]$:

$$\text{Aggregation: } P_A^{(f)} = \rho_A^{(f)}(s_a^{(f-1)}, s_w^{(f-1)} : w \in M(a)), \quad (4. 14)$$

$$\text{Combination: } S_A^{(f)} = \phi_A^{(f)}(p_a^{(f)}), \quad (4. 15)$$

where for each layer $f - 1$, the aggregated edge ρ_A includes the edge's feature vector Equation (4. 16) s_a and the terminal vertex feature vectors s_w , where $w \in M(a)$. The combination Equation (4. 17) ϕ_A uses this aggregation as input. Similarly, for nodes:

$$\text{Aggregation: } e_c^{(f)} = \rho_C^{(f)}(z_c^{(f-1)}, z_c^{(f-1)} : w \in M(c)). \quad (4. 16)$$

$$\text{Combination: } z_c^{(f)} = \rho_C^{(f)}(e_c^{(f)}), \quad (4. 17)$$

The final output feature vector $\hat{x}$ is obtained by applying a readout function Equation (4. 18):

$$\text{Readout: } \hat{x} = \phi_S(\rho_S(z_c^F, s_a^F : c, a \in S)) \quad (4. 18)$$

In the domain of speaker characteristic extraction, three distinct methodologies are employed: textual speaker embedding via LSTM networks, acoustic speaker embedding through DNNs, and visual speaker embedding utilizing s CNNs. The process of emotion recognition benefits substantially from the amalgamation of these diverse modalities, facilitated by a Cross-Attention Weighted Graph Neural Network CA-WGNN. The DNNs effectively models acoustic patterns derived from audio inputs, whereas the LSTM network is adept at discerning temporal relationships in textual data. Concurrently, the CNNs excels at extracting spatial features from video frames. Fusion of these modalities is achieved through the construction of a weighted graph, where the edges represent inter-modal correlations and weights are assigned based on the relative importance of each modality. The CA-WGNN leverages these weights to integrate data, enabling the detection of nuanced emotional cues.

This integrative approach fosters a robust and versatile framework for emotion recognition across multiple modalities, thereby enhancing classification accuracy by harnessing complementary information from varied sources. Table 4.1. delineates the performance of Emotion Recognition Models on the CMU-MOSI and IEMOCAP datasets, detailing their respective experimental configurations.

Table 4.1. Comparison of Experimental Configurations

Dataset Name	CMU-MOSI	IEMOCAP
Model Name	CA-WGNN	CA-WGNN
No. of Layers	7	7
Batch Size	32	32
Epochs	100	100
Random State	42	42
Learning Rate	0.001	0.001
Validation Split	0.2	0.5

4.2.4. Problem Statement

Robust and automated recognition of facial emotion is one of the stated problems. Other issues involve completing meaningful analysis of the taken image based on facial emotions, constructing test and training datasets, and ultimately, developing and implementing fitted classifiers to learn the underlying classifiers' vectors of facial descriptors and deliver a model design that can identify up to six models that are widely accepted across different cultures. Emotions that predominate include dread, joy, sorrow, surprise, disgust, and delight. If the system were in place, it could analyze a person's face and its features to determine their identification based on a weighted assumption. Accurate Sentiment Recognition and identification are made possible by our proposed CA-WGNN method, which combines facial expression analysis, culturally adaptive classifiers, and diverse dataset creation. It is achieved by using cumulative attribute weights, which provide nuanced and cross-cultural accuracy. When faced with the difficult task of automated and robust face emotion identification, the CA-WGNN is an essential tool. When it comes to analyzing the expressed emotions in photographs, CA-WGNN is crucial for solving the given challenge. Its relevance goes beyond only building fitted classifiers to include building complete training and test datasets. For a model to be able to recognize six commonly recognized emotions dread, pleasure, sadness, surprise, disgust, and delight across different cultures, this sophisticated neural network is crucial for learning the complex vectors of face descriptors. The proposed system is enhanced using CA-WGNN so that it can evaluate face characteristics and make reliable identifications using weighted assumptions.

4.2.5. Assessment Metrics for Classifier Performance

The performance of classifiers is evaluated using key metrics: Accuracy, Macro Precision, Macro Recall, and the F1 score. These metrics are defined as follows:

- **Accuracy:** This metric quantifies the proportion of correctly predicted instances in

the dataset and is calculated using the Formula (4. 19):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total}} \quad (4. 19)$$

TP (True Positives) and TN (True Negatives) are the numbers of correctly identified positive and negative instances, respectively, and Total represents the sum of all instances.

- **Precision:** Defined as the ratio of correctly predicted positive instances to the total predicted as positive, Precision indicates the model's ability to avoid false positives as (4. 20):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4. 20)$$

FP (False Positives) denotes instances incorrectly classified as positive.

- **Recall:** Recall, or sensitivity, measures the proportion of actual positives that are correctly identified and is given by (4. 21):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4. 21)$$

FN (False Negatives) are actual positives that were incorrectly classified as negative.

- **F1 Score:** The F1 score is the harmonic mean of Precision and Recall, providing a balance between them, especially important in the presence of class imbalances as (4. 22):

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4. 22)$$

This process involves collecting and merging feature vectors from every vertex and edge in the graph at the final iteration F . In this study, distinct methodologies were applied to each dataset to implement our algorithm. For audio data, DNNs is utilized to generate speaker embeddings. This involves extracting features from audio files using the librosa library and then processing them through a multi-layer DNNs to obtain embeddings. In contrast, for video data, a CNNs is employed for a similar purpose. This process includes the extraction and preprocessing of frames from video files, which are then input into the CNNs to derive embeddings. This research's core aspect lies in classifying these embeddings from each modality. In all three datasets text, audio, and video embeddings are separated into features (X) and target values (y). These are then input into a specially designed Cumulative Attribute Weighted Graph Neural Network (CA-WGNN) model. The model's performance is evaluated using metrics such as accuracy, precision, recall, and F1 score. These metrics collectively offer a comprehensive view of the classifier's performance, taking into account both error types and class distribution in the dataset.

4.3. Experimental Results and Evaluation

In this study, the experimental framework was meticulously designed to accommodate the rigors of machine learning applications, leveraging a harmonious blend of hardware and software configurations. Central to this framework was a Windows 10 operating system, chosen for its robustness and compatibility with advanced computational tasks. The experimental environment encompassed a dual-platform approach, utilizing both Google Colab and Jupyter Notebook. This methodology facilitated a comprehensive evaluation, allowing for a detailed analysis of various challenges, as enumerated in the accompanying Table 4.2.. A notable aspect of this setup was the dependency of Google Colab’s time efficiency on internet connectivity and hardware specifications. From a hardware perspective, the experiments were conducted on a system equipped with a Windows 10 64-bit operating system, complemented by 200 GB of storage in the C drive and 8 GB of RAM. This configuration ensured a seamless operational flow during the testing phases. Python, executed via Jupyter Notebook, was the primary programming language employed, underscoring its versatility and efficiency in handling complex datasets.

Integral to our research was the application of GNNs for sentiment analysis, targeting multimodal data encompassing text, audio, and video. This innovative approach, utilizing GNNs, was pivotal in analyzing the interconnected elements within these data types, thereby enhancing the accuracy and depth of sentiment analysis. The incorporation of GNNs allowed for a more nuanced understanding of the sentiment conveyed across different modalities, significantly augmenting the scope of our study in the domain of TER.

The CA-WGNN method is introduced for emotion recognition, leveraging trimodal inputs, including text, audio, and video modalities. This approach aims to enhance the interaction among emotional inputs, thereby fostering a more sophisticated and efficient system. The CA-WGNN is evaluated against established methodologies, such as the SVM [332], the Voting Classifier (integrating Logistic Regression and Stochastic Gradient Descent, denoted as VC(LR-SGD) [332], and the GBM [332]. Metrics, including accuracy, precision, recall, and the F1-score, are employed to determine the efficacy of inaccurate emotion categorization models. These metrics are essential for assessing the models’ robustness across various applications, confirming their reliability and capability to discern complex emotional nuances.

In their comprehensive study, Yousaf A. et al. [332] conducted a thorough evaluation of the Support Vector Machine (SVM) alongside six other prominent machine learning algorithms: Logistic Regression, Stochastic Gradient Descent, Naive Bayes, Decision Tree, Random Forest, and Voting Classifier, in recognizing emotions through textual tweet classification. This investigation focused on assessing and comparing the performance metrics of these models. Notably, a linear kernel function was employed for the SVM, indicating that the decision boundary is linear. The transformation of tweets into

numerical vectors was facilitated through the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization method, which emphasizes the significance and frequency of each word within the tweets. The study adhered to the default parameters for various SVM configurations, including the regularization parameter 'C', which balances the margin width and classification error, the kernel coefficient 'gamma', which affects the influence of individual training samples, the tolerance for the stopping criterion 'tol', and the maximum number of iterations 'maxiter.'

The research findings demonstrated that SVM was significantly effective, achieving an accuracy of 77% and an F1-score of 79% in a binary dataset that categorized tweets as happy or unhappy. Furthermore, in a multi-class dataset, where tweets were classified as happy, sad, angry, or neutral, SVM attained an accuracy of 67% and an F1 score of 64%. This in-depth analysis underscores the SVM's capability to efficiently recognize emotions from textual content on social media platforms, particularly Twitter, providing valuable contributions to sentiment analysis and emotion detection.

Table 4.2. Hardware and software configuration for system requirements

Component	Specification
RAM	8 GB
Hard Disk	200 GB in C Drive
OS	Windows 10 64 bit
Software	Individual Edition-Anaconda Jupyter Notebook
Source	https://www.anaconda.com/products/individual (accessed on 26 February 2024) https://docs.anaconda.com/anaconda/install/windows/ (accessed on 26 February 2024) https://jupyter.org (accessed on 26 February 2024)
Total Processing Time	(h, m, s): 00, 21, 31

4.3.1. Accuracy Analysis

Among the multiple factors considered in emotion detection, accuracy stands as a paramount measure, particularly when comparing the performance of models across multimodal datasets like IEMOCAP Table 4.3. and CMU-MOSI Table 4.4.. This comparative study focuses on the accuracy in recognizing emotions through text, audio, and video modalities. It delves into the model's proficiency in identifying and classifying emotional expressions accurately across these varied channels. The essence of this analysis is to gauge how effectively the trimodal method captures and aligns with the diverse emotional nuances present in both datasets. Such an evaluation provides pivotal insights into the model's overall capability to decipher emotions, whether conveyed through textual, auditory, or visual means.

Table 4.3. Performance Metrics for IEMOCAP Dataset

Dataset	Category	Precision (%)	Recall (%)	F1-Score (%)	Support (%)
IEMOCAP Text	Negative	0.91	0.90	0.91	69
	Neutral	0.96	0.96	0.96	68
	Positive	0.92	0.94	0.93	63
IEMOCAP Audio	Negative	0.91	0.91	0.91	69
	Neutral	0.94	0.96	0.95	68
	Positive	0.94	0.92	0.93	63
IEMOCAP Video	Negative	0.89	0.93	0.91	69
	Neutral	0.96	0.94	0.95	68
	Positive	0.93	0.90	0.92	63
Fusion Overall Accuracy					0.93 (200)
Fusion Macro Average		0.93	0.92	0.93	200
Fusion Weighted Average		0.93	0.93	0.93	200

Table 4.4. Performance Metrics for CMU-MOSI Datasets

Dataset	Category	Precision (%)	Recall (%)	F1-Score (%)	Support (%)
CMU-MOSI Text	Negative	0.88	0.94	0.91	69
	Neutral	0.97	0.90	0.93	68
	Positive	0.90	0.90	0.90	63
CMU-MOSI Audio	Negative	0.90	0.93	0.91	69
	Neutral	0.98	0.94	0.96	68
	Positive	0.91	0.92	0.91	63
CMU-MOSI Video	Negative	0.88	0.96	0.92	69
	Neutral	1.00	0.94	0.97	68
	Positive	0.92	0.89	0.90	63
Fusion Overall Accuracy		0.94 (200)			
Fusion Macro Average		0.94	0.94	0.94	200
Fusion Weighted Average		0.94	0.94	0.94	200

The results presented in the following sections compare the accuracy of the IEMOCAP and CMU-MOSI datasets, based on the criteria of Sentiment Recognition. The analysis revealed a marginally higher accuracy in the CMU-MOSI dataset (0.94%) compared to IEMOCAP (0.93%), indicating nuanced differences in the performance of these models in emotion detection across the three modalities. This comparison not only highlights the strengths and limitations of each dataset but also underscores the importance of multimodal approaches in enhancing the accuracy of sentiment recognition systems.

The provided table presents a comprehensive overview of the performance metrics for the CMU-MOSI dataset, categorized into three modalities: Text, Video, and Audio. Each modality is further subdivided into Negative, Neutral, and Positive sentiment categories. The table displays essential evaluation metrics, including Precision, Recall, and F1-Score, which are crucial for assessing the effectiveness of a sentiment classification model.

Notably, the model’s overall accuracy across all modalities is 94%

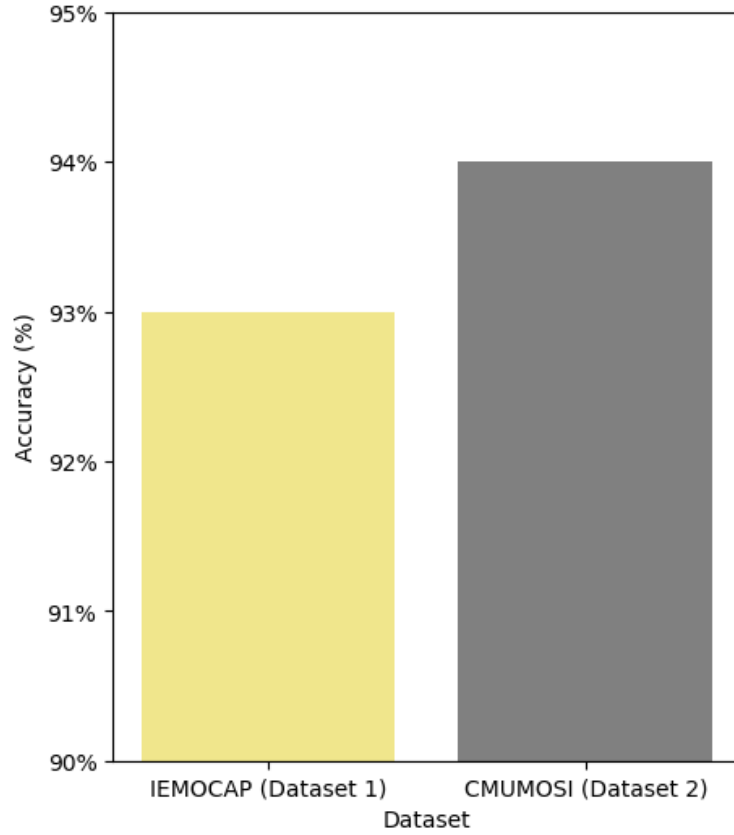


Figure 4.7. The accuracy of the IEMOCAP and CMU-MOSI datasets.

Following the analytical discourse, we present Table 4.5., which succinctly compares the accuracy percentages obtained from the IEMOCAP and CMU-MOSI datasets. This table serves as a quantitative testament to the findings discussed above, illustrating the slight yet significant variance in the accuracy of Sentiment Recognition between the two datasets.

Table 4.5. Comparison of accuracy between the IEMOCAP and CMU-MOSI datasets.

Dataset	Accuracy (%)
IEMOCAP (Dataset 1)	0.93
CMU-MOSI (Dataset 2)	0.94

The data presented in Table 4.5. provides a clear and concise reference for understanding the efficacy of each dataset in the context of TER. It reinforces the notion that while both datasets exhibit high levels of accuracy, subtle distinctions in their performance can be critical in selecting the appropriate dataset for specific research or application needs.

Across the field of emotion recognition, accurately determining the effectiveness of various computational methods is crucial for advancing the field. Table 4.6. and Figure 4.8. present the accuracy of both the proposed and existing methods. The CA-WGNN achieved 94%, compared with the existing methods of SVM, which attained 76%,GBM attained 74%, and VC(LR-SGD)attained 79%. It demonstrates that the suggested process achieves higher rates than existing methods, underscoring the superiority of the CA-WGNN approach in identifying emotions with greater accuracy. This notable improvement highlights the CA-WGNN in enhancing the accuracy and reliability of emotion recognition systems, marking a significant step forward in the field.

Table 4.6. Numerical outcomes of Accuracy.

Methods	Accuracy (%)
SVM [332]	76
GBM [332]	74
VC(LR-SGD) [332]	79
CA-WGNN [Proposed]	94

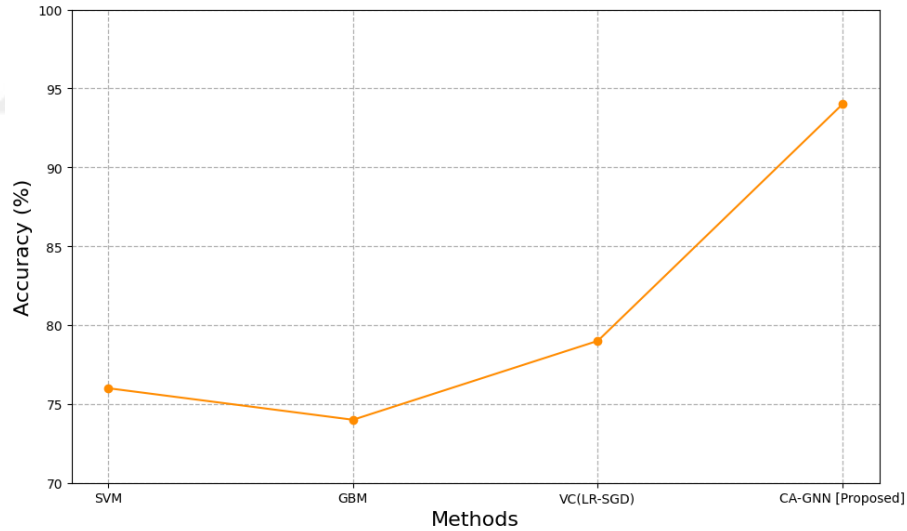


Figure 4.8. Graphical representation of Accuracy.

4.3.2. Precision Analysis

Delving into the intricacies of trimodal emotion detection, this study extends its focus to the precision aspect in the evaluation of text, audio, and video components within the IEMOCAP and CMU-MOSI datasets. Precision in this context is a measure of how accurately the trimodal emotion identification model discerns and categorizes emotions, taking into account the variances in emotional expressions manifesting through visual, auditory, and textual forms.

A crucial part of this comparative analysis involves assessing the diversity, quality, and annotation methodologies of the datasets. These elements are instrumental in understanding the nuances and the resulting precision in sentiment recognition. The detailed comparison of precision levels achieved by the models using IEMOCAP and CMU-MOSI datasets is systematically presented in Table 4.7. and Figure 4.9.. The findings indicate a marginally superior precision in the CMU-MOSI dataset (0.94%) as compared to IEMOCAP (0.93%).

Table 4.7. The comparison of precision between the IEMOCAP and CMU-MOSI datasets.

Dataset	Precision (%)
IEMOCAP (Dataset 1)	0.93
CMU-MOSI (Dataset 2)	0.94

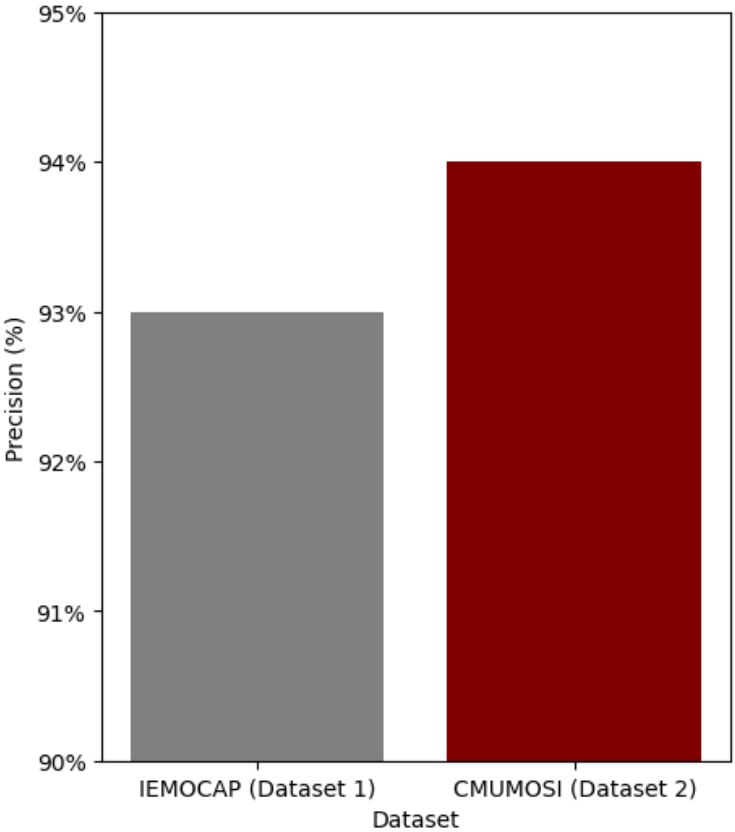


Figure 4.9. The comparison of precision between the IEMOCAP and CMU-MOSI datasets.

This precision comparison not only highlights the subtle yet significant differences in the datasets but also underscores the importance of detailed and nuanced evaluation in TER systems. An analysis such as this is vital to developing more accurate and reliable sentiment recognition models.

The comparative analysis of precision metrics significantly highlights the effectiveness of different computational approaches in emotion recognition. In Table 4.8. and Figure 4.10. we present the precision for both the proposed and existing methods. The CA-WGNN achieved a remarkable 92%, in comparison with the existing methods of SVM, which attained 76%,GBM attained 72%, and VC(LR-SGD)attained 78%. This demonstrates that the suggested process achieves higher precision rates than the existing methods, underlining the superiority of the CA-WGNN approach in accurately categorizing emotional states. The results reinforce the notion that leveraging advanced algorithms like CA-WGNN significantly enhances the precision of emotion recognition, setting a new benchmark for future research in this domain.

Table 4.8. Numerical Outcomes of Precision.

Methods	Precision (%)
SVM [332]	76
GBM [332]	72
VC(LR-SGD) [332]	78
CA-WGNN [Proposed]	92

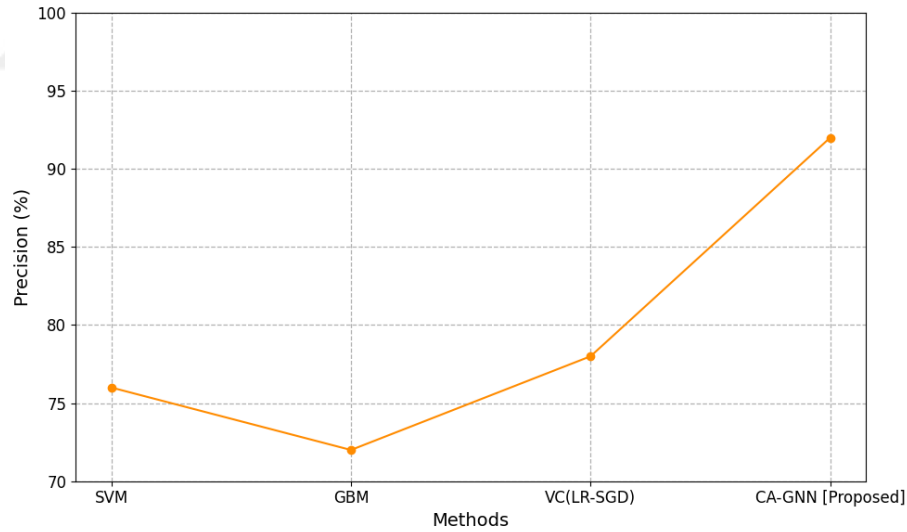


Figure 4.10. Graphical representation of Precision.

4.3.3. Recall Analysis

In the intricate landscape of Sentiment Recognition, recall emerges as a critical metric, reflecting the model's capacity to identify a wide range of emotions accurately. This study delves into a comparative analysis of recall within the IEMOCAP and CMU-MOSI datasets, each known for their unique strengths and focuses. IEMOCAP excels in capturing a broad spectrum of realistic emotional expressions in audiovisual

data, whereas CMU-MOSI distinguishes itself with an emphasis on multilingual Sentiment Recognition. The application of a tri-modal emotion identification approach in this study is pivotal. It facilitates a more comprehensive and context-aware categorization of emotions, adaptable to a variety of situations. This methodology enhances the depth and breadth of Sentiment Recognition, moving beyond conventional single-modal approaches. Table 4.9. and Figure 4.11. present a side-by-side comparison of the recall capabilities of models trained on these datasets. The results indicate a slightly higher recall for the CMU-MOSI dataset (0.94%) compared to IEMOCAP (0.93%), underscoring the nuanced differences in the datasets' abilities to accurately recognize a diverse range of emotions.

Table 4.9. Comparison of Recall.

Dataset	Recall (%)
IEMOCAP (Dataset 1)	0.93
CMU-MOSI (Dataset 2)	0.94

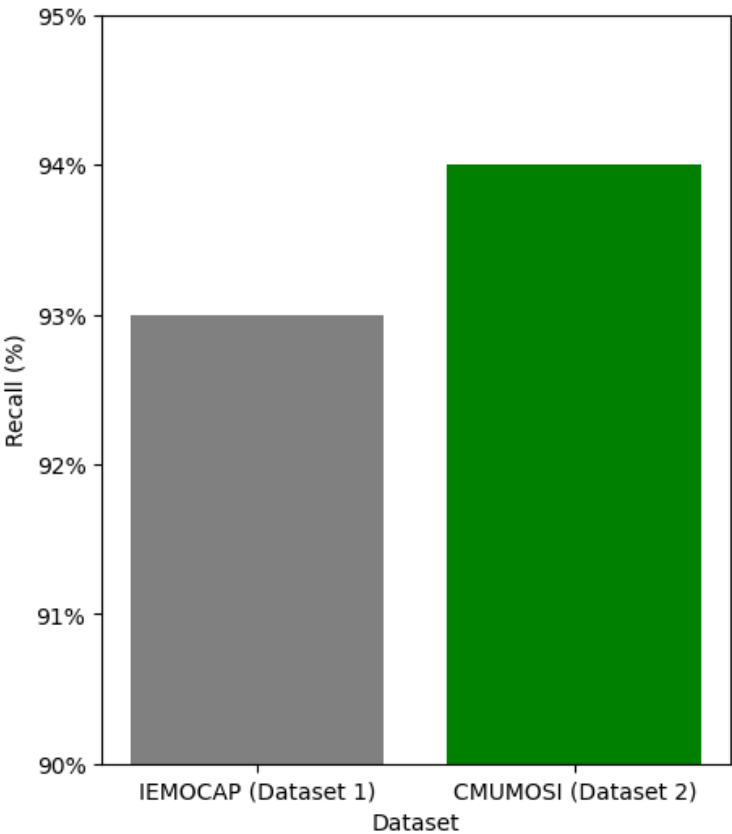


Figure 4.11. The recall comparison between the IEMOCAP and CMU-MOSI datasets.

This comparison not only highlights the specific strengths of each dataset but also emphasizes the importance of employing multimodal Sentiment Recognition frameworks

for more accurate and versatile ER systems. The findings bolster the case for the ongoing refinement and enhancement of these technologies in the field of ER research.

The evaluation of precision across various methods is critical in understanding their effectiveness in emotion recognition. Table 4.10. and Figure 4.12. present the precision metrics for both the proposed and existing methods. The CA-WGNN method demonstrates superior performance by achieving a 93% precision rate, in comparison with the existing methods where SVM attained 80%, GBM attained 79%, and VC(LR-SGD) attained 84%. This comparison clearly illustrates that the suggested CA-WGNN process outperforms the existing methods in terms of precision. The findings underscore the CA-WGNN's advanced capability in accurately identifying emotional nuances, setting a new benchmark for future research in the domain of emotion recognition.

Table 4.10. Numerical Outcomes of Recall.

Methods	Recall (%)
SVM [332]	80
GBM [332]	79
VC(LR-SGD) [332]	84
CA-GNN [Proposed]	93

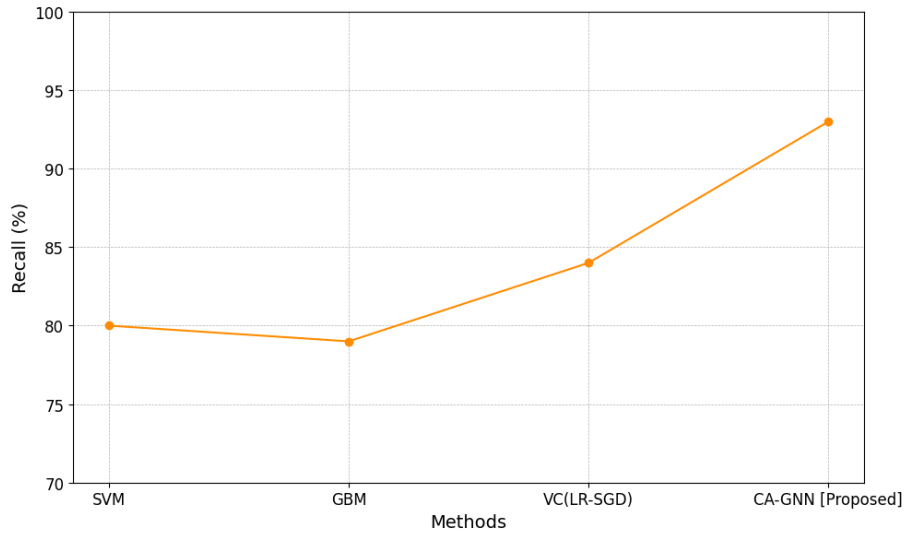


Figure 4.12. Graphical representation of Recall.

4.3.4. F1-Score Analysis

The F1-score, a crucial statistic in Sentiment Recognition systems, serves as a harmonious balance between precision and recall, encapsulating the overall accuracy of the system. In the context of trimodal emotion detection, the F1-score acquires an elevated

significance as it reflects the system’s ability to effectively integrate and interpret data across text, audio, and video modalities.

In this study, the robustness and adaptability of the trimodal emotion detection technique are scrutinized by comparing the F1-scores obtained from the IEMOCAP and CMU-MOSI datasets. Achieving high F1 scores is indicative of a model’s precision in identifying emotions, thereby underscoring the system’s efficiency across diverse datasets and modalities. This is crucial for gaining a comprehensive understanding of emotional expressions.

Table 4.11. and Figure 4.13. present a comparative analysis of the F1-scores for the IEMOCAP and CMU-MOSI datasets, based on their performance in ER. The data reveals a marginally higher F1-score for the CMU-MOSI dataset (0.94%) as compared to that of IEMOCAP (0.93%), highlighting the nuanced efficacy of the models in processing emotional data.

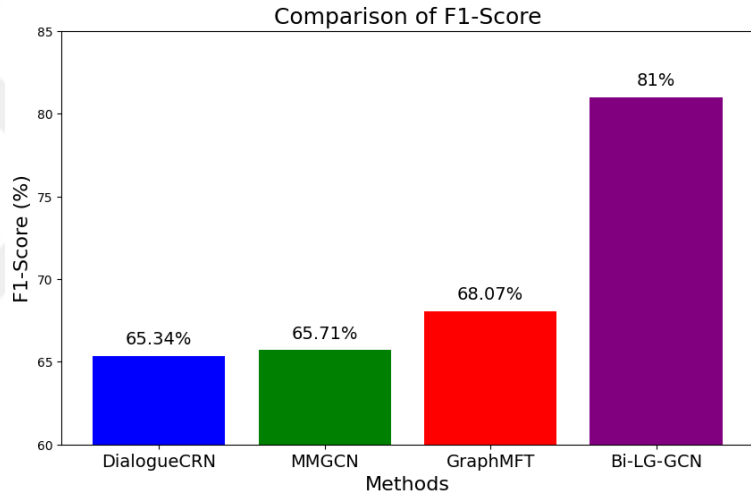


Figure 4.13. The F1-scores comparison between the IEMOCAP and CMU-MOSI datasets.

Table 4.11. Comparison of F1-score.

Dataset	F1-Score (%)
IEMOCAP (Dataset 1)	0.93
CMU-MOSI (Dataset 2)	0.94

This comparative exploration of F1 scores not only sheds light on the performance metrics of the datasets but also provides invaluable insights for future advancements in TERsystems, aiming for models that are both accurate and versatile.

The meticulous comparison of precision metrics underscores the advancements and effectiveness of emotion recognition technologies. In Table 4.12. and Figure 4.14., we detail the precision rates for both the proposed and existing methods. The CA-WGNN method

stands out by achieving a remarkable 91% precision, in contrast to the existing methods where SVM and GBM each attained 76%, and VC(LR-SGD) reached 81%. This comparison vividly illustrates the superior performance of the proposed CA-WGNN method, affirming its capacity to achieve higher precision rates than the conventional approaches. The results highlight the efficiency of the CA-WGNN and pave the way for future innovations in the field of emotion recognition, emphasizing the potential for further advancements and the continuous need for research and development in this dynamic area.

Table 4.12. Numerical Outcomes of F1-score.

Methods	F1-Score (%)
SVM [332]	78
GBM [332]	76
VC(LR-SGD) [332]	81
CA-GNN [Proposed]	91

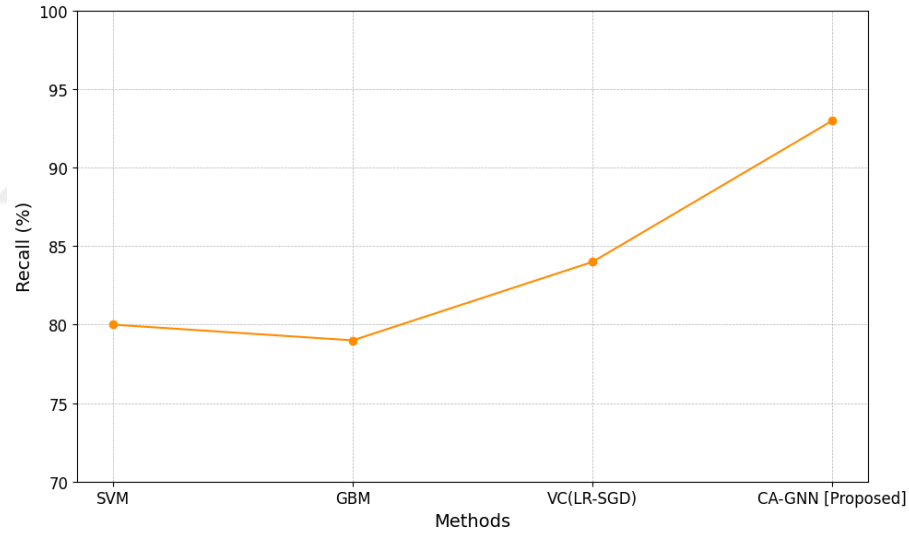


Figure 4.14. Graphical representation of F1-score.

4.3.5. Advancements Brought by CA-WGNN

The diagram depicts the architecture of the CA-WGNN model, highlighting the sequential data flow through various layers, each contributing significantly to the model's classification capabilities.

- **Cumulative Attribute Weighting:** The CA-WGNN model introduces a novel cumulative weighting approach for attributes within a graph structure. This technique is instrumental in deepening the understanding of node interrelations and dependencies within the graph.

- **Enhanced Contextual Understanding:** By focusing on cumulative attributes, CA-WGNNs achieve a heightened comprehension of the broader context surrounding each node. This aspect is crucial for making accurate predictions in complex graph structures.
- **Applicability in Complex Network Analysis:** CA-WGNNs are particularly advantageous in scenarios involving intricate graph structures. In such contexts, a node's influence often extends beyond its immediate neighbors, necessitating a more sophisticated analysis approach.

The effectiveness of the CA-WGNN model is further evidenced by its performance metrics on the IEMOCAP dataset, as shown in the Table 4.13.. The model demonstrates high precision, recall, and F1-score across different categories, underscoring its robustness and accuracy in sentiment analysis. Notably, the model achieves an overall accuracy of 94%, with a balanced performance across all categories, as indicated by the macro and weighted averages. These results highlight the model's capability in accurately classifying sentiments, even in the challenging context of multimodal datasets.

Table 4.13. Performance metrics of the CA-WGNN model on the IEMOCAP dataset.

Category	Precision	Recall	F1-Score	Support
Negative	0.86	0.93	0.90	69
Neutral	0.96	0.97	0.96	68
Positive	0.96	0.87	0.92	63
Accuracy	0.93 (200)			
Macro Average		0.93	0.92	0.93
Weighted Average		0.93	0.93	0.93

The CA-WGNN model's performance was rigorously assessed using the CMU-MOSI dataset Table 4.14., known for its comprehensive and multimodal nature. This evaluation aimed to ascertain the model's proficiency in accurately analyzing and classifying sentiments across diverse categories. The table below presents a detailed breakdown of the model's performance metrics, including precision, recall, and F1-score for Negative, Neutral, and Positive sentiment categories, along with the overall accuracy, macro, and weighted averages.

Table 4.14. Performance Metrics

Category	Precision (%)	Recall (%)	F1-Score (%)	Support (%)
Negative	0.93	0.94	0.94	69
Neutral	0.97	0.94	0.96	68
Positive	0.92	0.94	0.93	63
Accuracy	0.94 (200)			
Macro Average		0.94	0.94	0.94
Weighted Average		0.94	0.94	0.94

The data in the table reveal the CA-WGNN model’s exceptional accuracy and consistency in sentiment classification. The model demonstrates high precision and recall rates across all categories, a testament to its robust algorithmic structure. The precision of 0.97 in the Neutral category underscores the model’s capability to discern nuanced emotional expressions, a challenging aspect in affective computing. The overall accuracy of 94% is indicative of the model’s effectiveness in dealing with the intricacies of multimodal data, as presented in the CMU-MOSI dataset. These results validate the CA-WGNN model’s theoretical underpinnings and establish its practical applicability in real-world scenarios, particularly in the domain of multimodal affective computing. The comprehensive performance analysis using the CMU-MOSI dataset offers valuable insights into the model’s potential for broader applications in sentiment analysis and emotional recognition. The table above provides a quantitative evaluation of the CA-WGNN model’s performance. The reported metrics, namely precision, recall, and F1-score, reflect the model’s effectiveness in various sentiment categories (Negative, Neutral, and Positive). With an overall accuracy of 94%, the CA-WGNN demonstrates its robustness in analyzing and classifying complex network structures. Notably, its superior performance in the Neutral category, with a precision of 98% and an F1-score of 96%, underscores the model’s nuanced understanding of subtle contextual variations. These results validate the efficacy of the cumulative attribute weighting and enhanced contextual understanding principles inherent in the CA-WGNN design.

In the evolving field of emotion recognition, the IEMOCAP and CMU-MOSI datasets are essential, encompassing diverse text, audio, and video modalities. Their extensive multimodal data are crucial for those studying emotional expression. These datasets’ varied participant demographics, emotional contexts, and data collection methods impact the applicability of models across different datasets, highlighting challenges in effectively merging data from these varied modalities. Traditional approaches often fall short of understanding the complex interplay among different modalities. Exploring the landscape of emotional recognition technologies, we encounter several

challenges across different models. Analyzing complex emotional data with SVM is difficult due to its sensitivity to kernel, parameter choices, and computational demands. Likewise, the VC(LR-SGD) faces hurdles, including hyperparameter sensitivity, potential biases in unbalanced datasets, and the presupposition that all factors contribute equally. TheGBM also grapples with issues like overfitting, handling unbalanced datasets, and significant computational expenses. In conclusion, while each model offers potential for emotional recognition, their limitations underscore the need for careful consideration in model selection and the ongoing pursuit of more adaptable and efficient solutions.

This research introduces the Cumulative Attribute- Weighted Graph Neural Network CA-WGNN model, an innovative breakthrough in Sentiment Recognition. The CA-WGNN’s graph-based structure excels in attributing weights to features across modalities, adeptly recognizing inter-modal connections, and assigning relevant weights to each modality’s features. Its versatility with different modalities and specific dataset characteristics, like those of IEMOCAP and CMU-MOSI, showcases its capability in universalizing Sentiment Recognition models across various datasets. The CA-WGNN’s sophisticated grasp of intricate relationships and effective feature integration marks notable progress in this field. The study also offers a detailed comparison of experimental outcomes in subsequent Tables, discussing the CA-WGNN algorithm’s benefits, challenges, and environmental impacts. Table 4.15. presents the accuracy, precision, recall, and F1-scores from IEMOCAP and CMU-MOSI, confirming the algorithm’s proficiency in Sentiment Recognition. Table 4.16. outlines the key benefits and obstacles in implementing CA-WGNN, along with its environmental effects in real-world applications. These tables collectively illuminate the algorithm’s performance and flexibility in various environments, underscoring the importance of multimodal integration in affective computing. This comparison particularly emphasizes CA-WGNN’s skill in managing complex, cross-modal interactions, vital to the evolution of affective computing. In the context of emotion detection in the text, particularly when examining datasets such as CMU-MOSI, which encompass textual, audio, and visual data, the ROC as shown in the Figure 4.15. curve emerges as a crucial analytical tool. If we take a closer look at the ROC curve using examples from the text portion of two distinct CMU-MOSI datasets, we can gain insights into the model’s capability to differentiate between various emotional states expressed in the data.

Figure 4.15. curve graphically illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate ($1 - \text{specificity}$) at different thresholds. By analyzing the happiness and sadness, present in the text data. The area under the ROC curve, we can evaluate how effectively the model distinguishes between different emotional states, such as happiness and sadness, present in the text data. The area under the ROC curve AUC offers a concise measure of the model’s overall performance, with a higher AUC indicating greater accuracy. This evaluation is particularly vital in scenarios where maintaining a balance between correctly identifying emotions (true positives) and

minimizing false identifications (false positives) is paramount.

Thus, the ROC curve serves as an indispensable benchmark in the optimization of emotion detection models, ensuring they accurately and reliably interpret the complex array of human emotions embedded in text, as well as in multimodal data like that of the AUC datasets.

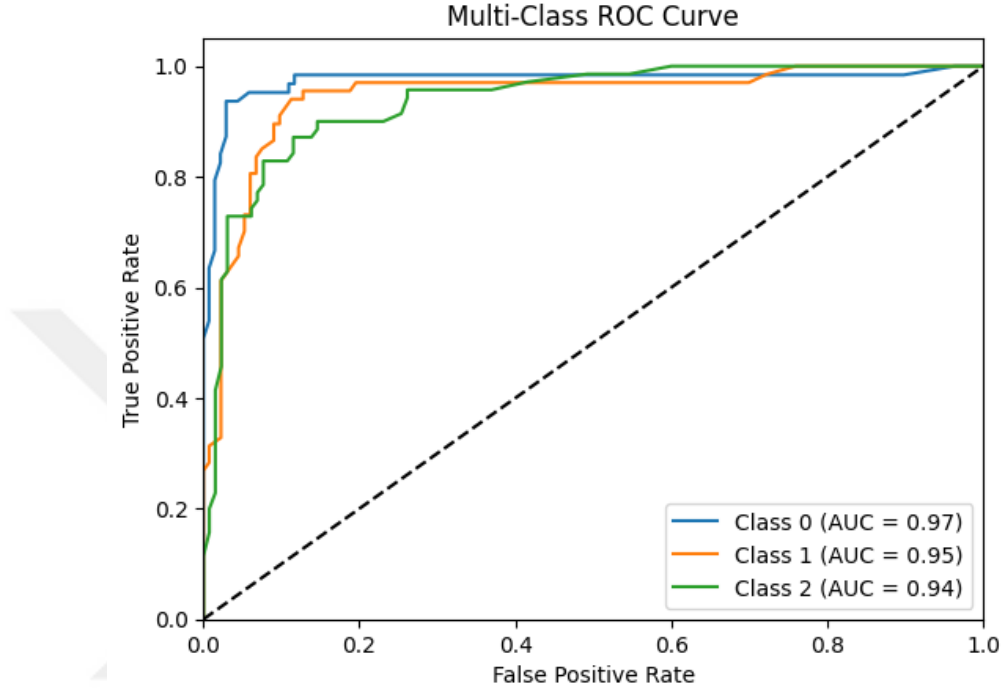


Figure 4.15. ROC Curve Analysis for Text Data from the CMU-MOSI Dataset.

The CA-WGNN model offers numerous advantages in the GNNs domain for emotion recognition and sentiment analysis. Its graph-based framework enables a deeper insight into the interconnected nature of multimodal data, effectively modeling the complex dynamics of emotional expression for more accurate sentiment analysis. The model's capacity to apply different weights to various modalities allows for customized approaches for each dataset, ensuring effectiveness across datasets with diverse structures and content. Moreover, CA-WGNN's focus on cumulative weighting enhances emotion recognition accuracy. By prioritizing impactful features in an emotional context, the model ensures that more relevant data are emphasized, improving emotion and sentiment detection precision while reducing misinterpretation from less pertinent or noisy data. Overall, the CA-WGNN model signifies a substantial advancement in GNNs applications for emotion recognition and sentiment analysis. Its cutting-edge structure and strategic weighting mechanism adeptly capture human emotional nuances, leading to more dependable and insightful affective computing analyses.

Table 4.15. Comparison of Performance Metrics

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
IEMOCAP (Dataset 1)	93.00	93.00	93.00	93.00
CMU-MOSI (Dataset 2)	94.00	94.00	94.00	94.00

Table 4.16. Challenges of CA-WGNN

Advantages	Challenges	Environmental Effects
Through effective text, audio, and video integration, CA-WGNN captures complex interactions for comprehensive ER. CA-WGNN overcomes standard models' inefficient fusion and captures cross-modal interactions using a graph-based design.	The model's training and inference processes can require more computer resources due to the graph-based design's complexity.	CA-WGNN's enhanced performance and robustness to help multimodal emotion identification systems in real-world applications.
Using two separate datasets, IEMOCAP and CMU-MOSI, the model shows that it can generalize to varied emotional circumstances and demographics of participants.	Although CA-WGNN is more adaptable, datasets can remain with different participant demographics and data-collection methods, making generalization difficult.	The model's flexibility and integration of multiple modalities can require more computing resources in resource-constrained contexts.
CA-WGNN uses a cumulative weighing method to improve the model's data interpretation by assigning relative values to features in each modality.	The cumulative weighing method may make it difficult to determine each attribute's decision-making impact.	In varied emotional settings, the model's capacity to generalize across datasets makes it useful.

The Trimodel emotion identification system, powered by the innovative CA-WGNN method, marks a significant leap in the realm of Sentiment Recognition. This system excels in harmonizing the unique strengths of textual, audio, and visual modalities, weaving them together through a sophisticated weighted graph structure. Its prowess lies not just in the seamless integration of these diverse data sources but also in its superior performance in detecting nuanced emotional states. This effectiveness is vividly demonstrated in comparative experiments, where it eclipses existing state-of-the-art models. A key aspect of this system is its adept modality-specific feature extraction, complemented by a comprehensive approach to emotional categorization. This dual strategy doesn't just enhance accuracy; it also illuminates the intricate interplay between different emotional expressions and the modalities through which they're perceived. The potential applications of this approach are vast and varied, spanning from mental health assessments to enriching HCI. When benchmarked against two prominent datasets, IEMOCAP and CMU-MOSI, the system showcases its robustness, particularly with the

CMU-MOSI dataset. Here, it achieves an impressive accuracy of 0.94%, along with equally remarkable precision, recall, and f1-score, all clocking in at 0.94%. This level of precision underscores the system's ability to offer a deeper, more nuanced understanding of emotions, paving the way for more advanced affective computing applications. Looking ahead, this model is poised to revolutionize the field of affective computing. In an ever-evolving technological landscape, the TER system stands out as a groundbreaking development, pushing the boundaries of emotion detection research. Its introduction heralds a new era in multimodal affective computing methodologies, setting the stage for further explorations and innovations.



5. MULTIMODAL EMOTION RECOGNITION USING BI-LG-GCN FOR MELD DATASET

Emotions are central to how humans interact. With more multimedia data available, our understanding of computer recognition of these emotions has greatly improved. Expression of emotions, whether written or spoken, deeply influences behavior and markedly affects decision-making, learning, and cognitive functions. In contemporary periods, the domain of artificial intelligence has strived to develop platforms proficient in comprehending and expressing emotions, progress majorly influenced by innovations in affective computing. Audio and written content integration has evolved, enhancing user experience and comprehension. As a result, dealing with singular modes of texts without this combination has become even more tedious and intricate [46]. The integration of audio and written content has improved, enhancing user comprehension and experience. Consequently, handling text without this combination has become more cumbersome and complex, underscoring the importance of bimodal content [47]. Challenges persist in communication via speech, including variability in the precision of emotional speech detection and a shortage of foundational understanding centered on emotions. Such dimensions are quintessential for articulating intent, interpreting social cues, and nurturing profound connections. Traditionally, emotion discernment predominantly relied upon facial lines [48]. Humans employ many means for emotional conveyance, encompassing gestures, postural indications, tonal variations, and facial nuances. A comprehensive approach that contemplates these diverse indicators is essential for precise emotion discernment and assimilation [49]. Stemming from this understanding, multimodal emotion recognition has emerged as a captivating field. It involves aggregating and examining emotional indicators from multiple channels. A more holistic and accurate representation of human emotions can be achieved by amalgamating information from varied sources, such as facial cues, vocal rhythms, physical gestures, and physiological responses. This interdisciplinary domain, converging research from affective computing, voice analytics, computer visualizations, and machine intelligence, endeavors to formulate algorithms and models adept at adeptly processing and interpreting multimodal emotional stimuli [50].

Improving human-computer interaction is an important factor behind multimodal emotion detection. To provide more individualized and interesting encounters, emotionally intelligent systems can modify their behavior and replies according to the user's emotional state [51]. Similarly, multimodal emotion recognition may be used in healthcare applications to assess a patient's emotional health and provide important insights for individualized care and treatment. Multimodal emotion recognition faces a variety of difficulties. It is a challenging effort to integrate and synchronize data from several modalities, each with distinct properties and noise sources [52]. Deep learning algorithms

have been developed to mix data from many modalities efficiently. Because emotional displays can vary greatly across people, creating solid models that can generalize across diverse people and cultural backgrounds is another difficulty. Further obstacles come from the subjectivity and context-dependence of emotions, which individual experiences may impact.

While GCNs, or Graph Convolutional Networks, have proven indispensable for visually understanding intricate systems through graphs, the importance of textual data in human interactions cannot be overlooked, especially in the realm of sentiment analysis. Graphs, as outlined by the rising popularity of infographics and knowledge graphs, can convey information efficiently[53]. On the other hand, textual data plays a significant role in sentiment analysis, as evidenced by the work of Mucahit Pinar Savci et al.[54]. Their research, which focused on evaluating deep learning and machine learning methods using e-commerce corpora in different languages, emphasized the power of pre-trained language models in discerning various sentiments. As for emotion extraction from speech, Pulatov et al., [55] present a notable advancement by employing dual feature extraction techniques, combining CNNs, MFCC, and Speech2Vec. While their methodology displays impressive accuracy enhancements on two renowned datasets, the omission of 10-fold cross-validation might prompt questions on generalizability. Nevertheless, their findings, suggesting the potential integration of diverse features and modalities, significantly enrich the current understanding of Speech Emotion Recognition research. Therefore, while graphical representation simplifies our comprehension of complex systems, textual data analysis remains crucial for understanding human emotions and intentions. Future studies might explore how integrating these two domains—graphical and textual—can bring more holistic insights across various sectors. Multimodal emotion identification has a lot of potential despite the difficulties in several areas. It may benefit from affective computing, human-robot interaction, virtual reality, gaming, market research, and psychology. Researchers and practitioners may design intelligent systems that are more sympathetic, intuitive, and responsive by analyzing and interpreting the intricate interplay of emotional signals [56]. There are many different ways that people might express and interpret their emotions. It tries to give a comprehensive knowledge of human emotions by including a variety of modalities. By developing emotionally intelligent systems, the discipline has the potential to transform human-computer interaction, healthcare, and other industries completely. We proposed a bi-lateral gradient graph convolutional network (Bi-LG-GCN) method to recognize the emotions of human beings.

5.1. Contributions of Bi-LG-GCN

- We provide the MELD dataset, which was utilized for accurate emotion categorization on both textual and audio modalities.
- To improve the multimodal dataset’s quality and consistency, we use pre-processing

techniques (linguistic and normalization).

- We apply K-PCA as a feature extraction method to get specific characteristics from each modality.
- To assess the effectiveness of our suggested strategy, we carry out thorough metrics such as accuracy, precision, recall, and f1-score using the multimodal dataset.

The following sections of the paper are written as follows: Section 2 contains reviews of the relevant literature. Section 3 explains the proposed approach, Section 4 discusses the findings, and Section 5 presents the conclusion. sectionBackground The field of research into understanding human emotions is rapidly growing and attracting the attention of both industries as it delves deep beyond facial expressions and body language to explore how emotions resonate within the nuances of our spoken words. Emotion, an intricate aspect of human communication, is pivotal in conveying sentiments, thoughts, and intentions. Our speech encodes information through elements like text meaning, speech rate, rhythm, pitch, volume, voice quality, articulation, duration, and inflection. Accordingly, the research objectives outlined of [89], a novel multi-modal approach for recognizing human emotions was introduced. The proposed method employed a deep learning architecture, specifically the "3D-Convolutional Neural Network (3D-CNN)," to extract spatiotemporal features from both electroencephalogram (EEG) signals and facial recordings. Then, a mix of data enhancement and ensemble learning approaches was presented to obtain the final fusion projections. The suggested scheme's multi-modality fusion was accomplished using data and score fusion approaches. The "3D-CNN output characteristics of the face chunks were then classified using the Support Vector Machine (SVM) classifier". In a parallel vein, wi and Liang [88] proposed a fusion-based approach to detect emotions in affective speech. This method integrated acoustic-prosodic information and semantic labels with multiple classifiers.

The text serves as a cornerstone in emotion recognition, crucial for enhancing human interaction within affective computing systems across languages. Mucahit Soylu et al.,[105] delved into attitude markers (AMs) in academic writing by English and Turkish authors. They identified 'significance' as a key functional category and 'adjective' as the top form. Their innovative approach, incorporating Java for data cleaning and a radial knowledge graph for visualization, offers promising avenues for emotion recognition in diverse linguistic landscapes. A key component of their methodology was the MDT technique, which is responsible for selecting the most suitable classifier based on recognition confidence. This work, although focused primarily on audio and semantic information, highlighted the effectiveness of using multi-modal data and sophisticated classifiers for enhanced emotion recognition. They emphasized the synergistic effects of combining distinct data types, achieving an impressive recognition accuracy of up to 85.79% when considering the personality traits of speakers. Building on the exploration of acoustic features, Jin et al. [107] centered on the fusion of features from both acoustic and lexical

levels to enhance emotion recognition in speech. At the acoustic level, a gamut of features, including intensity, F0, and others, was extracted, with novel representations such as Gaussian supervectors introduced. The lexical dimension saw the proposal of the "emotion vector" (eVector) feature, which was predicated on emotion lexicons, giving words weights based on their emotion-expressive inclination. Applied to the USC-IEMOCAP database, the study demonstrated that the late fusion of both these feature realms culminated in a four-class emotion recognition accuracy of 69.2%. This further solidified the potential of combining various feature sets to hone the precision of emotion recognition, marking it as a noteworthy contribution to the field. The researchers in [333] introduced a new approach for multimodal emotion recognition utilizing raw waveforms and cross-modal attention mechanisms. By integrating raw audio processing through a one-dimensional convolutional model and establishing a cross-modal attention network between audio and text features, they optimized their emotion detection system. This approach showed that features extracted from raw audio, when coupled with a cross-modal attention mechanism, can effectively capture the interplay between audio and textual cues, enhancing emotion classification capabilities.

The investigators of [334] presented an innovative emotion detection system according to various expressions and other modalities of the face, Galvanic Skin Response (GSR), and electroencephalogram (EEG). Emotional Analysis Databases for variable numbers of emotional classes, four on a median, including angry, disgusted, fearful, joyful, neutral, sad, and confused [26]. In natural deceptive facial expressions, the suggested model demonstrated the benefit of detecting the proper state of emotions. In the study by [335], a method for multi-modal emotion detection was introduced, which harnesses speech-visual correlation features. This approach utilized two-dimensional convolutional neural networks (2D-CNN) to extract speech characteristics and 3D-CNN for visual attributes. During multi-modal fusion, a feature correlation analysis technique was employed to analyze both audio and visual data. Results from experiments performed on diverse datasets show that the approach is comparable to other cutting-edge algorithms in terms of recognition effectiveness. The focus is on deploying deep learning methodologies to discern underlying emotional tones in textual data. A notable advancement in this domain is the Emo2Vec model, which, when integrated with Logistic Regression and GloVe, demonstrates competitive performance in emotion detection. Ragheb et al. have made a significant contribution by developing a learning-based model that recognizes six emotions as described by Paul Ekman; their method employs a two-phase approach of encoding and classification, utilizing tokenization, encoders, and Bi-LSTM units trained via average stochastic gradient descent [104]

According to Lian et al., [94], they proposed domain adversarial neural networks (DANN) for emotion detection. The main objective was to forecast emotion categories and develop a common representation where speaker's identity could not be discriminated. The depictions of different speakers were closer together thanks to their use of this technique.

Using unlabeled data during the training phase, they reduced the effect of low-resource sample collections. The researchers of [97] introduced a deep learning (DL) based method for accessing and combining text and perceptual data to classify emotions. To extract acoustic characteristics from unprocessed audio, they added a DCNN layer after using a SincNet layer based on a customizable sinc value with band-pass filtering. When compared to performing convergence over the raw voice signal, the method develops filter banks modified for the recognition of emotions and provides superior aspects. To infer N-gram level association on hidden models derived from the Bi-RNN, they used two parallel streams for text analysis "(a DCNN and a Bidirectional RNN followed by a DCNN)" with cross attentiveness. The researchers of [75] demonstrated M3ER, a learning-based approach for emotion identification from several input modalities. Their system incorporated inputs from many co-occurring models and is more resistant to any of the sensors are noisy various approaches than previous methods. "M3ER model an innovative, data-driven multiplicative fusion strategy" for combining the models, which learns to enhance the extra accurate cues. M3ER was resistant to sensor noise by incorporating a check step that used Canonical Correlational Assessment to distinguish between ineffective and effective models.

Liu et al., in their study [90], proposed utilizing deep canonical correlation analysis (DCCA) for recognizing emotions across multiple modalities. DCCA primarily functions by independently transforming each modality and then correlating the varied modalities into a unified hyperspace, adhering to predefined canonical statistical constraints. The efficacy of DCCA was assessed across five different datasets. The research results revealed that DCCA obtained state-of-the-art detection accuracy rates across all datasets. Mittal et al. introduced "EmotiCon," a learning-based system designed for identifying perceived emotions from photos and video clips by considering context [98]. This approach incorporates three different perspectives of context for emotion detection. The initial interpretation was based on the use of various senses for emotion identification. In the second analysis, they obtained semantic information from the input image and used a self-attention-based CNNs to encode the information. At last, depth maps were employed to simulate the third interpretation, which was connected to socio-dynamic interactions and agent proximity. They showed the effectiveness of their network by running tests on datasets.

The researchers [101] suggested an innovative cross-representation speech model for detecting emotions on "wav2vec 2.0 voice characteristics and also trained a CNN-based model to distinguish emotions from text data extracted with Transformer-based models. A score fusion method was used to merge the speech-based and text-based findings". Authors of [336] developed a method for multimodal emotion identification named "deep generalized canonical correlation analysis with an attention mechanism (DGCCA-AM)." The model established multimodal adaptive fusion with a focus system, extending the usual "canonical correlation analysis (CCA) from two modalities" to arbitrarily numerous

modalities. According to Zhang [337] suggested a deep automated encoder-based expression-EEG interaction multi-modal recognition of emotions approach. In the beginning, a decision tree was used as a feature-based selection approach. The solution vector values were then examined to establish the expression classification for the test sample based on the facial expression characteristics detected by sparse representation. The bimodal deep automated encoder was then used to combine the EEG. The 3rd layer of BDAE extracts characteristics for the supervised training phase. To complete the categorization, the LIBSVM classifier was employed.

An approach to identifying age, gender, and state of mind from audio was developed by Zaman et al. [102]. All audio recordings in our system were transformed into 20 statistical attributes, and the transformed numerical dataset was utilized to build several prediction models in order to achieve the goal. "Knearest neighbors KNN, XGBoost, AdaBoost, and Decision Tree, Artificial neural networks (ANN), Naive Bayes, and Support vector machine (SVM)." The authors of I will schedule some time for us to connect.[99] suggested an innovative emotion-relevant crucial subnetwork identification method and examined 3 EEG functional connectivity network features. On three open emotion EEG databases, the identifying capacity of the EEG connectivity characteristics in emotion detection was examined. By single-channel analysis, the strength feature surpassed the state-of-the-art differential entropy feature in the accuracy of classification. The findings from the study demonstrated that each of the five emotions had different functional conjunction characteristics. The researchers [72] proposed a deep multi-task learning system that analyzed emotions simultaneously. A video's multi-modal inputs provide unique and distinct information, and they typically do not contribute equally to decision-making. They suggested a context-level inter-modal attention method for evaluating an utterance's sentiment and conveying emotions at the same time. They tested the suggested method for multi-modal sentiment and evaluation of emotions using the CMU-MOSI database.

According to Nemati et al. [91] they suggested an integrated multimodal data fusion approach in which the audio and visual models were fused using a latent space linear map, and their presented characteristics into the cross-modal space were fused with the textual modality using a Dempster-Shafer (DS) theory-based evidential fusion approach. The suggested technique outperformed both decision-level and non-latent space fusion approaches when tested on DEAP dataset. The HED dataset was created by Fang et al. of [103] as a sizable multimodal emotion dataset to aid in the emotion detection challenge. The method for multimodal emotion recognition was then suggested. The "HED" dataset was far bigger than previous datasets and comprises emotion-aligned face, body, and text samples that express many emotions, including happiness, sadness, disgust, anger, and fear. The researchers of [100] utilized Muse-CaR, a vehicle review database, to make continuous emotional predictions. To do this, they initially extract handmade characteristics and complex depictions from several modes. They next used the "Long Short-Term Memory (LSTM) RNNs and the self-attention system to represent the

sequence's complicated temporal relationships. The Concordance Correlation Coefficient (CCC) loss was used to direct the model's learning of local fluctuations as well as the global trend of emotion". At last, two fusion procedures, early fusion and late fusion, were used to improve the model's performance even more by complementing information from various modalities.

To enhance the efficiency of the emotional identification framework, researchers presented a multimodal emotion detection model based on voice and text [92]. Additionally, the study confirmed that the intersection of sentiment analysis of textual data and the domain of artificial intelligence yields effective outcomes for classification and prediction. Various neural network architectures, including CNNs, LSTM, and GRU, have demonstrated their efficacy in this regard [95]. To learn acoustic emotion features, "CNN and LSTM (long short-term memory) were combined in the form of binary channels; in the meantime, an effective Bi-LSTM (bidirectional long short-term memory) network" was used to record textual elements. In addition, a Deep Neural Network DNNs was employed to classify and gain knowledge of the fusion characteristics. The results of both voice and text emotion analysis were used to establish the final emotional state. The authors of [93] introduced the Visual Aspects Attention Network, or VistaNet, which integrates text and visual elements. They demonstrated that, in many situations, visuals support text in emotion identification, emphasizing significant elements of an entity rather than conveying emotions independent of the text. As a result, rather than using visual data as features, VistaNet uses visual information as alignment to highlight the relevant phrases in a text using attention.

5.2. Methodology

Leveraging bimodal datasets in affective computing is gaining traction due to its ability to enhance human interactions and multiple use cases. Discerning emotions with multimodal information significantly elevates the caliber of human engagements. This approach is drawing widespread attention and is under exploration in various domains. By amalgamating details from sources such as facial cues, vocal nuances, gestures, and physiological indicators, multimodal emotion detection offers a deeper understanding of emotional states. This is particularly salient in areas like healthcare, gaming, virtual reality, and human-machine interfaces [338]. We compiled the MELD dataset for this specific emotion detection purpose. Data preprocessing was executed with normalization techniques and linguistic methodologies. Kernel Principal Component Analysis (K-PCA) was introduced for feature derivation. Subsequently, we employed the proposed Bi-Lateral Gradient Graphical Conventional Network (Bi-LG-GCN) algorithm for emotion discernment. The results were assessed based on metrics including accuracy, precision, recall, f1-score, Mean Squared Error (MSE), and Mean Absolute Error (MAE). The comprehensive methodological approach is illustrated in Figure 5.1..

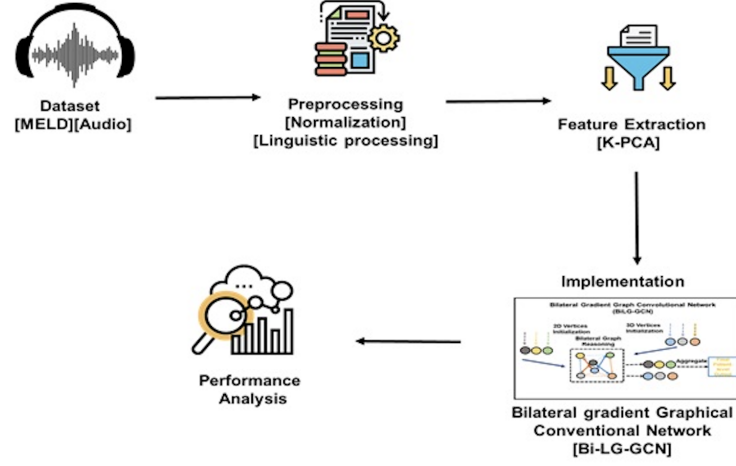


Figure 5.1. Multimodal Emotion Recognition using Bi-LG-GCN system

5.2.1. Data Collection

In our daily interactions, emotions are frequently communicated and understood through a mix of signals from various modalities. Figure 5.2. provides a glimpse into multimodal databases that accumulate both auditory and textual information. Given the significance of reliable data, many researchers have dedicated efforts to identify nuanced emotions, be it manifestly or in more subtle forms. These feelings typically originate from three main channels: text, voice, and visuals. As we embark on this data-gathering phase, it's pivotal to pinpoint the specific type of data to collect and the sources from which they'll be procured. The method of data collection can differ based on the application's nature [53]. A prime illustration of this is the MELD dataset, which was crafted by enhancing the core affective computing dataset.

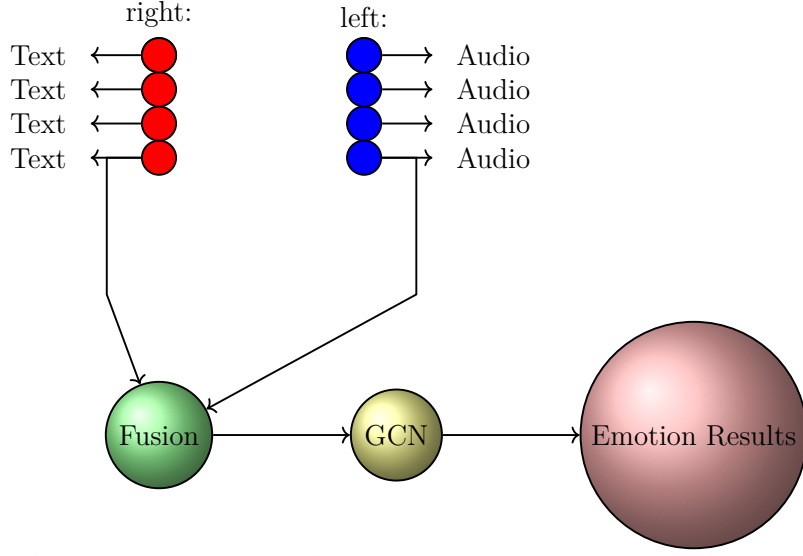


Figure 5.2. Design for Combining Multiple Text and Audio Datasets

The MELD was developed by improving and expanding upon the original Emotion Lines dataset. This multimodal dataset includes audio and visual media and text and consists of the same conversation examples found in Emotion Lines. For the purpose of multimodal emotion identification, we only employ the text and audio modalities. The dataset makes available more than 1400 conversations and 13000 utterances from the Friends TV series. The discussions included a number of speakers. Each remark in a discourse is given an identity according to one of the seven emotions: fear, anger, surprise, sorrow, neutral, disgust, and joy. Additionally, the dataset annotates each utterance's emotion (positive, negative, and neutral). The dataset was collected from (<https://www.kaggle.com/datasets/zaber666/meld-dataset>).

5.2.2. Data Preprocessing

Once the data has been collected, the primary task is to refine it for training the Bi-LG-GCN. This refinement may require feature extraction or modifications to account for variations in speaker attributes or recording quality. Such adjustments are crucial, particularly when dealing with datasets comprising both text and audio, to guarantee consistent normalization and effective language processing.

Normalization

Normalization describes standardizing and modifying input data from numerous modalities (text input, voice) to a single scale or range. This approach tries to eliminate differences and disparities between modalities, allowing for optimal emotional information

integration and comparison. By normalizing the data, information acquired from diverse modalities may be treated similarly and merged in a meaningful way. This normalization process improves the accuracy and reliability of the emotion detection system, allowing for more robust and precise recognition and interpretation of human emotions across many modalities.

Linguistic Processing

Linguistic processing is the computerized examination and interpretation of linguistic information with the purpose of identifying and detecting emotions in a multimodal environment. Enhancing the durability and precision of identifying emotional systems entails synthesizing emotional information from many language modalities. To evaluate linguistic elements like emotions, semantic content, and structural patterns, this discipline utilizes techniques from Machine learning, AI, and the processing of natural language. The objective is to enhance human emotion interpretation and processing across several modalities, enabling more precise and thorough emotion detection[6].

5.2.3. Feature Extraction Using Kernel Principle Component Analysis

Feature extraction is the process of extracting significant features from large amounts of data. It involves evaluating many modalities and translating raw input signals into compact and useful summaries that incorporate relevant emotional cues. These retrieved properties are input for machine learning algorithms to classify emotions. In our work, we proposed Kernel Principle Component Analysis (K-PCA) for feature extraction. For reducing the dimension of vast quantities of data, PCA is a popular multimodal estimate technique. Typically, dimensionality reduction is done by randomly selecting the linear relationship between the parameters. However, as was previously said, conventional PCA offers linear dimensional reduction. KPCA is still a more effective method for mapping a non-linear process in a data set. Contrary to other non-linear techniques, the ability of the kernel algorithm to run without any non-linear optimization is essential. This approach includes altering the input variables and using them as independent PC parameters. Kaiser-Meyer-Olkin (KMO) [339] is one of the most often used statistics for determining the amount of data in any factor analysis (FA), a quick explanation of the procedure to establish KPCA for feature extraction is provided with Eq. 5. 1.

$$KMO = \frac{\sum \sum q_{ji}^2}{\sum \sum q_{ji}^2 + \sum \sum b_{ji}^2} \quad (5. 1)$$

Where q_{ji} is the correlation value between variables j and i , and b_{ji} is their partial correlation value. Considering that the non-linear change $\phi(w)$ from the initial in sample covariance matrix D in F space should meet the Eq. 5. 3, the projected novel characteristics have zero

mean Eq. 5. 2:

$$\frac{1}{M} \sum_{j=1}^M \Phi(\Phi(W_j)) = 0 \quad (5. 2)$$

$$D = \frac{1}{M} \sum_{j=1}^M (\phi(W_j) \cdot (\phi(W_j))^S) \quad (5. 3)$$

If the kernel function is formatted as follows Eq. 5. 4:

$$l(W_j, W_i) = \phi(W_j)^S \phi(W_i) \quad (5. 4)$$

$$L_{bl}^2 = \lambda_l M L_{bl} \quad (5. 5)$$

$$L_{ji} = l(W_j, W_i) \quad (5. 6)$$

Where $\mathbf{b}_l$ is an N -dimensional column vector and $\mathbf{b}_{lj}^s$ as follows in the Eq. 5. 7:

$$\mathbf{b}_l = [b_{l1}, b_{l2}, \dots, b_{lM}]^S \quad (5. 7)$$

The theory b_{lis} is solvable by the Eq. 5. 8

$$L_{bl} = \lambda_l M_{b_l}, \quad (5. 8)$$

and the corresponding kernel primary components can be determined by the Eq. 5. 9

$$z_l(W) = \phi(W)^S \quad u_l = \sum_{j=1}^M b_{lj} \cdot \ell(W, W_j) \quad (5. 9)$$

If the predicted dataset $\{\varphi(w_j)\}$ does not have a zero mean, the kernel matrix L can be replaced with the Gram matrix $\tilde{L}$. The Gram matrix is denoted by Eq. 5. 10:

$$\tilde{L} = L - 1_M L - L 1_M + 1_M L M \quad (5. 10)$$

Where 1_M is the $M \times M$ matrix with all components equivalent to $\frac{1}{M}$. The positive aspect

of kernel approaches is that it is not essential to calculate $\{\varphi(w_j)\}$ explicitly; the kernel matrix may be rapidly established from the training dataset $\{w_j\}$ in Eq. 5. 11.

$$k(W, Z) = \exp\left(-\frac{1}{2\sigma^2}\|W - Z\|^2\right) \quad (5. 11)$$

In our work, KPCA analyzes and understands human emotions by combining numerous sources of data. KPCA is a dimensionality reduction approach that converts the data provided into a lower-dimensional space while retaining vital details. KPCA can manage nonlinear interactions between modalities by utilizing a kernel function. This method attempts to find the fundamental trends and patterns in multimodal data, enabling more accurate and robust emotion identification across several modalities improving human emotion comprehension and interpretation.

5.2.4. Implementation of Bilateral gradient graphical Conventional Network

An innovative method for multimodal emotion identification is the bi-lateral gradient graph convolutional network (Bi-LG-GCN). Emotion recognition is the process of identifying and labeling the emotions that are sent by humans through a variety of modalities such as audio and text. The Bi-LG-GCN needs the advantage of graph convolutional networks (GCNs) to detect the complex relationships between multiple modalities and increase recognition of emotional performance. Figure 5.3. shows the GCN architecture.

The input data to the Bi-LG-GCN consists of various modalities, each represented as a graph. The nodes in each graph represent the individual samples in the dataset, while the edges among the nodes record the relationships between them. In the context of multimodal emotion detection, the nodes may represent distinct face, voice characteristics, or textual information, and the edges could reflect similarities between them. The Bi-LG-GCN's primary concept is to efficiently capture multimodal interactions by combining local and global information from graphs. A bilateral gradient approach that makes use of the advantages of both local and global graph convolution processes achieves this. While the global graph convolution covers the general framework of the network and captures global interactions, the local graph convolution concentrates on gathering information about each node's nearest area.

Bi-LG-GCN can be represented as follows:

Local graph convolution in the Eq. 5. 12:

$$g_j k = \text{ReLU}\left(V_k \cdot Y_j + \sum_{j \in M_j} \alpha_{ji} \cdot Y_i\right) \quad (5. 12)$$

A hidden instance of the i^{th} node after the l^{th} local convolution of the graph layer is denoted by the symbol g_j^k in this equation. The k^{th} local convolution layer's weight matrix is denoted

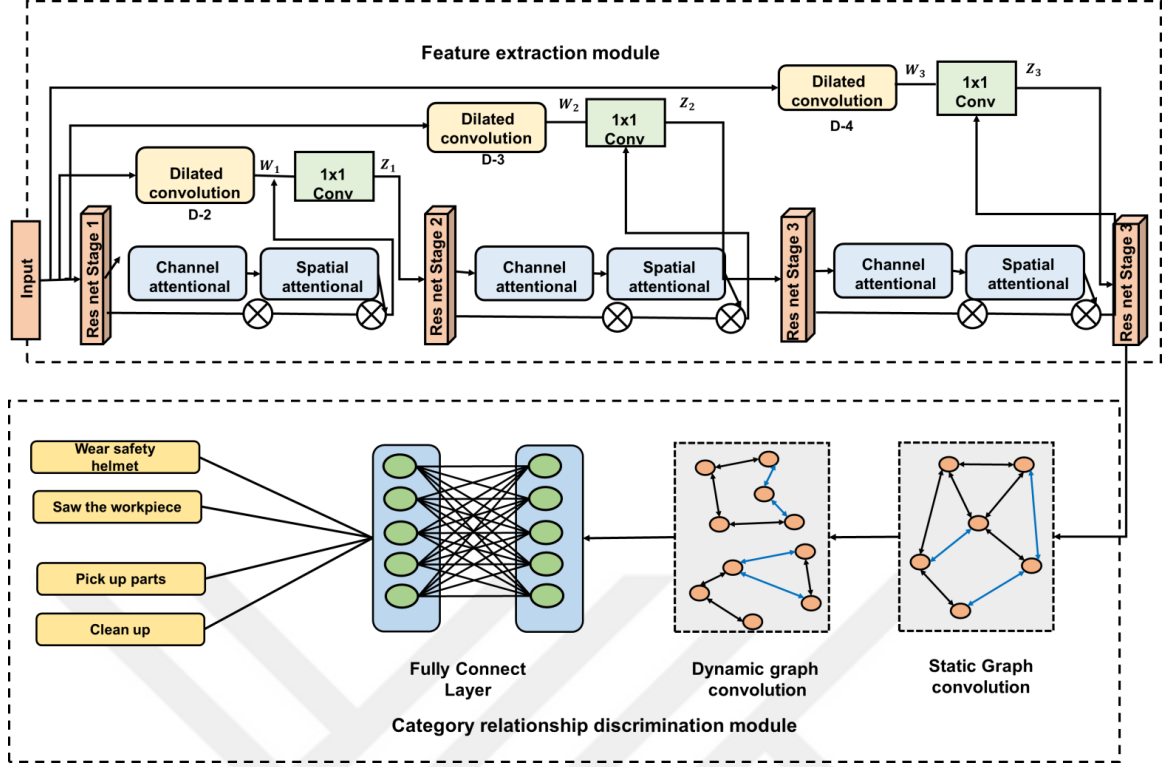


Figure 5.3. GCN Architecture

by V_k , while the i^{th} node's surrounding nodes are represented by M_j . The attention-related elements that control the information flow between the nodes are illustrated by the α_{ji} . Global graph convolution in the Eq. 5. 13:

$$g_j h = \text{ReLU} \left(V_h \cdot g_j k + \sum_{j \in w} \beta_{ji} \cdot g_i k \right) \quad (5. 13)$$

After global graph convolution, g_j delivers the hidden representation of the i^{th} node. W is for the set of all nodes in the graph, V_h stands for the weight matrix of the global convolution layer, and β_{ji} stands for the attention coefficients that capture the global connections between nodes. The bi-lateral gradient technique combines the local and global representations to generate the final node representations. It is achieved using an iterative optimization procedure that updates the attention coefficients α_{ij} and β_{ij} while minimizing the loss function.

Overall, the Bi-LG-GCN offers a strong foundation for multimodal emotion recognition by effectively capturing the intricate interactions between various modalities. The bi-lateral gradient approach combines local neighborhood information and the global graph structure to combine local and global information, improving the accuracy of recognized emotion. By adopting this methodology, the accuracy of emotion classification tools can be enhanced, and the broader domain of multimodal emotion recognition could also

witness significant advancements.

5.3. Experimental Results and Evaluation

Result evaluation in our context relies on accuracy, precision-recall, and the F1-score metrics. While accuracy offers a general overview, precision and recall prioritize positive predictions. The F1-score strikes a balance. The method employs Bi-LG-GCN for multimodal emotion recognition, assessing performance with these metrics to ensure precise emotion detection across diverse data sources.

5.3.1. Experimental Setup

We designed our experimental setup based on a combination of both hardware and software configurations tailored for efficient and taxing machine learning applications. The experiments were performed on a Windows 11 operating system. The chosen programming environment is Python 3.11, which supports the deployment of PyTorch 2.0. We also incorporated the use of Colab Google software to complement the setup. The underlying hardware and software infrastructure ensures the seamless implementation and execution of our technique.

Table 5.1. Hardware and Software Configuration

Component	Specification
OS	Windows 11
Processor	11th Gen Intel(R) Core(TM) i7-11370H @ 3.30GHz
Memory (RAM)	16.0 GB
Storage	SSD
Graphics Card	GPU 0 Intel(R) Iris(R) Xe Graphics GPU 1 NVIDIA GeForce RTX 3050 Ti Laptop GPU
Network Adapter	Ethernet (Turk Telecom) TTNET 24Mbps
Software	Pytorch 2.0 Python 3.11 Colab Google
Total Processing Time	(h, m, s): 00, 21, 31

5.3.2. Existing Methods

A crucial aspect of classification involves evaluating the models to gauge their precision and dependability in classifying data. Examining the performance metrics helps pinpoint the optimal model for classification assignments. Insights from model training and testing play a pivotal role in evaluating their efficacy. For the performance assessment of classification models, the confusion matrix is a standard tool. This matrix encompasses four components: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Using the confusion matrix, we can derive various performance metrics like accuracy, precision, recall, F1-score, etc., as illustrated in the color-coded axis above. Model predictions undergo testing and training processes, categorizing outcomes into the following segments: **True Positive** - A successful classification of an actual positive outcome, **False Negative** - An erroneous classification where a positive outcome is predicted as negative, **False Positive** - A misclassification where a negative outcome is predicted as positive, and **True Negative** - A successful classification of an actual negative outcome [55]. To provide a clearer visualization of these categorizations, the

following Table 5.2. shows how TP, FP, TN, and FN are represented.

Table 5.2. Representation of TP, FP, TN, and FN Explanation Matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

The accuracy, f1-score, precision, and recall metrics are used in this section to assess the performance of the suggested technique Bi-LG-GCN. For calculating accuracy and f1-score, we compare the Bi-LG-GCN with Dialogue Contextual Reasoning Networks (DialogueCRN) [204], Multimodal fused Graph Convolutional Networks (MMGCN), and Graph network-based Multimodal Fusion Technique [204]. For calculating precision and recall, we compare the Bi-LG-GCN with Support Vector Machine (SVM) [103], CentralNet, and Low-rank Multimodal Fusion (LMF).

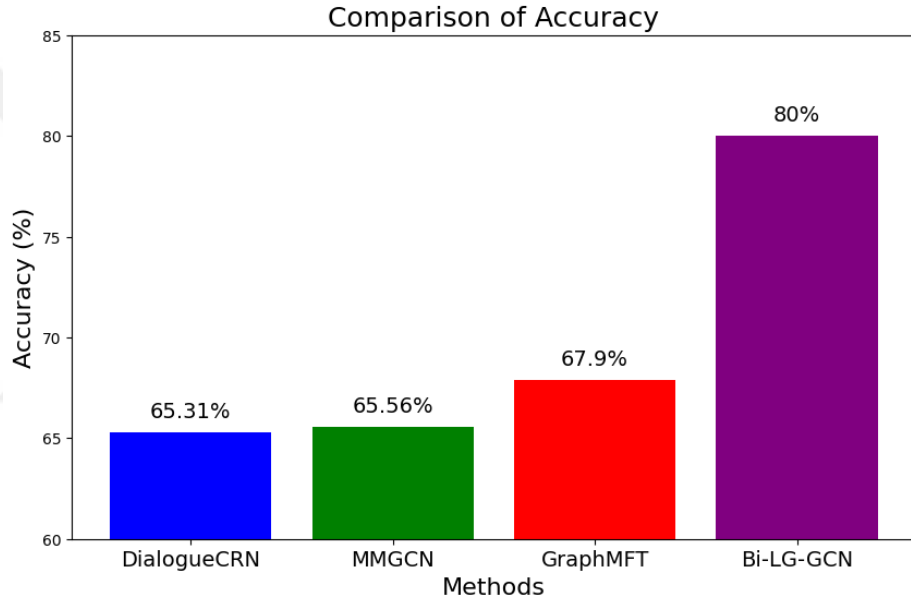
5.3.3. Comparison Phase

The degree of accuracy reflects how effectively these systems can identify and categorize emotions. It assesses how well the system can correlate the observed multimodal input data with the relevant emotional states, indicating how well it can decode and identify human emotions. Table 5.3. and Figure 5.4. show the accuracy comparison of the suggested Bi-LG-GCN and alternative procedures. Whereas Dialogue CRN, MMGCN, and Graph MFT only succeed with 65.31, 65.56, and 67.90% accuracy, respectively, the suggested technique Bi-LG-GCN achieves 80% accuracy. The Bi-LG-GCN achieved higher accuracy than the other methods. The accuracy value is expressed in the following Eq. 5. 14.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Number of Instances}} \quad (5. 14)$$

Table 5.3. Comparison of accuracy

Methods	Accuracy (%)
DialogueCRN	65.31
MMGCN	65.56
GraphMFT	67.9
Bi-LG-GCN	80

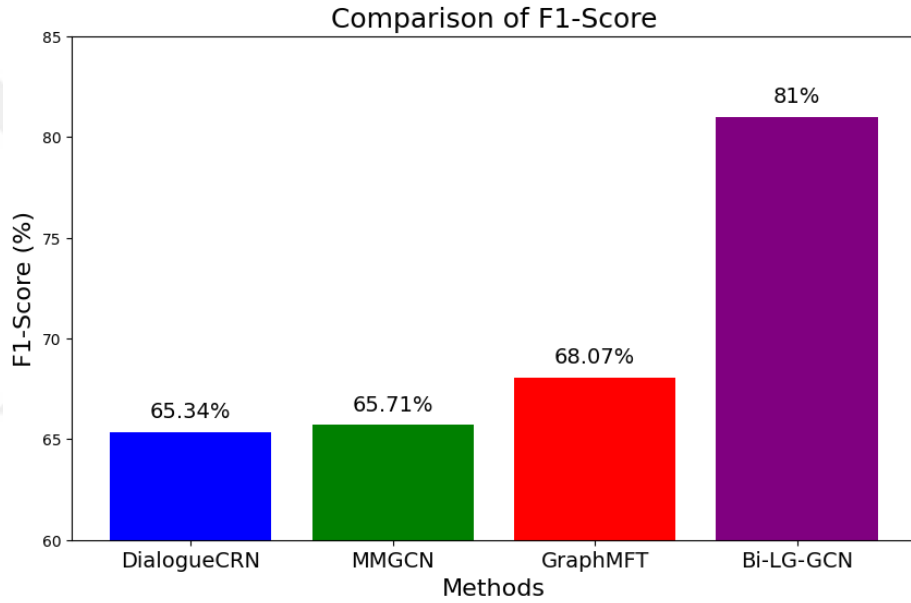
**Figure 5.4.** Accuracy comparison of Bi-LG-GCN with existing methods

The effectiveness of a model that attempts to identify and categorize emotions utilizing different forms of media combines recall and precision into a single measurement that considers the precision and comprehensiveness of emotion forecasts. The overall ability to capture emotions across several modalities is greater when the F1 score is higher. The F1 score comparison of the suggested and other approaches is displayed in Table 5.4. and Figure 5.5.. Whereas DialogueCRN, MMGCN, and GraphMFT only succeed at 65.34, 65.71, and 68.07%, respectively, the suggested technique Bi-LG-GCN achieves 81% accuracy. Comparing the Bi-LG-GCN method to other conventional approaches provides a higher F1 score. The F1-score with the Eqn. 5. 15 is described below.

$$\text{F1-score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \quad (5. 15)$$

Table 5.4. Comparison of F1-score

Methods	F1-Score (%)
DialogueCRN	65.34
MMGCN	65.71
GraphMFT	68.07
Bi-LG-GCN	81

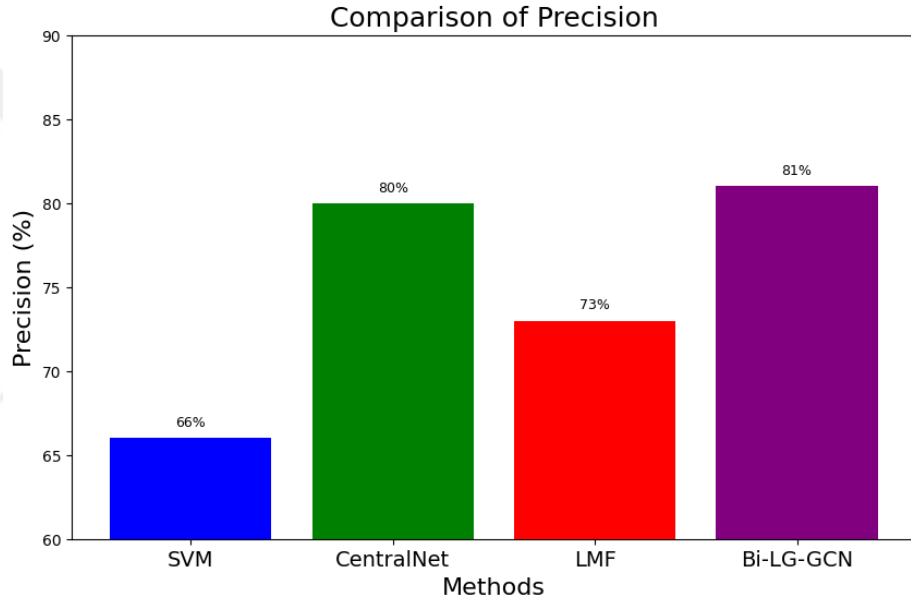
**Figure 5.5.** F1-score comparison of Bi-LG-GCN with existing methods

Precision is the ability to reliably and accurately recognize particular emotions from various signals; it calculates the percentage of emotions that were properly categorized out of all the emotions a system predicts. The system is more accurate at recognizing the desired emotions when the precision is higher since a lower percentage of false positives is shown. Precision is a key indicator of how well multimodal emotion recognition systems execute. Table 5.5. and Figure 5.6. demonstrate the precision comparison between the suggested and alternative procedures. A precision value of 81% is achieved by the recommended technique, Bi-LG-GCN, as opposed to merely 66%, 80%, and 73% for SVM, CentralNet, and LMF. The Bi-LG-GCN concept offers a high precision value when compared to earlier techniques. The following equation is used to determine the precision value with the Eqn. 5. 16.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5. 16)$$

Table 5.5. Comparison of precision

Methods	Precision (%)
SVM	66
CentralNet	80
LMF	73
Bi-LG-GCN	81

**Figure 5.6.** Precision comparison of Bi-LG-GCN with existing methods

Recall is a characteristic of a system that enables it to identify and classify emotions through multiple modalities consistently. The efficacy of the system in accurately discerning emotions from diverse data inputs is quantified by the ratio of genuinely positive emotions it has identified. Table 5.6. and Figure 5.7. illustrate the recall distinctions between the proposed and other methods. The suggested method, Bi-LG-GCN, attains a recall score of 81%, which surpasses the scores of 65%, 79%, and 69% achieved by SVM, CentralNet, and LMF respectively. Consequently, Bi-LG-GCN outperforms conventional methods in terms of recall with the Eqn. 5. 17. The ensuing formula determines the recall metric.

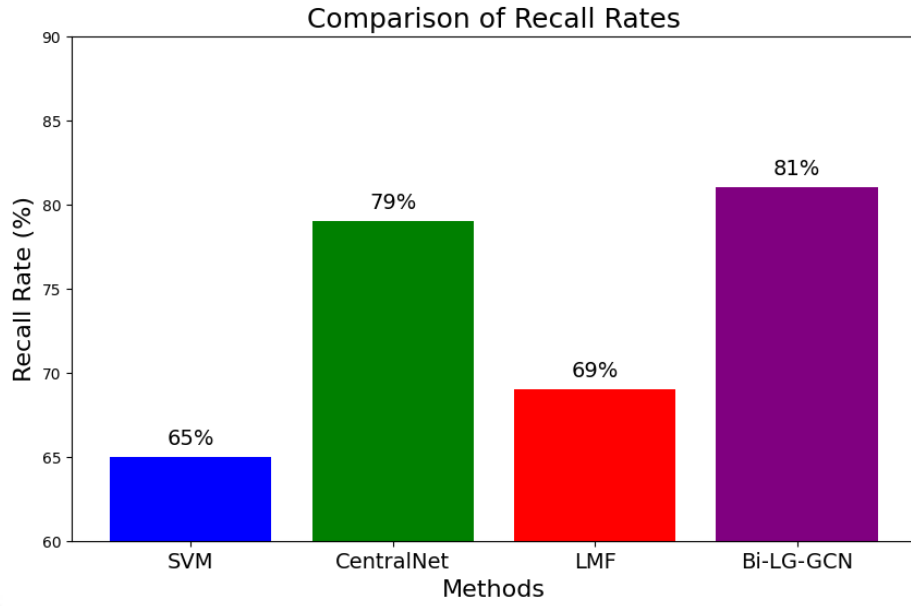


Figure 5.7. Recall comparison of Bi-LG-GCN with existing methods

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5.17)$$

Table 5.6. Comparison of recall

Technique	Recall Rate (%)
SVM	65
CentralNet	79
LMF	69
Bi-LG-GCN	81

In comparing the results, it becomes evident that Bi-LG-GCN demonstrates superior recall capabilities over other evaluated techniques. This dominance in performance reinforces the potential of Bi-LG-GCN for applications that prioritize recall. Furthermore, the detailed evaluation as per the provided formula offers a comprehensive perspective on the recall differences among these methods.

In our study, we proposed a novel method, a bi-lateral gradient graph convolutional network (Bi-LG-GCN) for multimodal emotion recognition. To recognize multimodal

emotions using Bi-LG-GNN, we collected the MELD dataset using textual and audio modalities. And applied linguistic and normalization pre-processing methods to improve the standard and standardization of both datasets. As a feature extraction technique, we used K-PCA to extract certain traits from each modality. The experiment is done on various parameters such as "accuracy (80%), precision (81%), recall (81%), F1-score (81%)" using Bi-LG-GCN for the MELD dataset. Several existing methods were used in the comparing phase. The experiment showed our proposed Bi-LG-GCN method performed efficiently when compared to earlier techniques. Its performance may be impacted by the quantity and caliber of multimodal input, which may result in unreliable emotion predictions. The approach may also have trouble generalizing to various cultural and linguistic environments, which makes it challenging to use in real-world situations involving a range of emotional expressions. For better user experiences, more studies can examine real-time emotion identification, cross-domain generalization, and its incorporation into virtual assistants and emotion-aware systems.

6. EMOTION RECOGNITION FROM TEXT AND AUDIO DATASET USING CUMULATIVE ATTRIBUTE GRAPH NEURAL NETWORKS

Envision a future where deaf and Hard-of-Hearing (HoH) individuals effortlessly overcome communication barriers. In our digital age, technological integration is crucial, especially for persons with disabilities. Approximately 20% of the global population experiences deafness and muteness, according to the World Health Organization (WHO) [340]. Traditional communication methods like sign language require mutual proficiency, often a rarity. Our study presents a groundbreaking AI solution designed to understand how emotion transforms how people with hearing and speech impairments communicate. This system is multilingual and utilizes sophisticated machine-learning methods to enhance communication skills greatly.

Organizations such as the World Association of Sign Language Interpreters and the World Federation of the Deaf, as highlighted in the United Nations' 2019 release [341], have strived to deploy human interpreters. The growing dependence on technology for translation and comprehension underscores the importance of accessible solutions. These tools should focus on the literal translation and capture the emotional context of words for clearer understanding. Our study introduces the Bimodal-Cumulative Attribute Graph Neural Network (Bi-CAGNN) algorithm, adept at interpreting and responding to both vocal and textual cues. This surpasses current technologies like O'Connor's gesture-sensing glove and Islam et al.'s sensor-equipped glove for the Bengali alphabet [342]. Our approach integrates wearable technology with advanced machine learning, as seen in the Automated Sign Language Interpreter (SLI) [343], providing a comprehensive solution for the HoH community. This article explores the Bi-CAGNN development and its potential as a transformative tool in enhancing communication accessibility for the deaf and HoH.

The Bi-CAGNN algorithm utilizes a neural network framework to process textual and vocal inputs, enabling the detection of a wide emotional spectrum. This is vital for the HoH community, offering insights into the subtleties of spoken language that extend beyond text. Additionally, the algorithm conducts an in-depth analysis of vocal characteristics such as tone, pitch, and rhythm, coupled with textual analysis. This dual approach provides a comprehensive understanding of communication, empowering HoH individuals with a deeper grasp of the speaker's emotions, going beyond simple text analysis capabilities. This adaptability is instrumental in making it a dynamic tool in the field of ER technology. Furthermore, it aims to diminish communication barriers for deaf individuals by facilitating a nuanced understanding of emotional nuances in conversations. This fosters greater inclusivity in social interactions and bridges a vital gap in text-only communication methods. Emotion recognition and audio-to-text technologies enhance communication for the hard of hearing by transcribing speech and capturing emotions,

making conversations more accessible. Tools like Otter.ai and Ava provide live transcription, especially useful when lip-reading is obstructed, by offering features such as shared notes and instant captioning, thereby facilitating better engagement in various scenarios [344].

6.1. Study contributions and novelties

Our research investigates the limitations inherent in current text-based communication platforms designed for individuals with hearing impairments. While these tools proficiently support functional dialogue, they notably fall short in conveying emotional nuance and depth. We emphasize the untapped potential of affective computing and AI to improve emotion recognition and expression in digital communications, particularly noting the current methods' limited emotional range. By introducing a novel approach that combines text and audio analysis using the Bi-CAGNN model and analyzing the CMU-MOSI dataset, we offer a promising solution to enhance emotion interpretation for individuals with hearing impairments. This research moves towards bridging the gap in accurately conveying and interpreting emotional cues in text-based communication, aiming to develop more effective assistive technologies for the deaf and hard of hearing.

6.2. Literature review

In 2009, significant strides were made with Daassi-Gnaba and Jaime Lopez Krahe's creation of an artificial talking head to aid lip readers, combining speech and emotion recognition (ER) technologies [345]. The same year, Millett P. evaluated speech-to-text captioning's accuracy in educational settings for the deaf or HoH students, emphasizing the ongoing challenges and advancements [346]. Progressing to 2020, M. Shezi and A. Ade-Ibijola introduced Deaf Chat, a mobile app enhancing speech-to-text conversion for multiple speakers using advanced technologies [347]. By 2021, Alam F. et al. developed an integrated sign language translation system, marking a significant leap in communication tools for the deaf and HoH [342]. The year 2022 saw the introduction of the Deep Emotional Arousal Network (DEAN) for multimodal sentiment analysis and ER by Zhang et al. [348], alongside Arun Prasath G and Annapurna Panaiyappan's voice recognition system for the HoH community [349].

In 2024, Alsaadawi and Das demonstrated the critical role of GNNs in affective computing, enhancing emotion detection capabilities [19]. Additionally, the integration of Natural Language Processing NLP with ER technologies significantly advanced communication aids for the deaf or HoH, as highlighted in research utilizing models like BERT and RestNet50v2 [350][106]. Recently, the CA-WGNN model emerged as a breakthrough in emotion recognition, demonstrating superior performance in classifying emotions from multimodal data [24]. Additionally, S. Anithaelavarasi evaluated the

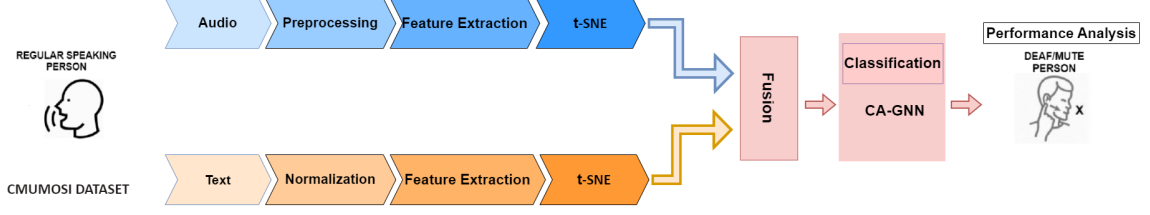


Figure 6.1. HoH: Analysis of Bi-CAGNN Algorithm.

effectiveness of various methodologies in recognizing MNIST Sign Language, stressing the significance of language and sentiment understanding for the HoH population [351]. In multimodal affective computing, A. Shahid et al. showed how t-SNE is instrumental in ER, particularly in evaluating emotional audio in synthesized speech [352].

6.3. Methodology and Implementation

In this section, we elaborate on our methodology, which is anchored in the principles of affective computing and focuses on a multimodal analysis framework. Utilizing the CMU-MOSI Dataset, comprising audio and textual data, Our methodology synthesizes insights drawn from seminal works in the field to refine computational models for emotion recognition.

Our analysis integrates extracted features into a single array for further examination or use in machine learning models. This integration is reflected in our classification algorithms, demonstrating a comprehensive approach to handling multimodal data. Figure 6.1. depicts the structure of the algorithm. Drawing on principles from [24], our algorithm is designed to convert text and audio into formats that are more accessible to the hard of hearing, using emotional cues to improve understanding. This approach underscores the importance of a multifaceted methodology in affective computing, especially for increasing accessibility and comprehension for those with hearing impairments.

6.3.1. Dataset

The CMU-MOSI (Multimodal Corpus of Sentiment Intensity) dataset [235], forms a comprehensive resource for research in multimodal sentiment analysis and emotion recognition. The CMU-MOSI dataset comprises 2,199 opinion video clips, each presenting an individual’s viewpoint. Sentiment labels in this dataset range from -3 (strongly negative) to 3 (strongly positive), accompanied by extensive annotations. These include subjectivity (objective vs. subjective opinions), sentiment intensity, visual features (annotated per frame and opinion), and detailed millisecond-level audio features. These pre-processed audio files are pivotal for extracting acoustic features, serving as inputs for deep-learning algorithms to classify emotions effectively. Expanding the research scope, the dataset encompasses over 23,500 sentence utterances from more than 1,000 YouTube

speakers, featuring video snippets from YouTube movie reviews annotated for sentiments and a range of emotions like happiness, sadness, anger, fear, disgust, and surprise. This dataset is particularly noteworthy for its combination of text, audio, and visual data, making it a rich resource for sentiment analysis, emotion recognition, and multimodal machine learning tasks.

Recent studies [235] highlight the significance of these datasets, particularly the CMU-MOSI, in advancing the Emotion Recognition (ER) field. The CMU-MOSI datasets, which incorporate both visual and auditory data, are crucial for their rich emotional content, making them essential for researchers in multimodal affective computing, natural language processing, and machine learning. These datasets are key for developing sophisticated emotion recognition models that interpret real-world human communication.

6.3.2. Emotion Classification Using BERT

After data preprocessing, the study leverages the BERT (Bidirectional Encoder Representations from Transformers) model, a preeminent model in the field of natural language processing, for emotion classification. The model, adept at understanding the nuances of language, categorizes each textual document into distinct emotions. This process showcases the application of pre-trained models in emotion analysis and underscores the model's adaptability across diverse linguistic datasets. BERT (Bidirectional Encoder Representations from Transformers) represents a seminal advancement in the domain of natural language processing NLP as introduced by [353]. This model significantly enhances machine understanding of human language by leveraging bidirectional training, wherein it simultaneously evaluates the context of each word in a sentence from both sides. Predicated on the Transformer architecture, which eschews traditional sequential processing methods for a mechanism focused on global dependencies between words, BERT is distinguished by its ability to grasp the nuanced meanings of words within their specific contexts.

The training methodology for BERT encompasses two stages: pre-training and fine-tuning. In the pre-training phase, the model is exposed to a vast corpus of text and learns language patterns through two unsupervised tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). These tasks aid the model in understanding language structure and flow. After pre-training, BERT can be fine-tuned with just an additional output layer, enabling its application to a wide array of NLP tasks, such as question answering and sentiment analysis, with minimal modifications. BERT's introduction has precipitated a paradigm shift in NLP, facilitating the development of models that achieve unprecedented accuracy in language understanding tasks. Its methodology of contextual analysis, pre-training on extensive text data, and adaptability to various tasks have set new benchmarks in the field. This breakthrough has paved the way for the emergence of numerous BERT-inspired models and underscored the model's versatility across different linguistic challenges, affirming its significant impact on advancing machine comprehension

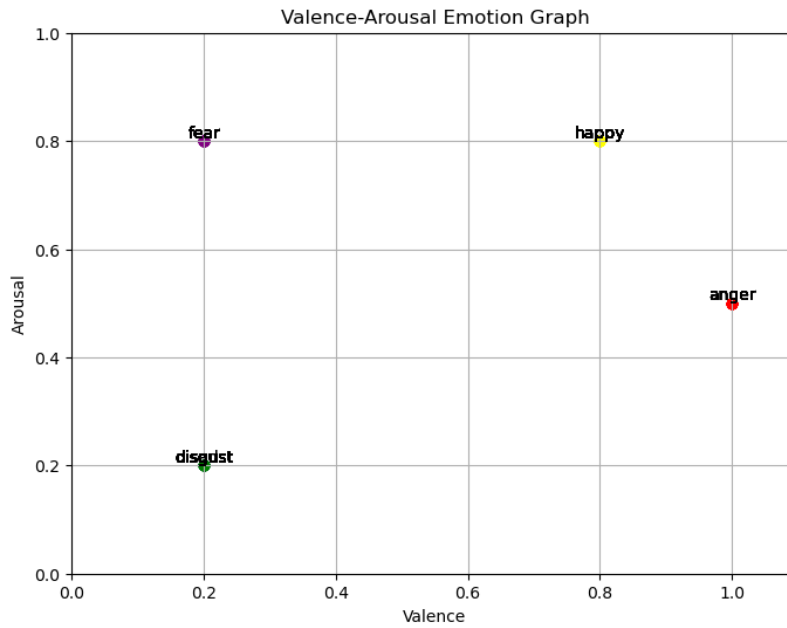


Figure 6.2. Text-Valence-Arousal Emotion Graph.

of human language [353].

6.3.3. Visualization and Dimensionality Reduction: Creating a Valence-Arousal Graph

This outlines the process of categorizing emotions, retrieving emotional scores, employing analysis techniques, and graphically representing emotional spectrums. The core of this computational analysis is establishing an *emotion_mapping* dictionary. This dictionary assigns discrete emotions, such as "anger", "happiness", and "sadness", to specific valence and arousal scores. Valence scores represent the emotion's positivity or negativity, and arousal scores indicate the emotion's intensity. These scores are based on a widely accepted understanding of emotional experiences, offering a standard approach to quantifying emotions in textual data. Following the mapping of emotions, the process involves scanning a text corpus to extract emotions from each document. Using the *emotion_mapping*, it assigns the relevant valence and arousal scores. These scores are collected into two arrays: *valence_scores* and *arousal_scores*, converting qualitative emotional expressions into quantitative data for analysis. This methodology enhances understanding of emotional content in text datasets through visualization and detailed analysis, improving digital communication for applications in sentiment analysis, psychology, and marketing. It makes the complex emotional content within large datasets more accessible, giving us a deeper understanding of human emotions. Figure 6.2. illustrates the distribution of valence and arousal scores as derived from our analysis.

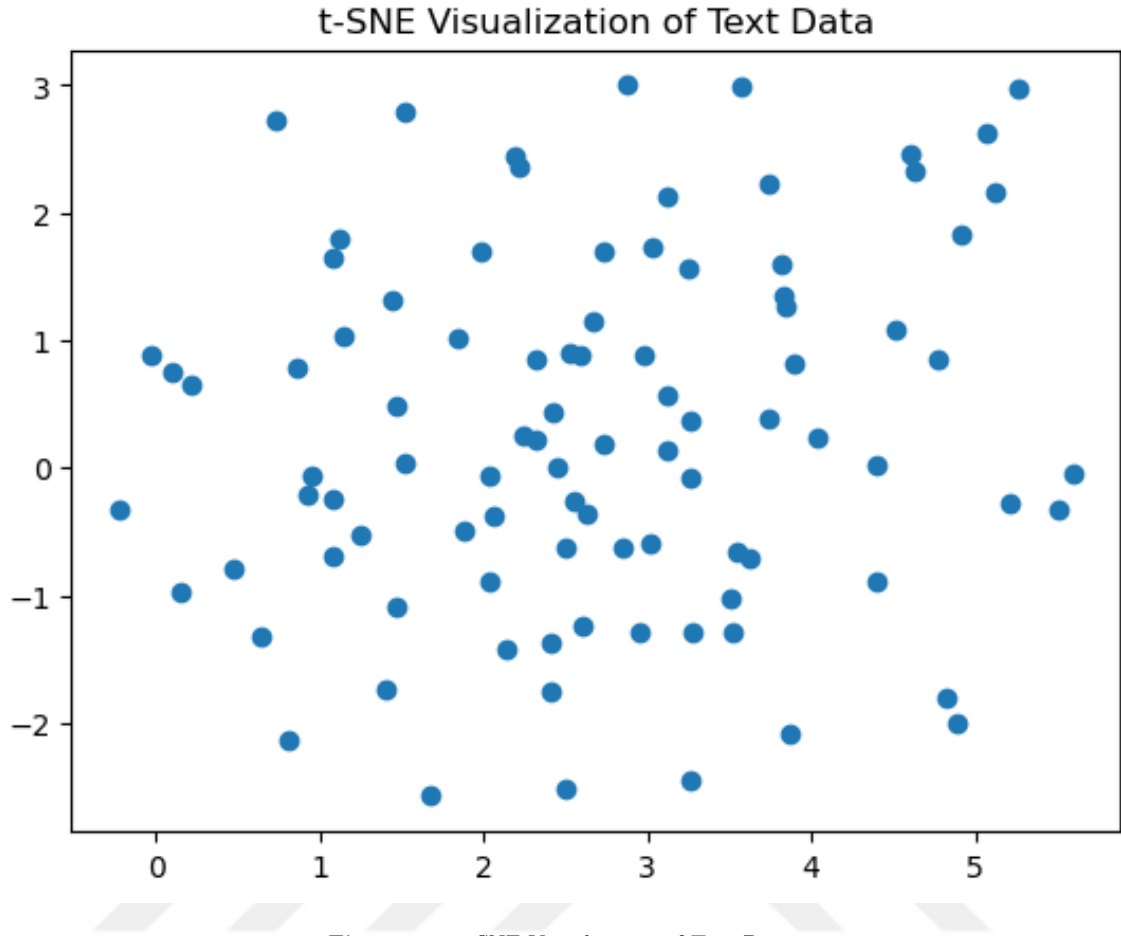


Figure 6.3. t-SNE Visualization of Text Data.

6.3.4. Dimensionality Reduction in Text Data Using t-SNE

The procedure for reducing the dimensionality of text data commences by encoding the textual information numerically through the Term Frequency-Inverse Document Frequency (TF-IDF) method. This step prepares the data for the subsequent reduction in dimensionality.

The t-distributed Stochastic Neighbor Embedding (t-SNE) technique is applied. Renowned for its capability to illustrate high-dimensional data in a comprehensible manner, t-SNE effectively models the similarities between data points as joint probabilities and strives to minimize the Kullback-Leibler divergence between the embeddings in the original high-dimensional space and the reduced-dimensional space. Despite its sensitivity to initial conditions and the non-convex nature of its cost function, which may lead to variable results, t-SNE unveils intricate structures within the data. Such revelations are particularly beneficial for the domain of affective computing, which seeks to decode human emotions through diverse channels. Specifically, t-SNE facilitates the discernment of intricate relationships among various emotional states, thereby augmenting our understanding of emotional expressions. Figure 6.3. displays the

dimensionality reduction of text data, illustrating the distribution and relationships within the dataset in a 2D space. The t-SNE method effectively preserves the data's local structures while simplifying its representation. In the context of affective computing, the efficacy of t-SNE can be attributed to several factors:

- **Perplexity:** This critical parameter influences the accuracy of the visualization by indicating the anticipated density of data around a point, thereby affecting the balance between local and global aspects of the data structure.
- **Data Quality:** The fidelity of the initial feature extraction be it from text, audio, or visual data—significantly impacts the resulting visualization. The presence of noise and irrelevant features may obscure the genuine categories of emotions.
- **Emotional Variability:** t-SNE's proficiency in clustering similar emotions while distinctly separating dissimilar ones underscores the diversity within emotional expressions, necessitating a refined interpretation of these variations.
- **Visualization and Hyperparameters:** Additional visualization techniques, such as the valence-arousal graph, and meticulous adjustment of t-SNE hyperparameters, including the learning rate and number of iterations, are crucial for achieving optimal visual representations.
- **Comparative Analysis:** While t-SNE facilitates the comparative study of emotional responses elicited by different stimuli, its stochastic nature can pose challenges to straightforward interpretation.

Through its ability to unveil latent structures within data, t-SNE significantly contributes to our comprehension of complex datasets, thereby propelling further analytical and modeling endeavors in emotion recognition, as it identifies semantic correlations within neural word embeddings [354].

6.3.5. Data Fusion and Feature Integration

Multimodal data fusion and analysis The cornerstone of this research lies in its multimodal approach, integrating audio data with textual analysis. Audio files are transcribed, and the transcriptions mirror the textual data processing steps. Using both textual and auditory data in parallel highlights the multifaceted nature of emotion recognition. Feature-level fusion, a pivotal technique in data integration, plays a significant role in multimodal affective computing as shown in Figure 6.4.. This method combines text, audio, and visual data into a unified feature set, starting with feature extraction using t-SNE for dimensionality reduction and visualization. The process involves validating and concatenating these features to form a comprehensive 'fused features' array. Feature-level fusion enhances machine learning efficacy by merging unique insights from multiple modalities, significantly improving predictive and classification

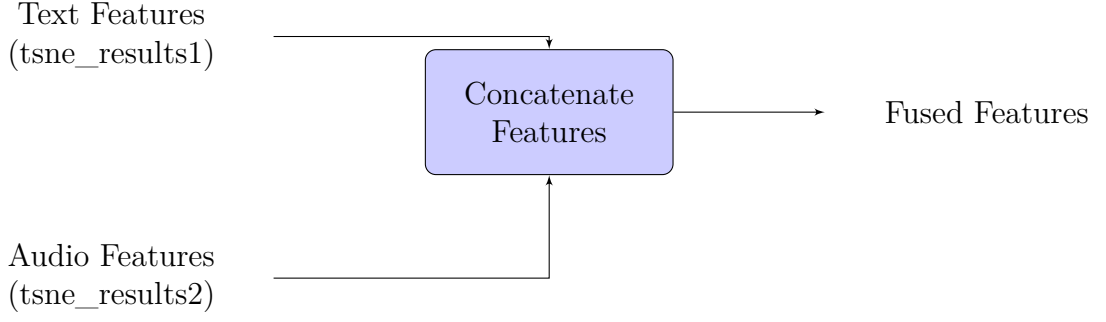


Figure 6.4. Block diagram illustrating the feature-level fusion process.

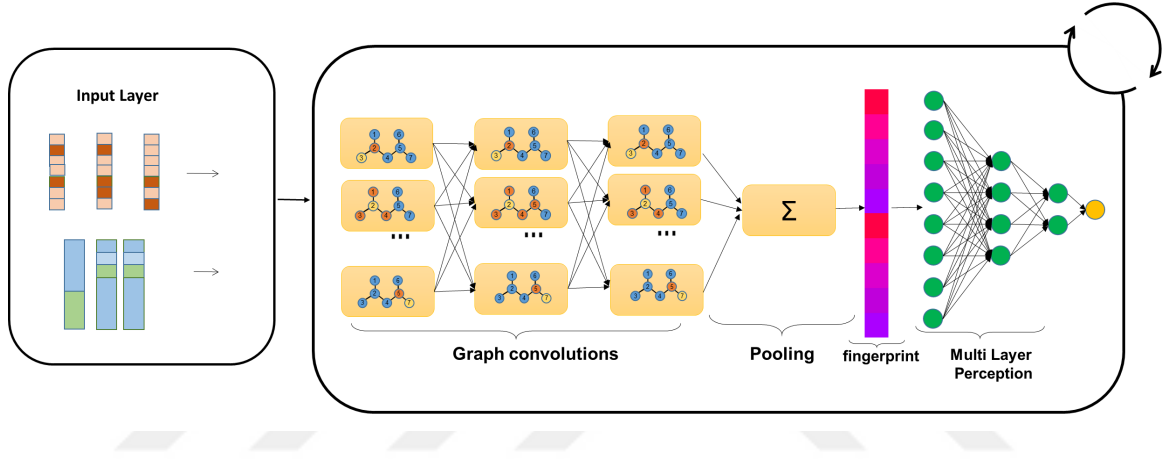


Figure 6.5. The CAGNN architecture [24].

model performance. It is crucial for effectively advancing multimodal data analysis and integrating diverse data sources.

6.3.6. Bi-CAGNN Approach for Emotion Classification

CA-GNN is a powerful machine learning model that integrates multimodal data with emotion identification. First, a neural network is built for binary classification, and then it is trained and evaluated. Cumulative Attribute Graph Neural Networks (CA-GNNs) detailed in Figure 6.5., aggregate attributes from a node's direct neighbors and cumulatively from nodes at greater distances at each layer. This aggregation can be weighted based on distance, node importance, or other graph metrics, allowing for a nuanced integration of attributes representing an advanced framework within the domain of graph neural networks (GNNs), specifically designed to enhance the learning of node representations by incorporating local and non-local structural and attribute information. This academic discussion delves into the core concepts, methodologies, and potential applications of CA-GNNs, offering a comprehensive understanding of their significance in

machine learning and network analysis.

Graph Neural Networks (GNNs) are a significant method for analyzing graph-structured data, useful in fields like social network analysis, bioinformatics, and recommendation systems. Traditional GNN models, like Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs), mainly gather features from direct neighbors to update node representations. However, these models often fall short in utilizing the full hierarchical structure of graphs, especially in recognizing long-range dependencies and incorporating global information.

Concepts:

CA-GNNs extend traditional GNN architectures by incorporating mechanisms to aggregate and integrate attributes from direct neighbors and nodes further away in the graph, effectively capturing cumulative attribute information. This approach enables the model to leverage a more comprehensive set of features, encompassing the immediate locality and broader structural contexts. The key components of CA-GNNs include:

- **Attribute Aggregation:** CA-GNNs aggregate attributes from both direct neighbors and distant nodes at each layer, with weightings based on distance or node significance, enabling nuanced attribute integration.
- **Feature Transformation:** CA-GNNs apply learnable transformations to node features to refine the aggregated data for better prediction accuracy.
- **Hierarchical Representation Learning:** By layering, CA-GNNs develop hierarchical representations that capture broader graph dynamics, essential for analyzing complex structures.

Applications:

CA-GNNs find applications across a wide array of domains where graph-structured data is prevalent:

- **Social Network Analysis:** Understanding community structures, influence propagation, and user behavior prediction.
- **Bioinformatics:** Protein function prediction, gene expression analysis, and biomolecular interaction modeling.
- **Recommendation Systems:** Leveraging user-item interaction graphs to improve recommendation accuracy and personalization.
- **Knowledge Graphs:** Entity resolution, relation prediction, and knowledge graph completion for enhancing semantic web applications.

Quadrant II
(-, +)

Quadrant I
(+, +)

Quadrant III
(-, -)

Quadrant IV
(+, -)

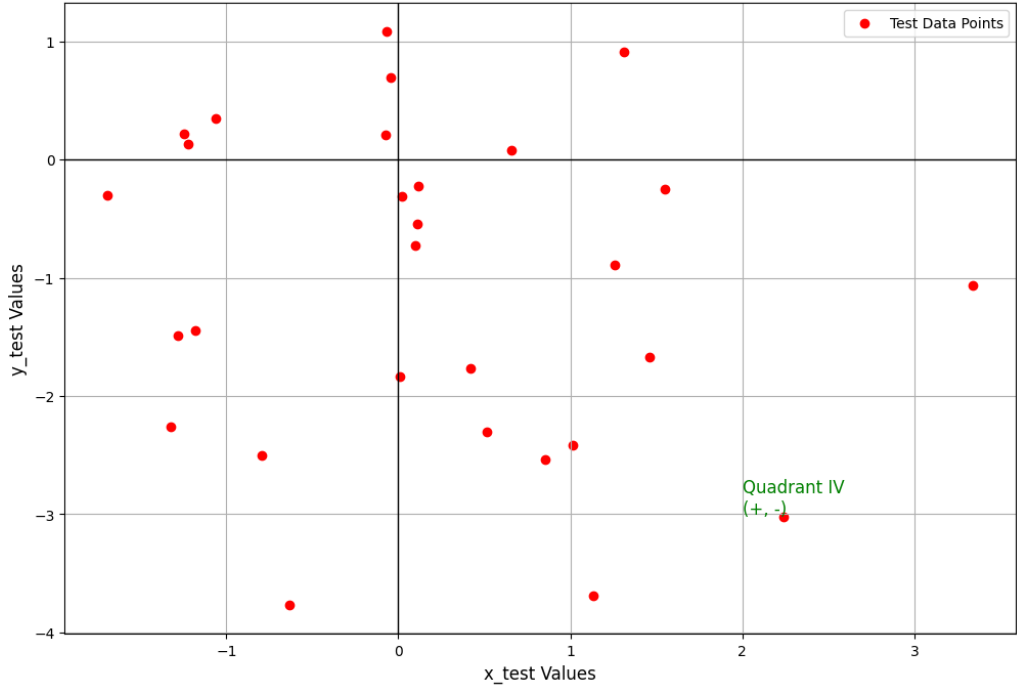


Figure 6.6. X,Y Plot of Test Data.

Cumulative Attribute Graph Neural Networks represent a significant advancement in graph-based learning, providing a robust framework for integrating comprehensive attribute information. CA-GNNs promise to unlock deeper insights and achieve superior performance across many complex graph structure applications through their enhanced representational capabilities. The scatter plot depicted in Figure 6.6. elucidates the relationship between two variables, x_{test} and y_{test} . Each red point on the plot signifies a pair of corresponding values from these arrays. The horizontal axis represents the x_{test} values, while the vertical axis corresponds to the y_{test} values. The data distribution, as shown, does not exhibit a distinct linear pattern, indicating an absence of a straightforward linear relationship between the variables.

This is further evidenced by the spread of data points across all four quadrants of the plot. Specifically, Quadrant I (top right) contains points where both x_{test} and y_{test} are positive, Quadrant II (top left) with negative x_{test} but positive y_{test} , Quadrant III (bottom left) where both are negative, and Quadrant IV (bottom right) with positive x_{test} and negative y_{test} . This variability indicates a diverse dataset and suggests that a simple linear model may not be sufficient for accurate prediction. Additionally, the size of the training data,

Table 6.1. Model Performance Metrics

Metric Name	Value (%)
Accuracy	0.82
Precision	0.8224
Recall	0.82
F1-Score	0.8202

denoted as $x_{\text{train}}.\text{shape}$ and $y_{\text{train}}.\text{shape}$, is reported as (114,), indicating that each training dataset comprises 114 samples. This size is a crucial aspect in determining the learning capability of the model, where generally, a larger dataset is preferable for enhanced model performance.

6.4. Experimental Results and Evaluation

This study introduces the Bi-CAGNN algorithm, an advanced ER system enhancing text and audio analysis for the HoH community. It accurately classifies nuanced emotions and can be customized to meet specific research and application needs. The table delineates the model’s performance in classifying data into Negative, Neutral, and Positive categories. The model achieves an overall accuracy of 82%, indicating a high proficiency level in correctly identifying the class labels of unseen data. This accuracy is substantiated by the precision and recall metrics, which also stand at 82%, illustrating the model’s balanced ability to minimize false positives and false negatives across classes. As seen Table 6.1., the F1 score, a harmonic mean of precision and recall, mirrors these metrics closely with a value of 0.8202, underscoring the model’s consistent performance across both dimensions of precision and recall. This balance is crucial in applications where both the reduction of false positives and the capture of true positives are equally important. A closer examination of the classification report reveals nuanced performance differences across the categories. The model exhibits the highest precision (0.86) in identifying Negative instances, albeit with a slightly lower recall (0.78), suggesting a conservative approach in classifying instances as Negative. Conversely, while having the lowest precision (0.76), the Positive class enjoys the highest recall (0.83), indicating the model’s propensity to more liberally classify instances as Positive, albeit at the cost of a higher false positive rate. As seen Table 6.2., the Neutral category demonstrates the most balanced performance with closely matched precision (0.84) and recall (0.85) metrics, leading to the highest F1 score (0.85) among the three classes. This suggests that the model is particularly adept at identifying genuinely neutral instances, with a good balance between avoiding false positives and capturing true positives. The model may need to be tuned to improve precision and recall, and perhaps more sophisticated methods should be used.

Table 6.2. Detailed Classification Report

Sentiment	Precision	Recall	F1 Score
Negative	0.86	0.78	0.82
Neutral	0.84	0.85	0.85
Positive	0.76	0.83	0.79
Average/Total	0.82	0.82	0.82

The Bi-CAGNN algorithm enhances emotional recognition (ER) by integrating audio and text analysis, significantly aiding the hard-of-hearing (HoH) community. It innovatively captures emotions during audio-to-text conversion, improving traditional interpretation methods by offering a nuanced textual representation of spoken emotions. Performance assessments reveal the algorithm’s proficiency in sentiment classification, achieving an 82% accuracy. However, its tendency to over-identify positive sentiments suggests the need for refinement to balance precision and recall across different emotional categories, particularly by mitigating false positives in positive sentiment detection without compromising its overall recall capabilities. This research employs multimodal affective

Table 6.3. Consolidated Emotion Recognition Performance on CMU-MOSI Dataset

References	Authors	Year	Approach	Accurate(%)
[81]	Chudasama et al.	2022	M2FNET	66.52
[355]	Han et al.	2021	BBFN	86.20
[356]	Mai et al.	2022	MIB	84.40
[357]	Rahman et al.	2020	MAG-BERT	72.16
[358]	Si et al.	2024	SPACE+WavLM	20.90
This study	Alsaadawi et al.	2024	Bi-CAGNNs	82.00

computing to aid individuals with hearing impairments, leveraging the CMU-MOSI Dataset to develop a model that accurately classifies emotional states from text and audio. By utilizing advanced feature extraction, dimensionality reduction through t-SNE, and the Bi-CAGNN model, this study not only provides enhanced emotional comprehension for those with hearing loss but also opens new avenues in ER technology by suggesting potential adaptations for diverse languages and integrating visual data to enrich emotional recognition. Moreover, translating audio to text is outlined as a pivotal advancement in communication technologies, greatly benefiting accessibility, comprehension, and content analysis. It facilitates multitasking and extends content reach globally through easier translation into multiple languages. These capabilities are vital in educational, professional, and legal contexts, providing indispensable benefits across various scenarios. Future research could explore its extension to various linguistic contexts and possibly

integrate visual data to enhance its multimodal capabilities further. Using affective computing, assistive technologies could be revolutionized, reducing misunderstandings and improving communication for the hearing impaired.

This study represents a significant leap in multimodal affective computing, highlighting the field's dynamic evolution. The core of the research focused on applying the Bi-CAGNN algorithm for effective emotion recognition from both vocal and textual inputs, aiming to benefit the deaf community significantly. Utilizing the CMU-MOSI Dataset, the study analyzed the algorithm's performance across various metrics, including accuracy, precision, recall, and F1 score, achieving commendable outcomes. However, the research also illuminated challenges, such as class imbalances, pointing toward the necessity for continued model refinement. An integral aspect of the study was the utilization of the Valence-Arousal Emotion Graph, providing a sophisticated visual representation of emotional states. This innovative approach deepens our understanding of the complex interplay between valence and arousal in emotional responses and underscores the Bi-CAGNN's potential to revolutionize digital communication through precise emotion classification. The research delineates future pathways for enhancing the model, incorporating visual data, and broadening its application spectrum, thereby setting a new course for affective computing research and its societal implications.

7. CONCLUSION

This thesis explores the forefront application of GNNs in artificial intelligence, specifically focusing on enhancing emotion recognition technologies through the innovative CA-WGNN algorithm. As global scientific and technological advancements accelerate, comprehending human emotions has become increasingly vital for human-to-human interactions and human-computer interfaces. The complexities introduced by globalization significantly impact decision-making processes across various domains, including politics, medicine, education, security, and economics. The study of emotion recognition is indispensable for gaining deeper insights into nuanced human responses, critical for improving interaction dynamics in an interconnected world. Multimodal emotion recognition, which integrates multiple forms of data (text, audio, visual), is particularly crucial for achieving a more precise and comprehensive analysis of emotional states. Graph neural networks represent a novel and appropriate approach for the classification tasks involved in emotion recognition due to their proficiency in processing and modeling relational data. The structural nature of GNNs allows them to effectively capture the intricate patterns of emotional cues distributed across different communication channels. By leveraging the unique capabilities of GNNs, this research endeavors to expand the boundaries of machine comprehension of human emotions. Consequently, this contributes to developing more empathetic and effective technological solutions, addressing the challenges posed by globalization. In conclusion, the CA-WGNN algorithm demonstrates significant potential in advancing emotion recognition technologies. By integrating multimodal data and utilizing the strengths of graph neural networks, this research offers promising directions for future developments in artificial intelligence aimed at better understanding and responding to human emotions. This work underscores the importance of continued exploration and innovation in this field to foster more sophisticated and human-centric technological advancements. Our review uncovers several exciting avenues for future research in trimodal affective computing, each holding the key to advancing the field. We propose to delve into the following: Advancements in technology, such as exploring cutting-edge machine learning technologies like Graph Neural Networks, to sharpen the accuracy and efficiency of emotion detection. Overcoming cross-modality hurdles by weaving together diverse data forms, including text, audio, visual, and physiological signals, to create more coherent affective computing systems. Developing real-time applications for the immediate processing of multimodal data in dynamic and critical decision-making environments. Emphasizing research on ethics and privacy to ensure data security and personal privacy in emotion recognition. Exploring the adaptation of these technologies to different cultural contexts and languages, considering the variance in emotional expressions. Investigating long-term studies on these systems' performance and adaptability through continuous user interaction for evolving, personalized solutions. Researching the potential of merging trimodal affective computing

with emerging technologies, such as augmented reality, virtual reality, and the Internet of Things (IoT), to open new realms of application. These focus areas promise to catapult advancements in trimodal affective computing, leading to more sophisticated, accurate, and adaptable emotion recognition systems. Such progress will enhance technological capabilities and deepen our comprehension of human emotions in a digital context.



REFERENCES

- [1] B. Pan, K. Hirota, Z. Jia, Y. Dai, A review of multimodal emotion recognition: From datasets, preprocessing, features, to fusion methods, *Neurocomputing* 561 (2023) 126866. doi:10.1016/j.neucom.2023.126866.
URL <https://www.sciencedirect.com/science/article/pii/S092523122300989X>
- [2] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, Y. Zong, A survey of deep learning-based multimodal emotion recognition: Speech, text, and face, *Entropy* 25 (10) (2023) 1440. doi:10.3390/e25101440.
URL <https://www.mdpi.com/1099-4300/25/10/1440>
- [3] C. Chen, H. Hong, J. Guo, B. Song, Inter-intra modal representation augmentation with trimodal collaborative disentanglement network for multimodal sentiment analysis, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023) 1476–1488. doi:10.1109/TASLP.2023.3263801.
URL <https://ieeexplore.ieee.org/document/10089492>
- [4] A. Chaddad, Y. Wu, R. Kateb, A. Bouridane, Electroencephalography signal processing: A comprehensive review and analysis of methods and techniques, *Sensors* 23 (14) (2023) 6434. doi:10.3390/s23146434.
URL <https://www.mdpi.com/1424-8220/23/14/6434>
- [5] W. Alsabhan, Human–computer interaction with a real-time speech emotion recognition with ensembling techniques 1d convolution neural network and attention, *Sensors* 23 (2023). doi:10.3390/s23031386.
- [6] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, *Information Fusion* 91 (2023) 424–444. doi:10.1016/j.inffus.2022.09.025.
URL <https://www.sciencedirect.com/science/article/pii/S1566253522001634>
- [7] J. Gondohanindijo, Muljono, E. Noersasongko, Pujiono, D. R. M. Setiadi, Multi-features audio extraction for speech emotion recognition based on deep learning, *International Journal of Advanced Computer Science and Applications (IJACSA)* 14 (6) (2023). doi:10.14569/IJACSA.2023.0140623.
URL <https://thesai.org/Publications/ViewPaper?Volume=14&Issue=6&Code=IJACSA&SerialNo=23>
- [8] K. R. Scherer, J. Sundberg, B. Fantini, S. Trznadel, F. Eyben, The expression of emotion in the singing voice: Acoustic patterns in vocal performance, *The Journal of the Acoustical Society of America* 142 (4) (2017) 1805–1815.
- [9] S. Poria, E. Cambria, A. Hussain, G.-B. Huang, Towards an intelligent framework for multimodal affective data analysis, *Neural Networks* 63 (2015) 104–116.
- [10] K. Fridkin, S. A. Gershon, Nothing more than feelings? how emotions affect attitude change during the 2016 general election debates, *Political Communication* 38 (4) (2021) 370–387.
- [11] I. Broch-Due, H. L. Kjaerstad, L. V. Kessing, K. Miskowiak, Subtle behavioural responses during negative emotion reactivity and down-regulation in bipolar disorder: A facial expression and eye-tracking study, *Psychiatry Research* 266 (2018) 152–159.
- [12] E. B. Setiawan, Song recommendation application using speech emotion recognition, *IJID (International Journal on Informatics for Development)* 10 (1) (2021) 15–22.

- [13] A. Agarwal, D. Toshniwal, Application of lexicon based approach in sentiment analysis for short tweets, in: 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), 2018, pp. 189–193.
- [14] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, *Information Fusion* 37 (2017) 98–125.
- [15] E. A. Clark, J. Kessinger, S. E. Duncan, M. A. Bell, J. Lahne, D. L. Gallagher, S. F. O’Keefe, The facial action coding system for characterization of human affective response to consumer product-based stimuli: A systematic review, *Frontiers in Psychology* 11 (2020) 920.
- [16] S. Gu, F. Wang, J. A. Bourgeois, J. H. Huang, N. P. Patel, A model for basic emotions using observations of behavior in drosophila, *Frontiers in Psychology* 10 (2019) 781.
- [17] L. Barrett, R. Adolphs, S. Marsella, A. Martinez, S. Pollak, Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements, *Psychological Science in the Public Interest* 20 (1) (2019) 1–68.
- [18] M. Taboada, Sentiment analysis: An overview from linguistics, *Annual Review of Linguistics* 2 (2016) 325–347.
- [19] H. Alsaadawi, R. Das, Multimodal emotion recognition using bi-lg-gcn for the meld dataset, *Balkan Journal of Electrical and Computer Engineering (BAJECE)* 12 (1) (2024).
- [20] A. Kennedy, D. Inkpen, Sentiment classification of movie reviews using contextual valence shifters, *Computational Intelligence* 22 (2) (2006) 110–125.
- [21] D. Datcu, L. J. Rothkrantz, Emotion recognition using bimodal data fusion, in: *Proceedings of the 12th International Conference on Computer Systems and Technologies*, 2011, pp. 122–128.
- [22] X. Wang, X. Chen, C. Cao, Human emotion recognition by optimally fusing facial expression and speech feature, *Signal Processing: Image Communication* 84 (2020) 115831.
- [23] B. T. Atmaja, K. Shirai, M. Akagi, Speech emotion recognition using speech feature and word embedding, in: *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 519–523.
- [24] H. F. T. Al-Saadawi, R. Das, TER-CA-WGNN: Trimodel Emotion Recognition Using Cumulative Attribute-Weighted Graph Neural Network, *Applied Sciences* 14 (6) (2024) 2252, number: 6 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/app14062252.
URL <https://www.mdpi.com/2076-3417/14/6/2252>
- [25] T. Zhang, Z. Tan, Deep emotion recognition using facial, speech and textual cues: A survey (2021). doi:10.36227/techrxiv.15184302.v2.
URL https://www.techrxiv.org/articles/preprint/Deep_Emotion_Recognition_in_Dynamic_Data_using_Facial_Speech_and_Textual_Cues_A_Survey/15184302/2
- [26] T. Dalgleish, M. Power, *Handbook of Cognition and Emotion*, John Wiley & Sons, 2000-11-21, google-Books-ID: vsLvrhohXhAC.
URL https://www.google.com.tr/books/edition/Handbook_of_Cognition_and_Emotion/vsLvrhohXhAC?hl=en&gbpv=1&dq=isbn:9780470842218&printsec=frontcover&pli=1
- [27] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, *Information Fusion* 37 (2017) 98–125. doi:10.1016/j.inffus.2017.02.003.
URL <https://www.sciencedirect.com/science/article/pii/S1566253517300738>
- [28] J. Tao, T. Tan, *Affective Computing: A Review*, Springer, 2005, pages: 995. doi:10.1007/11573548_125.

- [29] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, K. Karpouzis, The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data, in: A. C. R. Paiva, R. Prada, R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science*, Springer, 2007, pp. 488–500. doi:10.1007/978-3-540-74889-2_43.
- [30] Y. Susanto, A. G. Livingstone, B. C. Ng, E. Cambria, The hourglass model revisited, *IEEE Intelligent Systems* 35 (5) (2020) 96–102, conference Name: IEEE Intelligent Systems. doi:10.1109/MIS.2020.2992799.
- [31] D. C. Rubin, J. M. Talarico, A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words, *Memory* 17 (8) (2009) 802–808, publisher: Routledge _eprint: <https://doi.org/10.1080/09658210903130764>. doi:10.1080/09658210903130764.
URL <https://doi.org/10.1080/09658210903130764>
- [32] J. A. Russell, Core affect and the psychological construction of emotion, *Psychological Review* 110 (2003) 145–172, place: US Publisher: American Psychological Association. doi:10.1037/0033-295X.110.1.145.
- [33] P. Kumar, M. Vardhan, Plutchik wheel of emotion and machine learning-based hybrid sentiment analysis for the hindi language with minimum dependency on high computation resources, *SN Computer Science* 4 (6) (2023) 797. doi:10.1007/s42979-023-02237-7.
URL <https://doi.org/10.1007/s42979-023-02237-7>
- [34] E. Cambria, D. Hazarika, S. Poria, A. Hussain, R. B. V. Subramanyam, Benchmarking multimodal sentiment analysis, in: A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science*, Springer International Publishing, 2018, pp. 166–179. doi:10.1007/978-3-319-77116-8_13.
- [35] S. Poria, E. Cambria, N. Howard, G.-B. Huang, A. Hussain, Fusing audio, visual and textual clues for sentiment analysis from multimodal content, *Neurocomputing* 174 (2016) 50–59. doi:10.1016/j.neucom.2015.01.095.
URL <https://www.sciencedirect.com/science/article/pii/S0925231215011297>
- [36] M. Jeon, Chapter 1 - emotions and affect in human factors and human–computer interaction: Taxonomy, theories, approaches, and methods, in: M. Jeon (Ed.), *Emotions and Affect in Human Factors and Human-Computer Interaction*, Academic Press, 2017, pp. 3–26. doi:10.1016/B978-0-12-801851-4.00001-X.
URL <https://www.sciencedirect.com/science/article/pii/B978012801851400001X>
- [37] Z. Sun, P. Sarma, W. Sethares, Y. Liang, Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, *Proceedings of the Association for the Advancement of Artificial Intelligence AAAI Conference on Artificial Intelligence* 34 (5) (2020) 8992–8999, number: 05. doi:10.1609/aaai.v34i05.6431.
URL <https://ojs.aaai.org/index.php/AAAI/article/view/6431>
- [38] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, M. Pantic, A survey of multimodal sentiment analysis, *Image and Vision Computing* 65 (2017) 3–14. doi:10.1016/j.imavis.2017.08.003.
URL <https://www.sciencedirect.com/science/article/pii/S0262885617301191>
- [39] Z. Ahmad, N. Khan, A survey on physiological signal based emotion recognition (2022). *arXiv:2205.10466[cs]*, doi:10.48550/arXiv.2205.10466.
URL <http://arxiv.org/abs/2205.10466>

- [40] M. F. H. Siddiqui, P. Dhakal, X. Yang, A. Y. Javaid, 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) (2022) 47, number: 6 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/mti6060047.
URL <https://www.mdpi.com/2414-4088/6/6/47>
- [41] J. Zhang, Z. Yin, P. Chen, S. Nichele, Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review, *Information Fusion* 59 (2020) 103–126. doi:10.1016/j.inffus.2020.01.011.
URL <https://www.sciencedirect.com/science/article/pii/S1566253519302532>
- [42] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, X. Zhao, Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects, *Expert Systems with Applications* 237 (2024-03-01) 121692. doi:10.1016/j.eswa.2023.121692.
URL <https://www.sciencedirect.com/science/article/pii/S0957417423021942>
- [43] T. Praveen Kumar, B. Vardhan, A review on multi-model sentiment analysis using deep learning for text, speech, & emoji reorganization, *Harbin Gongye Daxue Xuebao/Journal of Harbin Institute of Technology* 54 (2022) 2022. doi:10.11720/JHIT.54042022.8.
- [44] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, W. Zhang, A systematic review on affective computing: emotion models, databases, and recent advances, *Information Fusion* 83-84 (2022) 19–52. doi:10.1016/j.inffus.2022.03.009.
URL <https://www.sciencedirect.com/science/article/pii/S1566253522000367>
- [45] P. Bhatt, A. Sethi, V. Tasgaonkar, J. Shroff, I. Pendharkar, A. Desai, P. Sinha, A. Deshpande, G. Joshi, A. Rahate, P. Jain, R. Walambe, K. Kotecha, N. K. Jain, Machine learning for cognitive behavioral analysis: datasets, methods, paradigms, and research directions, *Brain Informatics* 10 (1) (2023-07-31) 18. doi:10.1186/s40708-023-00196-6.
- [46] P. Savci, B. Das, Comparison of pre-trained language models in terms of carbon emissions, time and accuracy in multi-label text classification using AutoML, *Heliyon* 9 (5) (2023) e15670. doi:10.1016/j.heliyon.2023.e15670.
URL <https://www.sciencedirect.com/science/article/pii/S2405844023028773>
- [47] M. Aydogan, A hybrid deep neural network-based automated diagnosis system using x-ray images and clinical findings, *International Journal of Imaging Systems and Technology* 33 (4) (2023) 1368–1382, _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/ima.22856>. doi:10.1002/ima.22856.
URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/ima.22856>
- [48] D. Dupré, E. G. Krumhuber, D. Küster, G. J. McKeown, A performance comparison of eight commercially available automatic classifiers for facial affect recognition, *PLOS ONE* 15 (4) (2020) e0231968, publisher: Public Library of Science. doi:10.1371/journal.pone.0231968.
URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0231968>
- [49] E. Cameron, M. Green, *Making Sense of Change Management: A Complete Guide to the Models, Tools and Techniques of Organizational Change*, Kogan Page Publishers, 2019.
URL <https://www.example.com/your-book-url>
- [50] W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, T. R. Gadekallu, Cross corpus multi-lingual speech emotion recognition using ensemble learning, *Complex & Intelligent Systems* 7 (4) (2021) 1845–1854. doi:10.1007/s40747-020-00250-4.
URL <https://doi.org/10.1007/s40747-020-00250-4>
- [51] M. Imani, G. A. Montazer, A survey of emotion recognition methods with emphasis on E-Learning environments, *Journal of Network and Computer Applications* 147 (2019) 102423. doi:10.1016/j.jnca.2019.102423.
URL <https://www.sciencedirect.com/science/article/pii/S1084804519302759>

- [52] S. K. Yadav, K. Tiwari, H. M. Pandey, S. A. Akbar, A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions, *Knowledge-Based Systems* 223 (2021) 106970. doi:10.1016/j.knosys.2021.106970.
URL <https://www.sciencedirect.com/science/article/pii/S0950705121002331>
- [53] R. Das, M. Soylu, A key review on graph data science: The power of graphs in scientific studies, *Chemometrics and Intelligent Laboratory Systems* 240 (2023-09-15) 104896. doi:10.1016/j.chemolab.2023.104896.
URL <https://www.sciencedirect.com/science/article/pii/S0169743923001466>
- [54] P. Savci, B. Das, Prediction of the customers' interests using sentiment analysis in e-commerce data for comparison of arabic, english, and turkish languages, *Journal of King Saud University - Computer and Information Sciences* 35 (3) (2023-03-01) 227–237. doi:10.1016/j.jksuci.2023.02.017.
URL <https://www.sciencedirect.com/science/article/pii/S131915782300054X>
- [55] I. Pulatov, R. Oteniyazov, F. Makhmudov, Y.-I. Cho, Enhancing speech emotion recognition using dual feature extraction encoders, *Sensors* 23 (14) (2023-01) 6640, number: 14 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/s23146640.
URL <https://www.mdpi.com/1424-8220/23/14/6640>
- [56] M. Egger, M. Ley, S. Hanke, Emotion recognition from physiological signal analysis: A review, *Electronic Notes in Theoretical Computer Science* 343 (2019) 35–55. doi:10.1016/j.entcs.2019.04.009.
URL <https://www.sciencedirect.com/science/article/pii/S157106611930009X>
- [57] L. Qu, C. Weber, S. Wermter, Multimodal target speech separation with voice and face references (2020). arXiv:2005.08335[cs, eess], doi:10.48550/arXiv.2005.08335.
URL <http://arxiv.org/abs/2005.08335>
- [58] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2018, pp. 2236–2246. doi:10.18653/v1/P18-1208.
URL <https://aclanthology.org/P18-1208>
- [59] Q. Portes, J. M. Carvalho, J. Pinquier, F. Lerasle, Multimodal neural network for sentiment analysis in embedded systems, in: *Proceedings of the 16th International Conference on Computer Vision Theory and Applications (VISAPP 2021)*, Vol. 5: VISAPP, SCITEPRESS: Science and Technology Publications, 2017, p. 387. doi:10.5220/0010224703870398.
URL <https://hal.science/hal-03445482>
- [60] P. P. Liang, Z. Liu, A. Bagher Zadeh, L.-P. Morency, Multimodal language analysis with recurrent multistage fusion, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2017, pp. 150–161. doi:10.18653/v1/D18-1014.
URL <https://aclanthology.org/D18-1014>
- [61] M. G. Huddar, S. S. Sannakki, V. S. Rajpurohit, Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM, *Multimedia Tools and Applications* 80 (9) (2021) 13059–13076. doi:10.1007/s11042-020-10285-x.
URL <https://doi.org/10.1007/s11042-020-10285-x>
- [62] J. Wu, T. Zhu, X. Zheng, C. Wang, Multi-modal sentiment analysis based on interactive attention mechanism, *Applied Sciences* 12 (16) (2022) 8174, number: 16 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/app12168174.
URL <https://www.mdpi.com/2076-3417/12/16/8174>

- [63] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, S. Poria, Multimodal sentiment analysis using hierarchical fusion with context modeling, *Knowledge-Based Systems* 161 (2018) 124–133. doi:10.1016/j.knosys.2018.07.041.
URL <https://www.sciencedirect.com/science/article/pii/S0950705118303897>
- [64] A. Aslam, A. B. Sargano, Z. Habib, Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks, *Applied Soft Computing* 144 (2023-09-01) 110494. doi:10.1016/j.asoc.2023.110494.
URL <https://www.sciencedirect.com/science/article/pii/S1568494623005124>
- [65] J. Peng, T. Wu, W. Zhang, F. Cheng, S. Tan, F. Yi, Y. Huang, A fine-grained modal label-based multi-stage network for multimodal sentiment analysis, *Expert Systems with Applications* 221 (2023-07-01) 119721. doi:10.1016/j.eswa.2023.119721.
URL <https://www.sciencedirect.com/science/article/pii/S0957417423002221>
- [66] T. Wu, J. Peng, W. Zhang, H. Zhang, S. Tan, F. Yi, C. Ma, Y. Huang, Video sentiment analysis with bimodal information-augmented multi-head attention, *Knowledge-Based Systems* 235 (2022) 107676. doi:10.1016/j.knosys.2021.107676.
URL <https://www.sciencedirect.com/science/article/pii/S0950705121009382>
- [67] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos (2016). arXiv:1606.06259[cs], doi:10.48550/arXiv.1606.06259.
URL <http://arxiv.org/abs/1606.06259>
- [68] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, L.-P. Morency, Multi-level multiple attentions for contextual multimodal sentiment analysis, in: 2017 IEEE International Conference on Data Mining (ICDM), 2017, pp. 1033–1038, ISSN: 2374-8486. doi:10.1109/ICDM.2017.134.
- [69] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional MKL based multimodal emotion recognition and sentiment analysis, in: 2016 IEEE 16th International Conference on Data Mining (ICDM), 2016-12, pp. 439–448, ISSN: 2374-8486. doi:10.1109/ICDM.2016.0055.
- [70] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, K. Yang, CH-SIMS: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2017, pp. 3718–3727. doi:10.18653/v1/2020.acl-main.343.
URL <https://aclanthology.org/2020.acl-main.343>
- [71] S. Dham, A. Sharma, A. Dhall, Depression scale recognition from audio, visual and text analysis (2017-09-18). arXiv:1709.05865[cs], doi:10.48550/arXiv.1709.05865.
URL <http://arxiv.org/abs/1709.05865>
- [72] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal, P. Bhattacharyya, Multi-task learning for multi-modal emotion recognition and sentiment analysis (2019-05-14). arXiv:1905.05812[cs], doi:10.48550/arXiv.1905.05812.
URL <http://arxiv.org/abs/1905.05812>
- [73] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, L.-P. Morency, Words can shift: Dynamically adjusting word representations using nonverbal behaviors, *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence* 33 (1) (2019-07-17) 7216–7223, number: 01. doi:10.1609/aaai.v33i01.33017216.
URL <https://ojs.aaai.org/index.php/AAAI/article/view/4706>
- [74] D. Zhang, W. Zhang, S. Li, Q. Zhu, G. Zhou, Modeling both Intra- and Inter-modal Influence for Real-Time Emotion Detection in Conversations, *Conference on Multimedia (MM '20)*, 2020, pages: 511. doi:10.1145/3394171.3413949.

- [75] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, D. Manocha, M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues, *Proceedings of the Association for the Advancement of Artificial Intelligence AAAI Conference on Artificial Intelligence* 34 (2) (2020) 1359–1367, number: 02. doi:10.1609/aaai.v34i02.5492. URL <https://ojs.aaai.org/index.php/AAAI/article/view/5492>
- [76] Y. Mao, Q. Sun, G. Liu, X. Wang, W. Gao, X. Li, J. Shen, DialogueTRM: Exploring the intra- and inter-modal emotional behaviors in the conversation (2020). arXiv:2010.07637[cs], doi:10.48550/arXiv.2010.07637. URL <http://arxiv.org/abs/2010.07637>
- [77] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.-M. Meßner, E. Cambria, G. Zhao, B. W. Schuller, The MuSe 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress, in: *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge, MuSe '21*, Association for Computing Machinery, 2021, pp. 5–14. doi:10.1145/3475957.3484450. URL <https://doi.org/10.1145/3475957.3484450>
- [78] B. Xie, M. Sidulova, C. H. Park, Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion, *Sensors* 21 (14) (2021-01) 4913, number: 14 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/s21144913. URL <https://www.mdpi.com/1424-8220/21/14/4913>
- [79] M. C. Caschera, P. Grifoni, F. Ferri, Emotion classification from speech and text in videos using a multimodal approach, *Multimodal Technologies and Interaction* 6 (4) (2022) 28, number: 4 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/mti6040028. URL <https://www.mdpi.com/2414-4088/6/4/28>
- [80] S. Xing, S. Mai, H. Hu, Adapted dynamic memory network for emotion recognition in conversation, *IEEE Transactions on Affective Computing* 13 (3) (2022) 1426–1439, conference Name: IEEE Transactions on Affective Computing. doi:10.1109/TAFFC.2020.3005660.
- [81] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, N. Onoe, M2fnet: Multi-modal fusion network for emotion recognition in conversation, in: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 2022, pp. 4651–4660. doi:10.1109/CVPRW56347.2022.00511. URL <https://ieeexplore.ieee.org/document/9857351/>
- [82] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2017, pp. 873–883. doi:10.18653/v1/P17-1081. URL <https://aclanthology.org/P17-1081>
- [83] J. Liang, R. Li, Q. Jin, Semi-supervised multi-modal emotion recognition with cross-modal distribution matching, in: *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2852–2861. arXiv:2009.02598[cs,eess], doi:10.1145/3394171.3413579. URL <http://arxiv.org/abs/2009.02598>
- [84] Y. Jiang, W. Li, M. S. Hossain, M. Chen, A. Alelaiwi, M. Al-Hammadi, A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition, *Information Fusion* 53 (2020) 209–221. doi:10.1016/j.inffus.2019.06.019. URL <https://www.sciencedirect.com/science/article/pii/S1566253519301381>
- [85] S. Tang, Z. Luo, G. Nan, J. Baba, Y. Yoshikawa, H. Ishiguro, Fusion with hierarchical graphs for multimodal emotion recognition, in: *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 1288–1296, ISSN: 2640-0103. doi:10.23919/APSIPAASC55919.2022.9979932.

- [86] P. Kumar, S. Malik, B. Raman, Hybrid fusion based interpretable multimodal emotion recognition with insufficient labelled data (2022). [arXiv:2208.11450\[cs\]](#), doi:10.48550/arXiv.2208.11450.
URL <http://arxiv.org/abs/2208.11450>
- [87] S. Mai, Y. Zeng, S. Zheng, H. Hu, Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis, *IEEE Transactions on Affective Computing* 14 (3) (2023-07) 2276–2289, conference Name: IEEE Transactions on Affective Computing. doi:10.1109/TAFFC.2022.3172360.
URL <https://ieeexplore.ieee.org/document/9767560>
- [88] C.-H. Wu, W.-B. Liang, Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels, *T. Affective Computing* 2 (2011) 10–21. doi:10.1109/T-AFFC.2010.16.
URL <https://ieeexplore.ieee.org/document/5674019>
- [89] E. S. Salama, R. A. El-Khoribi, M. E. Shoman, M. A. W. Shalaby, A 3d-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition, *Egyptian Informatics Journal* 22 (2) (2021) 167–176.
URL <https://www.sciencedirect.com/science/article/pii/S1110866520301389>
- [90] W. Liu, J.-L. Qiu, W.-L. Zheng, B.-L. Lu, Multimodal emotion recognition using deep canonical correlation analysis (2019). doi:10.48550/arXiv.1908.05349.
URL <http://arxiv.org/abs/1908.05349>
- [91] S. Nemati, R. Rohani, M. E. Basiri, M. Abdar, N. Y. Yen, V. Makarenkov, A hybrid latent space data fusion method for multimodal emotion recognition, *IEEE Access* 7 (2019) 172948–172964, conference Name: IEEE Access. doi:10.1109/ACCESS.2019.2955637.
URL <https://ieeexplore.ieee.org/document/8911364>
- [92] L. Cai, Y. Hu, J. Dong, S. Zhou, Audio-Textual Emotion Recognition Based on Improved Neural Networks, *Mathematical Problems in Engineering* 2019 (2019) 1–9. doi:10.1155/2019/2593036.
- [93] Q.-T. Truong, H. Lauw, VistaNet: Visual aspect attention network for multimodal sentiment analysis, *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (2019-07-17) 305–312. doi:10.1609/aaai.v33i01.3301305.
URL <https://doi.org/10.1609/aaai.v33i01.3301305>
- [94] Z. Lian, J. Tao, B. Liu, J. Huang, Z. Yang, R. Li, Context-dependent domain adversarial neural network for multimodal emotion recognition., in: *Interspeech*, 2020, pp. 394–398.
URL https://www.iscaspeech.org/archive/interspeech_2020/lian20b_interspeech.html
- [95] M. Aydoğan, A. Karci, Improving the accuracy using pre-trained word embeddings on deep neural networks for turkish text classification, *Physica A: Statistical Mechanics and its Applications* 541 (2020-03) 123288. doi:10.1016/j.physa.2019.123288.
URL <https://linkinghub.elsevier.com/retrieve/pii/S0378437119318436>
- [96] N. J. Shoumy, L.-M. Ang, K. P. Seng, D. M. M. Rahaman, T. Zia, Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals, *Journal of Network and Computer Applications* 149 (2020) 102447. doi:10.1016/j.jnca.2019.102447.
URL <https://www.sciencedirect.com/science/article/pii/S1084804519303078>
- [97] D. Priyasad, T. Fernando, S. Denman, C. Fookes, S. Sridharan, Attention driven fusion for multi-modal emotion recognition (2020). [arXiv:2009.10991\[cs,eess,stat\]](#), doi:10.48550/arXiv.2009.10991.
URL <http://arxiv.org/abs/2009.10991>

- [98] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, D. Manocha, Emoticon: Context-aware multimodal emotion recognition using frege's principle, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14233–14242.
URL <https://ieeexplore.ieee.org/document/9156904>
- [99] X. Wu, W.-L. Zheng, B.-L. Lu, Investigating EEG-based functional connectivity patterns for multimodal emotion recognition (2020). [arXiv:2004.01973\[cs\]](https://arxiv.org/abs/2004.01973), doi:10.48550/arXiv.2004.01973.
URL <http://arxiv.org/abs/2004.01973>
- [100] L. Sun, Z. Lian, J. Tao, B. Liu, M. Niu, Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism, in: Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop, MuSe'20, Association for Computing Machinery, 2020-10-15, pp. 27–34. doi:10.1145/3423327.3423672.
URL <https://doi.org/10.1145/3423327.3423672>
- [101] M. R. Makiuchi, K. Uto, K. Shinoda, Multimodal emotion recognition with high-level speech and text features (2021). [arXiv:2111.10202\[cs, eess\]](https://arxiv.org/abs/2111.10202), doi:10.48550/arXiv.2111.10202.
URL <http://arxiv.org/abs/2111.10202>
- [102] S. R. Zaman, D. Sadekeen, M. A. Alfaz, R. Shahriyar, One source to detect them all: Gender, age, and emotion detection from voice, in: 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), 2021, pp. 338–343, ISSN: 0730-3157. doi:10.1109/COMPSAC51774.2021.00055.
URL <https://ieeexplore.ieee.org/document/9529731>
- [103] Z. Fang, A. He, Q. Yu, B. Gao, W. Ding, T. Zhang, L. Ma, FAF: A novel multimodal emotion recognition approach integrating face, body and text (2022). [arXiv:2211.15425\[cs\]](https://arxiv.org/abs/2211.15425), doi:10.48550/arXiv.2211.15425.
URL <http://arxiv.org/abs/2211.15425>
- [104] S. K. Bharti, S. Varadhaganapathy, R. K. Gupta, P. K. Shukla, M. Bouye, S. K. Hingaa, A. Mahmoud, Text-based emotion recognition using deep learning approach, Computational Intelligence and Neuroscience 2022 (2022) e2645381, publisher: Hindawi. doi:10.1155/2022/2645381.
URL <https://www.hindawi.com/journals/cin/2022/2645381/>
- [105] M. Soyly, A. Soyly, R. Das, A new approach to recognizing the use of attitude markers by authors of academic journal articles, Expert Systems with Applications 230 (2023-11) 120538. doi:10.1016/j.eswa.2023.120538.
URL <https://linkinghub.elsevier.com/retrieve/pii/S0957417423010400>
- [106] U. Rehman, A. Sohail, Exploring The Role of Verbal and Non-Verbal Communication Skills of Nursing in Patients Healing, Research Square, 2023. doi:10.21203/rs.3.rs-3376906/v1.
- [107] Q. Jin, C. Li, S. Chen, H. Wu, Speech emotion recognition with acoustic and lexical features, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 4749–4753, iSSN: 2379-190X. doi:10.1109/ICASSP.2015.7178872.
URL <https://ieeexplore.ieee.org/document/7178872>
- [108] N. R. Bhowmik, M. Arifuzzaman, M. R. H. Mondal, Sentiment analysis on bangla text using extended lexicon dictionary and deep learning algorithms, Array 13 (2022) 100123. doi:10.1016/j.array.2021.100123.
URL <https://www.sciencedirect.com/science/article/pii/S259000562100059X>

- [109] M. Maithri, U. Raghavendra, A. Gudigar, J. Samanth, P. D. Barua, M. Murugappan, Y. Chakole, U. R. Acharya, Automated emotion recognition: Current trends and future perspectives, *Computer Methods and Programs in Biomedicine* 215 (2022) 106646. doi:10.1016/j.cmpb.2022.106646.
URL <https://www.sciencedirect.com/science/article/pii/S0169260722000311>
- [110] A. Onan, Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification, *Journal of King Saud University - Computer and Information Sciences* 34 (5) (2022) 2098–2117. doi:10.1016/j.jksuci.2022.02.025.
URL <https://www.sciencedirect.com/science/article/pii/S1319157822000696>
- [111] O. Araque, S. Frenda, R. Sprugnoli, D. Nozza, V. Patti, EMit at EVALITA 2023: Overview of the categorical emotion detection in italian social media task, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, Vol. 3473 of CEUR Workshop Proceedings, CEUR, 2023, pp. 1613–0073.
URL <https://ceur-ws.org/Vol-3473/#1>
- [112] K. Lutsai, C. H. Lampert, Geolocation predicting of tweets using bert-based models (May 2023). arXiv:2303.07865[cs], doi:10.48550/arXiv.2303.07865.
URL <http://arxiv.org/abs/2303.07865>
- [113] W. Meng, Y. Wei, P. Liu, Z. Zhu, H. Yin, Aspect based sentiment analysis with feature enhanced attention cnn-bilstm, *IEEE Access PP* (2019) 1–1. doi:10.1109/ACCESS.2019.2952888.
- [114] X. Li, K. Mao, F. Lin, Z. Feng, Feature-aware conditional gan for category text generation, *Neurocomputing* 547 (2023) 126352. doi:10.1016/j.neucom.2023.126352.
URL <https://www.sciencedirect.com/science/article/pii/S0925231223004757>
- [115] Y. Xiong, M. Nor, Y. Li, H. Guo, L. Dai, Reinforcement learning-based aspect term extraction using dilated convolutions and differential equation initialization, *International Journal of Advanced Computer Science and Applications* 14 (January 2023). doi:10.14569/IJACSA.2023.0140518.
- [116] H. A. A. Hafez, E. Alabdulkreem, Language-independent text tokenization using unsupervised deep learning, *Intelligent Automation & Soft Computing* 35 (1) (2022) 321–334. doi:10.32604/iasc.2023.026235.
URL <https://www.techscience.com/iasc/v35n1/48133>
- [117] M. Bashir, A. R. Javed, M. Arshad, T. R. Gadekallu, W. Shahzad, M. Baig, Context aware emotion detection from low resource urdu language using deep neural network, *ACM Transactions on Asian and Low-Resource Language Information Processing* 22 (3) (2022) Article 37. doi:10.1145/3528576.
- [118] M. F. Bashir, A. R. Javed, M. U. Arshad, T. R. Gadekallu, W. Shahzad, M. O. Beg, Context-aware emotion detection from low-resource urdu language using deep neural network, *ACM Transactions on Asian and Low-Resource Language Information Processing* 22 (5) (2023) 131:1–131:30. doi:10.1145/3528576.
URL <https://doi.org/10.1145/3528576>
- [119] Y. Xu, X. Li, H. Wang, H. Chang, A short text sentiment analysis model combined with attention mechanism of bidirectional affective words, *Journal of Computing and Electronic Information Management* 11 (1) (2023) 16–27. doi:10.54097/jceim.v11i1.9474.
URL <https://drpress.org/ojs/index.php/jceim/article/view/9474>

- [120] H. Singh, N. Helian, R. Adams, Y. Sun, Sentiment analysis using blstm-resnet on textual images, in: 2022 International Joint Conference on Neural Networks (IJCNN), 2022, pp. 1–8, ISSN: 2161-4407. doi:10.1109/IJCNN55064.2022.9892883.
URL <https://ieeexplore.ieee.org/document/9892883>
- [121] J. Wang, X. Li, X. An, Modeling multiple latent information graph structures via graph convolutional network for aspect-based sentiment analysis, *Complex & Intelligent Systems* 9 (4) (2023) 4003–4014. doi:10.1007/s40747-022-00940-1.
URL <https://doi.org/10.1007/s40747-022-00940-1>
- [122] D. Tian, J. Shi, J. Feng, A self-attention-based multi-level fusion network for aspect category sentiment analysis, *Cognitive Computation* 15 (4) (2023) 1372–1390. doi:10.1007/s12559-023-10160-5.
URL <https://doi.org/10.1007/s12559-023-10160-5>
- [123] N. Wang, X. Kong, B. Ren, L. Hao, B. Han, Seabil: Self-attention-weighted ultrashort-term deep learning prediction of ship maneuvering motion, *Ocean Engineering* 287 (2023) 115890. doi:10.1016/j.oceaneng.2023.115890.
URL <https://www.sciencedirect.com/science/article/pii/S0029801823022746>
- [124] I. Ameer, N. Bölücü, M. H. F. Siddiqui, B. Can, G. Sidorov, A. Gelbukh, Multi-label emotion classification in texts using transfer learning, *Expert Systems with Applications* 213 (2023) 118534. doi:10.1016/j.eswa.2022.118534.
URL <https://www.sciencedirect.com/science/article/pii/S0957417422016098>
- [125] S. Siswanto, Z. Mar'ah, A. Sabir, T. Hidayat, F. Adhel, W. Amni, Sentiment analysis using naive bayes with lexicon-based feature on tiktok application, *Jurnal Varian* 6 (2022) 85–92. doi:10.30812/varian.v6i1.2205.
- [126] Q. Jin, Genetic algorithm and support vector machine application in english text classification for intelligent teaching, *Soft Computing* (August 2023). doi:10.1007/s00500-023-09084-x.
URL <https://doi.org/10.1007/s00500-023-09084-x>
- [127] J. Attieh, J. Tekli, Supervised term-category feature weighting for improved text classification, *Knowledge-Based Systems* 261 (2023) 110215. doi:10.1016/j.knosys.2022.110215.
URL <https://www.sciencedirect.com/science/article/pii/S0950705122013119>
- [128] R. Catelli, S. Pelosi, M. Esposito, Lexicon-based vs. bert-based sentiment analysis: A comparative study in italian, *Electronics* 11 (3) (2022) 374. doi:10.3390/electronics11030374.
URL <https://www.mdpi.com/2079-9292/11/3/374>
- [129] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, X. Kong, Multimodal sentiment analysis based on fusion methods: A survey, *Information Fusion* 95 (February 2023). doi:10.1016/j.inffus.2023.02.028.
- [130] A. Diwali, K. Saeedi, K. Dashtipour, M. Gogate, E. Cambria, A. Hussain, Sentiment analysis meets explainable artificial intelligence: A survey on explainable sentiment analysis, *IEEE Transactions on Affective Computing* (2023) 1–12Conference Name: IEEE Transactions on Affective Computing. doi:10.1109/TAFFC.2023.3296373.
URL <https://ieeexplore.ieee.org/abstract/document/10185138>
- [131] C. Selvi, R. PushpaLaksmi, Sa-msvm: Hybrid heuristic algorithm-based feature selection for sentiment analysis in twitter, *Computer Systems Science and Engineering* 44 (2023) 2439–2456. doi:10.32604/csse.2023.029254.
- [132] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Computation* 9 (8) (1997) 1735–1780.

- [133] B. Wang, J. Zhang, Z. Huo, Y. Lin, W. Chen, Q. Li, Twitter sentiment analysis using deep convolutional neural network and lstm, in: Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2016, pp. 767–770.
- [134] M. Huang, Z. Xu, Y. Wu, L. Zhao, X. Dong, B. Zhou, Attention-based lstm for aspect-level sentiment classification, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI), AAAI Press, 2017, pp. 5010–5016.
- [135] Y. Tay, A. T. Luu, S. C. Hui, J. Su, Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis, in: Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI), AAAI Press, 2018, pp. 5956–5963.
- [136] A. Chatterjee, S. Poria, E. Cambria, S. Sivaprasad, A. Gelbukh, Multimodal sentiment analysis using hierarchical fusion with context modeling, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, 2019, pp. 2225–2235.
- [137] S. Zhang, J. Zhang, Y. Liu, X. Wu, H. Wang, Bi-directional lstm with attention mechanism and convolutional layer for sentiment analysis, in: Proceedings of the IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 1234–1241.
- [138] B. Yang, Y. Zhang, X. Li, L. Wang, J. Xu, Hybrid cnn-lstm model for sentiment analysis of social media texts, in: Proceedings of the 15th IEEE International Conference on Semantic Computing (ICSC), IEEE, 2021, pp. 123–130.
- [139] J. Kim, H. Lee, S. Park, J. Choi, H. Kim, Transformer-lstm model for emotion recognition in text, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2022, pp. 456–465.
- [140] S. Iyer, R. Gupta, A. Kumar, P. Sharma, R. Singh, Cnn-lstm ensemble for eeg-based emotion recognition, in: Proceedings of the 2023 International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2023, pp. 789–798.
- [141] A. Hosseini, S. Rahimi, M. Akbari, R. Sharif, N. Fazeli, Multimodal emotion recognition using sound, image, and text fusion, in: Proceedings of the 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE, 2024, pp. 456–465.
- [142] K. Anadkat, H. Diwanji, S. Modasiya, M. Mehta, Creation and analysis of multimodal emotion recognition corpus with indian actors, *Reliability: Theory & Applications* 18 (3) (2023) 269–283.
URL <https://cyberleninka.ru/article/n/creation-and-analysis-of-multimodal-emotion-recognition-corpus-with>
- [143] N. Za'im, F. Al-Dhief, M. Azman, M. Mohamed, N. Latiff, M. Mat Baki, The accuracy of an online sequential extreme learning machine in detecting voice pathology using the malaysian voice pathology database, *Journal of Otolaryngology - Head & Neck Surgery* 52 (2023). doi: 10.1186/s40463-023-00661-6.
URL <https://doi.org/10.1186/s40463-023-00661-6>
- [144] M. Mohana, P. Subashini, M. Krishnaveni, Emotion recognition from facial expression using hybrid cnn-lstm network, *International Journal of Pattern Recognition and Artificial Intelligence* 37 (8) (2023) 2356008. doi:10.1142/S0218001423560086.
URL <https://www.worldscientific.com/doi/10.1142/S0218001423560086>
- [145] M. Saumard, Enhancing speech emotions recognition using multivariate functional data analysis, *Big Data and Cognitive Computing* 7 (3) (2023) 146. doi:10.3390/bdcc7030146.
URL <https://www.mdpi.com/2504-2289/7/3/146>

- [146] M. Xu, F. Zhang, W. Zhang, Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset (2021).
URL <https://ieeexplore.ieee.org/document/9381872>
- [147] M. Rayhan Ahmed, S. Islam, A. K. M. Muzahidul Islam, S. Shatabda, An ensemble 1d-cnn-lstm-gru model with data augmentation for speech emotion recognition, *Expert Systems with Applications* 218 (2023) 119633. doi:10.1016/j.eswa.2023.119633.
URL <https://www.sciencedirect.com/science/article/pii/S0957417423001343>
- [148] K. Kaur, P. Singh, Comparison of various feature selection algorithms in speech emotion recognition, *AIUB Journal of Science and Engineering (AJSE)* 22 (2023) 125–131. doi:10.53799/ajse.v22i2.357.
URL <https://doi.org/10.53799/ajse.v22i2.357>
- [149] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, L.-P. Morency, YouTube movie reviews: Sentiment analysis in an audio-visual context, *IEEE Intelligent Systems* 28 (3) (2013) 46–53, conference Name: IEEE Intelligent Systems. doi:10.1109/MIS.2013.34.
- [150] Y. Gao, L. Wang, J. Liu, J. Dang, S. Okada, Adversarial domain generalized transformer for cross-corpus speech emotion recognition, *IEEE Transactions on Affective Computing* (2023) 1–12doi:10.1109/TAFFC.2023.3290795.
URL <https://ieeexplore.ieee.org/document/10168224>
- [151] M. Khan, A. El Saddik, F. S. Alotaibi, N. T. Pham, AAD-net: Advanced end-to-end signal processing system for human emotion detection & recognition using attention-based deep echo state network, *Knowledge-Based Systems* 270 (2023) 110525. doi:10.1016/j.knosys.2023.110525.
URL <https://www.sciencedirect.com/science/article/pii/S0950705123002757>
- [152] A.-H. Jo, K.-C. Kwak, Speech emotion recognition based on two-stream deep learning model using korean audio information, *Applied Sciences* 13 (4) (2023) 2167. doi:10.3390/app13042167.
URL <https://www.mdpi.com/2076-3417/13/4/2167>
- [153] A. Amjad, L. Khan, N. Ashraf, M. B. Mahmood, H.-T. Chang, Recognizing semi-natural and spontaneous speech emotions using deep neural networks, *IEEE Access* 10 (2022) 37149–37163. doi:10.1109/ACCESS.2022.3163712.
URL <https://ieeexplore.ieee.org/document/9745600>
- [154] A. Jamshed, B. Mallick, R. K. Bharti, An efficient pattern mining convolution neural network (cnn) algorithm with grey wolf optimization (gwo), *The Imaging Science Journal* 70 (4) (2022) 238–252. doi:10.1080/13682199.2023.2166193.
URL <http://arxiv.org/abs/2204.04704>
- [155] A. S. Rao, A. P. Reddy, P. Vulpala, K. S. Rani, P. Hemalatha, Deep learning structure for emotion prediction using mfcc from native languages, *International Journal of Speech Technology* (2023). doi:10.1007/s10772-023-10047-8.
URL <https://doi.org/10.1007/s10772-023-10047-8>
- [156] S.-T. Pan, H.-J. Wu, Performance improvement of speech emotion recognition systems by combining 1d cnn and lstm with data augmentation, *Electronics* 12 (11) (2023) 2436. doi:10.3390/electronics12112436.
- [157] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, X. Zhao, Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and prospects, *Expert Systems with Applications* (2023) 121692doi:10.1016/j.eswa.2023.121692.

- [158] M. L. Ian Burnett, M. N. Stolar, Optimized multi-channel deep neural network with 2d graphical representation of acoustic speech features for emotion recognition, in: Proceedings of the 2014 8th International Conference on Signal Processing and Communication Systems (ICSPCS), 2014, pp. 1–6. doi:10.1109/ICSPCS.2014.7021083.
- [159] I. T. Kun Han, Dong Yu, Speech emotion recognition using deep neural network and extreme learning machine, INTERSPEECH 2014 (2014).
URL <https://www.microsoft.com/en-us/research/publication/speech-emotion-recognition-using-deep-neural-network-and-extreme-learning-machine>
- [160] C. Z. Seyedmahdad Mirsamadi, Emad Barsoum, Automatic speech emotion recognition using recurrent neural networks with local attention, in: Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2227–2231. doi:10.1109/ICASSP.2017.2227.
- [161] J. D. S. Ortega, M. Senoussaoui, E. Granger, M. Pedersoli, P. Cardinal, A. L. Koerich, Multimodal Fusion with Deep Neural Networks for Audio-Video Emotion Recognition, arXiv:1907.03196 [cs, eess] (2019). doi:10.48550/arXiv.1907.03196.
URL <http://arxiv.org/abs/1907.03196>
- [162] P. Heracleous, K. Takai, Y. Wang, K. Yasuda, A. Yoneyama, Y. Mohammad, An empirical study on feature extraction in dnn-based speech emotion recognition, Communications in Computer and Information Science (2020).
URL https://link.springer.com/chapter/10.1007/978-3-030-60700-5_40
- [163] H. Zhang, R. Gou, J. Shang, F. Shen, Y. Wu, G. Dai, Pre-trained deep convolution neural network model with attention for speech emotion recognition, Frontiers in Physiology 12 (2021) 643202. doi:10.3389/fphys.2021.643202.
- [164] L. Sun, Y. Huang, Q. Li, P. Li, Multi-classification speech emotion recognition based on two-stage bottleneck features selection and MCJD algorithm, Signal, Image and Video Processing 16 (5) (2022) 1253–1261. doi:10.1007/s11760-021-02076-0.
URL <https://doi.org/10.1007/s11760-021-02076-0>
- [165] E. Pranav, S. Kamal, C. Satheesh Chandran, M. Supriya, Facial Emotion Recognition Using Deep Convolutional Neural Network, in: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 317–320, iSSN: 2575-7288. doi:10.1109/ICACCS48705.2020.9074302.
URL <https://ieeexplore.ieee.org/document/9074302>
- [166] M. A. Adil, Facial emotion detection using convolutional neural networks (2021).
URL https://dspace.library.uvic.ca/bitstream/handle/1828/13388/Adil_Mohammed_Adnan_MEng_2021.pdf?sequence=3
- [167] R. Banerjee, S. De, S. Dey, A Survey on Various Deep Learning Algorithms for an Efficient Facial Expression Recognition System, International Journal of Image and Graphics 23 (03) (2023) 2240005, publisher: World Scientific Publishing Co. doi:10.1142/S0219467822400058.
URL <https://www.worldscientific.com/doi/10.1142/S0219467822400058>
- [168] S. k. Sana, G. Sruthi, D. Suresh, G. Rajesh, G. V. Subba Reddy, Facial emotion recognition based music system using convolutional neural networks, Materials Today: Proceedings 62 (2022) 4699–4706. doi:10.1016/j.matpr.2022.03.131.
URL <https://www.sciencedirect.com/science/article/pii/S2214785322014547>
- [169] H. M. Shahzad, S. M. Bhatti, A. Jaffar, S. Akram, M. Alhajlah, A. Mahmood, Hybrid facial emotion recognition using CNN-based features, Applied Sciences 13 (9) (2023) 5572, number: 9 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/app13095572.
URL <https://www.mdpi.com/2076-3417/13/9/5572>

- [170] D. Qu, S. Dhakal, D. Carrillo, Facial Emotion Recognition using CNN in PyTorch, arXiv:2312.10818 [cs] (Dec. 2023). doi:10.48550/arXiv.2312.10818.
URL <http://arxiv.org/abs/2312.10818>
- [171] K. N. Kumar Tataji, M. N. Kartheek, M. V. N. K. Prasad, Cc-cnn: A cross connected convolutional neural network using feature level fusion for facial expression recognition, Multimedia Tools and Applications (2023). doi:10.1007/s11042-023-16433-3.
URL <https://doi.org/10.1007/s11042-023-16433-3>
- [172] R. Singh, S. Saurav, T. Kumar, R. Saini, A. Vohra, S. Singh, Facial expression recognition in videos using hybrid cnn & convlstm, International Journal of Information Technology 15 (4) (2023) 1819–1830. doi:10.1007/s41870-023-01183-0.
URL <https://doi.org/10.1007/s41870-023-01183-0>
- [173] E. S. Smitha, S. Sendhilkumar, G. S. Mahalakshmi, Ensemble Convolution Neural Network for Robust Video Emotion Recognition Using Deep Semantics, Scientific Programming 2023 (2023) e6859284, publisher: Hindawi. doi:10.1155/2023/6859284.
URL <https://www.hindawi.com/journals/sp/2023/6859284/>
- [174] K. Sarvakar, R. Senkamalavalli, S. Raghavendra, J. Santosh Kumar, R. Manjunath, S. Jaiswal, Facial emotion recognition using convolutional neural networks, Materials Today: Proceedings 80 (2023) 3560–3564. doi:10.1016/j.matpr.2021.07.297.
URL <https://www.sciencedirect.com/science/article/pii/S2214785321051567>
- [175] A. Soultana, F. Benabbou, N. Sael, Enhanced CNN-Based Model for Facial Emotions Recognition in Smart Car Applications, Arabian Journal for Science and Engineering (Dec. 2023). doi:10.1007/s13369-023-08507-z.
URL <https://doi.org/10.1007/s13369-023-08507-z>
- [176] E. Bahk, M. Kaya, B. Kaya, A CNN-LSTM based approach for image captioning, in: 7th IET Smart Cities Symposium (SCS 2023), Vol. 2023, 2023, pp. 585–588. doi:10.1049/icp.2024.1019.
URL <https://ieeexplore.ieee.org/abstract/document/10510687>
- [177] Y. Yang, W. Hu, H. Lin, H. Hu, Robust cross-domain pseudo-labeling and contrastive learning for unsupervised domain adaptation nir-vis face recognition, IEEE Transactions on Image Processing 32 (2023) 5231–5244. doi:10.1109/TIP.2023.3309110.
URL <https://ieeexplore.ieee.org/abstract/document/10250166>
- [178] Z. Peng, H. Wu, Z. Song, H. Xu, X. Zhu, J. He, H. Liu, Z. Fan, Emotalk: Speech-driven emotional disentanglement for 3d face animation (2023). doi:10.48550/arXiv.2303.11089.
URL <http://arxiv.org/abs/2303.11089>
- [179] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, R. Horaud, Audio-visual speech enhancement using conditional variational auto-encoders, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 1788–1800. doi:10.1109/TASLP.2020.3000593.
URL <http://arxiv.org/abs/1908.02590>
- [180] F. Amro, H. Purohit, Integrated content-network analysis to discover influential collectives for studying social cyber-threats from online social movements, Social Network Analysis and Mining 13 (1) (2023) 120. doi:10.1007/s13278-023-01124-6.
URL <https://doi.org/10.1007/s13278-023-01124-6>
- [181] X.-B. Nguyen, C. N. Duong, X. Li, S. Gauch, H.-S. Seo, K. Luu, Micron-bert: Bert-based facial micro-expression recognition (2023). arXiv:2304.03195[cs], doi:10.48550/arXiv.2304.03195.
URL <http://arxiv.org/abs/2304.03195>

- [182] N. Singh, R. Kapoor, Facial micro-expression recognition (fmer) using model compression, in: 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET), 2022, pp. 1–4. doi:10.1109/ICECET55527.2022.9872920.
URL <https://ieeexplore.ieee.org/document/9872920>
- [183] F. Zhang, T. Zhang, Q. Mao, C. Xu, Joint pose and expression modeling for facial expression recognition, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 3359–3368. doi:10.1109/CVPR.2018.00354.
URL <https://ieeexplore.ieee.org/document/8578452>
- [184] A. F. Claret, K. R. Casali, T. S. Cunha, M. C. Moraes, Automatic classification of emotions based on cardiac signals: A systematic literature review, *Annals of Biomedical Engineering* 51 (11) (2023) 2393–2414. doi:10.1007/s10439-023-03341-8.
- [185] P. Zhang, F. Li, R. Zhao, R. Zhou, L. Du, Z. Zhao, X. Chen, Z. Fang, Real-time psychological stress detection according to ECG using deep learning, *Applied Sciences* 11 (9) (2021) 3838, number: 9 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/app11093838.
URL <https://www.mdpi.com/2076-3417/11/9/3838>
- [186] M. Alruwaili, R. Alruwaili, U. A. Kumar, A. M. Albarrak, N. H. Ali, R. Basri, Human emotion recognition based on brain signal analysis using fuzzy neural network, *Soft Computing* (2023). doi:10.1007/s00500-023-08224-7.
URL <https://doi.org/10.1007/s00500-023-08224-7>
- [187] J. Zhu, C. Jiang, J. Chen, Z. Zhang, X. Li, Y. Yang, Eeg based depression recognition using improved graph convolutional neural network, *Computers in Biology and Medicine* 148 (2022) 105815. doi:10.1016/j.combiomed.2022.105815.
URL <https://www.sciencedirect.com/science/article/pii/S0010482522005765>
- [188] D. W. Prabowo, H. A. Nugroho, N. A. Setiawan, J. Debayle, A systematic literature review of emotion recognition using eeg signals, *Cognitive Systems Research* 82 (2023) 101152. doi:10.1016/j.cogsys.2023.101152.
URL <https://www.sciencedirect.com/science/article/pii/S1389041723000803>
- [189] N. Kumari, S. Anwar, V. Bhattacharjee, A comparative analysis of machine and deep learning techniques for eeg evoked emotion classification, *Wireless Personal Communications* 128 (2023) 2869–2890. doi:10.1007/s11277-022-10076-7.
URL <https://doi.org/10.1007/s11277-022-10076-7>
- [190] M. Alalhareth, S.-C. Hong, An improved mutual information feature selection technique for intrusion detection systems in the internet of medical things, *Sensors* 23 (2023) 4971. doi:10.3390/s23104971.
URL <https://www.mdpi.com/1424-8220/23/10/4971>
- [191] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, Neural message passing for quantum chemistry, in: *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 1263–1272.
- [192] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: *International Conference on Learning Representations (ICLR)*, 2017, pp. 1–14.
- [193] K. Xu, W. Hu, J. Leskovec, S. Jegelka, How powerful are graph neural networks?, in: *International Conference on Learning Representations (ICLR)*, 2019, pp. 1–17. doi:10.48550/arXiv.1810.00826.
URL <https://arxiv.org/abs/1810.00826>
- [194] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, M. Sun, Graph neural networks: A review of methods and applications, *AI Open* 1 (2020) 57–81. doi:10.1016/j.aiopen.2021.09.001.
URL <https://www.sciencedirect.com/science/article/pii/S2666651021000120>

- [195] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: AAAI Conference on Artificial Intelligence, 2018, pp. 7444–7452.
- [196] R. Herzig, E. Levi, H. Xu, E. Brosh, A. Globerson, T. Darrell, Classifying collisions with spatio-temporal action graph networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2019, pp. 2347–2356.
URL https://openaccess.thecvf.com/content_ICCVW_2019/html/ADW/Herzig_Spatio-Temporal_Action_Graph_Networks_ICCVW_2019_paper.html
- [197] Z. Wang, L. Zheng, Y. Li, S. Wang, Linkage based face clustering via graph convolution network, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1117–1125.
- [198] C. R. Qi, H. Su, K. Mo, L. J. Guibas, Pointnet: Deep learning on point sets for 3d classification and segmentation, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 652–660.
- [199] D. Teney, L. Liu, A. van den Hengel, Graph-structured representations for visual question answering, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3233–3241.
- [200] O. Ashual, L. Wolf, Specifying object attributes and relations in interactive scene generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4561–4569.
URL https://openaccess.thecvf.com/content_ICCV_2019/html/Ashual_Specifying_Object_Attributes_and_Relations_in_Interactive_Scene_Generation_ICCV_2019_paper.html
- [201] S. Tripathi, S. N. Sridhar, S. Sundaresan, H. Tang, Compact scene graphs for layout composition and patch retrieval, in: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 676–683.
URL https://openaccess.thecvf.com/content_CVPRW_2019/html/CV-CMD/Tripathi_Compact_Scene_Graphs_for_Layout_Composition_and_Patch_Retrieval_CVPRW_2019_paper.html
- [202] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: A review of methods and applications, *AI open* 1 (2020) 57–81.
- [203] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation, *arXiv:1908.11540 [cs]* (Aug. 2019). doi:10.48550/arXiv.1908.11540.
URL <http://arxiv.org/abs/1908.11540>
- [204] J. Li, X. Wang, G. Lv, Z. Zeng, GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation (2023). *arXiv:2208.00339 [cs]*, doi:10.48550/arXiv.2208.00339.
URL <http://arxiv.org/abs/2208.00339>
- [205] E. Mansouri-Benssassi, J. Ye, Synch-graph: Multisensory emotion recognition through neural synchrony via gcns, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI Press, New York, New York, USA, 2020, pp. 1351–1358. doi:10.1609/aaai.v34i02.5491.
URL <https://aaai.org/ojs/index.php/AAAI/article/view/5491>
- [206] Z. Jia, Y. Lin, X. Xie, C. Chen, Hetemotionnet: Two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition, *CoRR abs/2108.03354* (2021).
URL <https://arxiv.org/abs/2108.03354>
- [207] H. Chen, T. Mao, A. Zadeh, S. Poria, E. Cambria, L. Shen, Dialoguernn: An attentive rnn for emotion detection in conversations, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27 (11) (2019) 1925–1939. doi:10.1109/TASLP.2019.2938252.

- [208] Y. Ren, Y. Wu, X. Liang, L. Lin, Graphmft: A graph network based multimodal fusion technique for emotion recognition, *IEEE Transactions on Affective Computing* 12 (2) (2019) 365–377. doi:10.1109/TAFFC.2019.2929036.
- [209] Z. Lian, J. Tao, B. Liu, J. Huang, Z. Yang, R. Li, Conversational emotion recognition using self-attention mechanisms and graph neural networks, in: *INTERSPEECH*, 2020, pp. 2347–2351. doi:10.21437/Interspeech.2020-1703.
- [210] P. Shenoy, R. Jyothi, B. Patra, R. Sahay, S. Agarwal, V. Reddy, S. Mohta, M. Shrivastava, Decn: Dialogical emotion correction network for conversational emotion recognition, *IEEE Transactions on Neural Networks and Learning Systems* 31 (10) (2020) 3958–3969. doi:10.1109/TNNLS.2020.2998470.
- [211] Z. Zhang, Y. Lei, L. Liu, H. Wang, Y. Shen, J. Wu, Aspect sentiment analysis with heterogeneous graph neural network, *IEEE Transactions on Knowledge and Data Engineering* 33 (9) (2021) 3215–3227. doi:10.1109/TKDE.2020.2969595.
- [212] Z. Zhang, J. Wu, Y. Lei, L. Liu, Y. Shen, H. Wang, Multi-modal fusion network for emotion recognition, *IEEE Transactions on Multimedia* 22 (1) (2020) 179–190. doi:10.1109/TMM.2020.2977048.
- [213] Y. Hu, M. Chen, L. Zhao, Z. Zhang, X. Li, J. Xu, Cogmen: Contextualized gnn based multimodal emotion recognition, *IEEE Transactions on Affective Computing* 12 (4) (2021) 987–998. doi:10.1109/TAFFC.2021.3063625.
- [214] W. Liu, J. Wu, X. Li, Y. Shen, H. Wang, Z. Zhang, Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations, *IEEE Transactions on Multimedia* 24 (2022) 1070–1082. doi:10.1109/TMM.2022.3145953.
- [215] M. H. R. Pereira, F. L. C. Pádua, A. C. M. Pereira, F. Benevenuto, D. H. Dalip, Fusing Audio, Textual and Visual Features for Sentiment Analysis of News Videos, *arXiv:1604.02612 [cs]* (Apr. 2016). doi:10.48550/arXiv.1604.02612. URL <http://arxiv.org/abs/1604.02612>
- [216] Y. Li, K. Zhang, J. Wang, X. Gao, A cognitive brain model for multimodal sentiment analysis based on attention neural networks, *Neurocomputing* 430 (2021-03-21) 159–173. doi:10.1016/j.neucom.2020.10.021. URL <https://www.sciencedirect.com/science/article/pii/S0925231220315174>
- [217] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, A database of German emotional speech, Vol. 5, Conference: *INTERSPEECH 2005 - Eurospeech*, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005, 2005-09-04, journal Abbreviation: 9th European Conference on Speech Communication and Technology Pages: 1520 Publication Title: 9th European Conference on Speech Communication and Technology. doi:10.21437/Interspeech.2005-446. URL 10.21437/Interspeech.2005-446
- [218] I. Sneddon, M. McRorie, G. McKeown, J. Hanratty, The belfast induced natural emotion database, *IEEE Transactions on Affective Computing* 3 (1) (2012-01) 32–41, conference Name: *IEEE Transactions on Affective Computing*. doi:10.1109/T-AFFC.2011.26. URL <https://ieeexplore.ieee.org/document/5975142>
- [219] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, <http://immortal.multicomp.cs.cmu.edu/IEMOCAP/>, accessed on June 6, 2023 (2008).
- [220] Carnegie Mellon University, ICT-MMMO Dataset, accessed: 2023-06-06 (2023). URL <http://multicomp.cs.cmu.edu/resources/ict-mmмо-dataset/>

- [221] V. Perez-Rosas, R. Mihalcea, L.-P. Morency, Multimodal opinion utterances dataset (moud), available through CMU Multimodal Data SDK on GitHub: <https://github.com/CMU-MultiComp-Lab/CMU-MultimodalSDK> (2013).
- [222] M. Aché, CMU-MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos, <https://www.kaggle.com/datasets/mathurinache/cmu-mosi>, kaggle Dataset (2021).
- [223] H.-C. Chou, W.-C. Lin, L.-C. Chang, C.-C. Li, H.-P. Ma, C.-C. Lee, NNIME: The NTHU-NTUA chinese interactive multimodal emotion corpus, in: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2017, pp. 292–298. doi:10.1109/ACII.2017.8273615.
URL <http://nnime.ee.nthu.edu.tw/>
- [224] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal sentiment intensity analysis in videos: Cmu-mosei, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 4796–4806.
- [225] A. Zadeh, E. M. Li, S. Poria, E. Cambria, L.-P. Morency, Affective meld: A multimodal dataset for emotion recognition in conversations, Online (2018).
URL <https://affective-meld.github.io/>
- [226] G. Shen, X. Wang, X. Duan, H. Li, W. Zhu, MEMoR: A Dataset for Multimodal Emotion Reasoning in Videos, IEEE, 2020-10-12, pages: 502. doi:10.1145/3394171.3413909.
URL <https://dl.acm.org/doi/10.1145/3394171.3413909>
- [227] A. Bagher Zadeh, Y. Cao, S. Hessner, P. P. Liang, S. Poria, L.-P. Morency, CMU-MOSEAS: A multimodal language dataset for spanish, portuguese, german, and french, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 1801–1812. doi:10.18653/v1/2020.emnlp-main.141.
URL <https://aclanthology.org/2020.emnlp-main.141>
- [228] L. Stappen, A. Baird, L. Schumann, S. Bjorn, The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements, IEEE Transactions on Affective Computing (2021) 1–1Conference Name: IEEE Transactions on Affective Computing. doi:10.1109/TAFFC.2021.3097002.
- [229] H. C. Ngo, A pipeline to data preprocessing for lipreading and audio-visual speech recognition, International Journal of Advanced Trends in Computer Science and Engineering 9 (2020) 4589–4596. doi:10.30534/ijatcse/2020/58942020.
- [230] J. L. Tracy, D. Randles, Four models of basic emotions: A review of ekman and cordaro, izard, levenson, and panksepp and watt, Emotion Review 3 (4) (2011) 397–405, publisher: SAGE Publications. doi:10.1177/1754073911410747.
URL <https://doi.org/10.1177/1754073911410747>
- [231] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, IEMOCAP: interactive emotional dyadic motion capture database, Language Resources and Evaluation 42 (4) (2008-12-01) 335–359. doi:10.1007/s10579-008-9076-6.
URL <https://doi.org/10.1007/s10579-008-9076-6>
- [232] V. Pérez-Rosas, R. Mihalcea, L.-P. Morency, Utterance-level multimodal sentiment analysis, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2013, pp. 973–982.
URL <https://aclanthology.org/P13-1096>

- [233] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, MELD: A multimodal multi-party dataset for emotion recognition in conversations (2019). [arXiv:1810.02508\[cs\]](https://arxiv.org/abs/1810.02508), doi:10.48550/arXiv.1810.02508.
URL <http://arxiv.org/abs/1810.02508>
- [234] O. Martin, I. Kotsia, B. Macq, I. Pitas, The eNTERFACE' 05 audio-visual emotion database, in: 22nd International Conference on Data Engineering Workshops (ICDEW'06), 2006, pp. 8–8. doi:10.1109/ICDEW.2006.145.
- [235] M. G. Huddar, S. S. Sannakki, V. S. Rajpurohit, Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification, International Journal of Multimedia Information Retrieval 9 (2) (2020-06-01) 103–112. doi:10.1007/s13735-019-00185-8.
URL <https://doi.org/10.1007/s13735-019-00185-8>
- [236] V. Pérez Rosas, R. Mihalcea, L.-P. Morency, Multimodal sentiment analysis of spanish online videos, IEEE Intelligent Systems 28 (3) (2013-05) 38–45, conference Name: IEEE Intelligent Systems. doi:10.1109/MIS.2013.9.
- [237] Y. Shao, S. Taylor, N. Marshall, C. Morioka, Q. Zeng-Treitler, Clinical text classification with word embedding features vs. bag-of-words features, in: 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 2874–2878. doi:10.1109/BigData.2018.8622345.
- [238] M. G. Huddar, S. S. Sannakki, V. S. Rajpurohit, A survey of computational approaches and challenges in multimodal sentiment analysis, International Journal of Computer Science and Engineering 7 (1) (2019-01) 876–883.
URL https://www.ijcseonline.org/pdf_paper_view.php?paper_id=3601&148-IJCSE-05569.pdf
- [239] B. Farnsworth, Facial action coding system (facs) – a visual guidebook, <https://imotions.com/blog/facial-action-coding-system/>, [Online]. Available: <https://imotions.com/blog/facial-action-coding-system/>. [Accessed 4 August 2021] (2019).
- [240] K. Sikka, L. Bramer, A. Divakaran, Deep Unified Multimodal Embeddings for Understanding both Content and Users in Social Media Networks, IJCV, 2019.
- [241] K. Raju, B. Chinna Rao, K. Saikumar, N. Lakshman Pratap, An optimal hybrid solution to local and global facial recognition through machine learning, in: P. Kumar, A. J. Obaid, K. Cengiz, A. Khanna, V. E. Balas (Eds.), A Fusion of Artificial Intelligence and Internet of Things for Emerging Cyber Systems, Intelligent Systems Reference Library, Springer International Publishing, 2022, pp. 203–226. doi:10.1007/978-3-030-76653-5_11.
URL https://doi.org/10.1007/978-3-030-76653-5_11
- [242] M. S. Hossain, G. Muhammad, An audio-visual emotion recognition system using deep learning fusion for a cognitive wireless framework, IEEE Wireless Communications 26 (3) (2019) 62–68. doi:10.1109/MWC.2019.1800419.
URL <https://ieeexplore.ieee.org/abstract/document/8752484>
- [243] n. Mustaqeem, S. Kwon, A CNN-assisted enhanced audio signal processing for speech emotion recognition, Sensors (Basel, Switzerland) 20 (1) (2019) 183. doi:10.3390/s20010183.
- [244] E. Ghaleb, M. Popa, S. Asteriadis, Metric learning-based multimodal audio-visual emotion recognition, IEEE MultiMedia (2019). doi:10.1109/MMUL.2019.2960219.
- [245] S. Jothimani, S. N. Sangeethaa, K. Premalatha, R. Sathishkannan, A new spatio-temporal neural architecture with bi-lstm for multimodal emotion recognition, in: ICCES, 2023, pp. 257–262. doi:10.1109/ICCES57224.2023.10192713.
URL <https://ieeexplore.ieee.org/abstract/document/10192713>

- [246] L. Schoneveld, A. Othmani, H. Abdelkawy, Leveraging Recent Advances in Deep Learning for Audio-Visual Emotion Recognition, *Pattern Recognition Letters* 146 (2021) 1–7, arXiv:2103.09154 [cs, eess]. doi:10.1016/j.patrec.2021.03.007.
URL <http://arxiv.org/abs/2103.09154>
- [247] M. Tang, M. Ling, J. Tang, J. Hu, A micro-expression recognition algorithm based on feature enhancement and attention mechanisms, *Virtual Reality* 27 (3) (2023) 2405–2416. doi:10.1007/s10055-023-00808-w.
URL <https://doi.org/10.1007/s10055-023-00808-w>
- [248] Q. Huang, J. Zhang, Z. Zeng, D. He, X. Ye, Y. Chen, PPDF-FedTMI: A federated learning-based transport mode inference model with privacy-preserving data fusion, *Simulation Modelling Practice and Theory* 129 (2023) 102845. doi:10.1016/j.simpat.2023.102845.
URL <https://www.sciencedirect.com/science/article/pii/S1569190X23001223>
- [249] N. A. Ab Razak, S. Sahran, Lightweight micro-expression recognition on composite database, *Applied Sciences* 13 (3) (2023) 1846. doi:10.3390/app13031846.
URL <https://www.mdpi.com/2076-3417/13/3/1846>
- [250] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, Cambridge, MA, 2016.
URL <http://www.deeplearningbook.org/>
- [251] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, The Graph Neural Network Model, *IEEE Transactions on Neural Networks* 20 (1) (2009) 61–80, conference Name: IEEE Transactions on Neural Networks. doi:10.1109/TNN.2008.2005605.
URL <https://ieeexplore.ieee.org/document/4700287>
- [252] M. C. Nwadiugwu, *Neural Networks, Artificial Intelligence and the Computational Brain*, arXiv, 2020.
URL <https://arxiv.org/abs/2101.08635>
- [253] P. S. Dhillon, P. P. Talukdar, K. Crammer, Learning better data representation using inference-driven metric learning, in: *Proceedings of the ACL 2010 Conference Short Papers*, Association for Computational Linguistics, 2010, pp. 377–381.
- [254] M. H. Rohban, H. R. Rabiee, Supervised neighborhood graph construction for semi-supervised classification, *Pattern Recognition* 45 (4) (2012) 1363–1372.
- [255] K. Ozaki, M. Shimbo, M. Komachi, Y. Matsumoto, Using the mutual k-nearest neighbor graphs for semi-supervised classification on natural language data, in: *Proceedings of the fifteenth conference on computational natural language learning*, 2011, pp. 154–162.
- [256] L. Berton, A. D. A. Lopes, Graph construction based on labeled instances for semisupervised learning, in: *2014 22nd International Conference on Pattern Recognition, IEEE*, 2014, pp. 2477–2482. doi:10.1109/ICPR.2014.428.
- [257] L. Berton, T. de Paulo Faleiros, A. Valejo, J. Valverde-Rebaza, A. de Andrade Lopes, Rgcli: Robust graph that considers labeled instances for semi-supervised learning, *Neurocomputing* 226 (2017) 238–248.
- [258] L. Zhuang, Z. Zhou, S. Gao, J. Yin, Z. Lin, Y. Ma, Label information guided graph construction for semi-supervised learning, *IEEE Transactions on Image Processing* 26 (9) (2017) 4182–4192. doi:10.1109/TIP.2017.2713941.
- [259] L. Waikhom, R. Patgiri, *Graph Neural Networks: Methods, Applications, and Opportunities*, arXiv, 2021.
URL <https://arxiv.org/abs/2108.10733>

- [260] F. Azhari, C. Sennersten, C. Lindley, E. Sellers, Deep learning implementations in mining applications: a compact critical review, *Artificial Intelligence Review* 56 (2023) 1–36. doi:10.1007/s10462-023-10500-9.
- [261] B. Khemani, S. Patil, K. Kotecha, S. Tanwar, A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions, *Journal of Big Data* 11 (1) (2024) 18. doi:10.1186/s40537-023-00876-4.
URL <https://doi.org/10.1186/s40537-023-00876-4>
- [262] M. Gori, G. Monfardini, F. Scarselli, A new model for learning in graph domains, in: *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, Vol. 2, 2005, pp. 729–734 vol. 2, ISSN: 2161-4407. doi:10.1109/IJCNN.2005.1555942.
URL <https://ieeexplore.ieee.org/document/1555942>
- [263] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph Neural Networks: A Review of Methods and Applications, *arXiv:1812.08434 [cs, stat]* (Oct. 2021). doi:10.48550/arXiv.1812.08434.
URL <http://arxiv.org/abs/1812.08434>
- [264] T. N. Kipf, M. Welling, Semi-Supervised Classification with Graph Convolutional Networks, *arXiv:1609.02907 [cs, stat]* (Feb. 2017). doi:10.48550/arXiv.1609.02907.
URL <http://arxiv.org/abs/1609.02907>
- [265] Y. LeCun, Y. Bengio, G. Hinton, Deep Learning, *Nature* 521 (2015) 436–44. doi:10.1038/nature14539.
- [266] F. R. K. Chung, *Spectral Graph Theory*, Vol. 92 of CBMS Regional Conference Series in Mathematics, American Mathematical Society, Providence, RI, 1997.
- [267] *A Guide to Convolutional Neural Networks for Computer Vision | Springer eBooks | IEEE Xplore* (2018).
URL <https://ieeexplore.ieee.org/book/8295029>
- [268] Yildirim, A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification, *Computers in Biology and Medicine* 96 (2018) 189–202. doi:10.1016/j.combiomed.2018.03.016.
URL <https://www.sciencedirect.com/science/article/pii/S0010482518300738>
- [269] P. Khorrami, T. Le Paine, K. Brady, C. Dagli, T. S. Huang, How deep neural networks can improve emotion recognition on video data, in: *2016 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2016, pp. 619–623.
- [270] L. Chao, J. Tao, M. Yang, Y. Li, Z. Wen, Long short term memory recurrent neural network based multimodal dimensional emotion recognition, in: *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC '15*, ACM, New York, NY, USA, 2015, pp. 65–72.
- [271] Y. Bengio, A. Courville, P. Vincent, Representation Learning: A Review and New Perspectives, *IEEE transactions on pattern analysis and machine intelligence* 35 (2013) 1798–1828. doi:10.1109/TPAMI.2013.50.
- [272] W. L. Hamilton, R. Ying, J. Leskovec, Inductive representation learning on large graphs, *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017).
URL <https://proceedings.neurips.cc/paper/2017/hash/5dd9db5e033da9c6fb5ba83c7a7e9-Abstract.html>
- [273] P. Cui, X. Wang, J. Pei, W. Zhu, A Survey on Network Embedding, *IEEE Transactions on Knowledge and Data Engineering PP* (Nov. 2017). doi:10.1109/TKDE.2018.2849727.

- [274] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, Proceedings of Workshop at ICLR 2013 (Jan. 2013).
- [275] W. L. Hamilton, R. Ying, J. Leskovec, Representation Learning on Graphs: Methods and Applications, arXiv:1709.05584 [cs] (Apr. 2018). doi:10.48550/arXiv.1709.05584.
URL <http://arxiv.org/abs/1709.05584>
- [276] P. V. Rouast, M. T. P. Adam, R. Chiong, Deep learning for human affect recognition: Insights and new developments, IEEE Transactions on Affective Computing 12 (2) (2021-04) 524–543, conference Name: IEEE Transactions on Affective Computing. doi:10.1109/TAFFC.2018.2890471.
- [277] M. Mansoorizadeh, N. Moghaddam Charkari, Multimodal information fusion application to human emotion recognition from face and speech, Multimedia Tools and Applications 49 (2) (2010-08-01) 277–297. doi:10.1007/s11042-009-0344-2.
URL <https://doi.org/10.1007/s11042-009-0344-2>
- [278] D. Hazarika, S. Gorantla, S. Poria, R. Zimmermann, Self-attentive feature-level fusion for multimodal emotion detection, in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2018, pp. 196–201. doi:10.1109/MIPR.2018.00043.
- [279] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, O. S. Neffati, Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network, Applied Sciences 13 (8) (2023) 4750, number: 8 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/app13084750.
URL <https://www.mdpi.com/2076-3417/13/8/4750>
- [280] Devopedia, Confusion matrix, <https://devopedia.org/confusion-matrix>, accessed: 2024-02-12 (2024).
- [281] T. Wang, Y. Hou, D. Zhou, Q. Zhang, A contextual attention network for multimodal emotion recognition in conversation, in: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021, pp. 1–7. doi:10.1109/IJCNN52387.2021.9533718.
- [282] Y.-J. Choi, Y.-W. Lee, B.-G. Kim, Residual-based graph convolutional network for emotion recognition in conversation for smart internet of things, Big Data 9 (4) (2021) 279–288. doi:10.1089/big.2020.0274.
URL <https://www.liebertpub.com/doi/10.1089/big.2020.0274>
- [283] J. Hu, Y. Liu, J. Zhao, Q. Jin, Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 5666–5675. doi:10.18653/v1/2021.acl-long.440.
URL <https://aclanthology.org/2021.acl-long.440>
- [284] Y. Shou, T. Meng, W. Ai, S. Yang, K. Li, Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis, Neurocomputing 501 (2022) 629–639. doi:10.1016/j.neucom.2022.06.072.
- [285] X. Tao, J. D. Velásquez, Multi-source information fusion for smart health with artificial intelligence, Information Fusion 83-84 (2022-07-01) 93–95. doi:10.1016/j.inffus.2022.03.010.
URL <https://www.sciencedirect.com/science/article/pii/S1566253522000343>
- [286] S. Tokuno, G. Tsumatori, S. Shono, E. Takei, T. Yamamoto, G. Suzuki, S. Mituyoshi, M. Shimura, Usage of emotion recognition in military health care, in: 2011 Defense Science Research Conference and Expo (DSR), IEEE, 2013, pp. 1–5. doi:10.1109/DSR.2011.6026823.
URL <http://ieeexplore.ieee.org/document/6026823/>

- [287] B. Das, A deep learning model for identification of diabetes type 2 based on nucleotide signals, *Neural Computing and Applications* 34 (2022) 12587–12599. doi:<https://doi.org/10.1007/s00521-022-07121-8>.
URL <https://link.springer.com/article/10.1007/s00521-022-07121-8>
- [288] S. Rajaram, M. Geetha, Vision-based human emotion recognition using HOG-KLT feature, in: *Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019)*, *Lecture Notes in Networks and Systems* 121, *Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019)*, 2020, pp. 261–272. doi:[10.1007/978-981-15-3369-3_20](https://doi.org/10.1007/978-981-15-3369-3_20).
- [289] M. Sajjad, M. Nasir, F. U. M. Ullah, K. Muhammad, A. K. Sangaiah, S. W. Baik, Raspberry pi assisted facial expression recognition framework for smart security in law-enforcement services, *Information Sciences* 479 (2019-04-01) 416–431. doi:[10.1016/j.ins.2018.07.027](https://doi.org/10.1016/j.ins.2018.07.027).
URL <https://www.sciencedirect.com/science/article/pii/S0020025518305425>
- [290] S. P. A. Dewi, Y. Fitriani, D. Effendi, Improving ability in identifying procedure text elements using visual audio media, in: *International Conference on Education Universitas PGRI Palembang (INCoEPP 2021)*, Atlantis Press, 2021, pp. 1011–1015, ISSN: 2352-5398. doi:[10.2991/assehr.k.210716.201](https://doi.org/10.2991/assehr.k.210716.201).
URL <https://www.atlantis-press.com/proceedings/incoepp-21/125958904>
- [291] H. S. Ali, F. ul Hassan, S. Latif, H. U. Manzoor, J. Qadir, Privacy enhanced speech emotion communication using deep learning aided edge computing, in: *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2021-06, pp. 1–5, ISSN: 2694-2941. doi:[10.1109/ICCWorkshops50388.2021.9473669](https://doi.org/10.1109/ICCWorkshops50388.2021.9473669).
URL <https://ieeexplore.ieee.org/document/9473669>
- [292] R. Cowie, Ethical issues in affective computing, *Frontiers in Robotics and AI* 7 (2020) 52. doi:[10.3389/frobt.2020.00052](https://doi.org/10.3389/frobt.2020.00052).
URL <https://www.frontiersin.org/articles/10.3389/frobt.2020.00052>
- [293] Z. Wang, Z. Hu, F. Li, S.-B. Ho, E. Cambria, Learning-based stock trending prediction by incorporating technical indicators and social media sentiment, *Cognitive Computation* 15 (3) (2023) 1092–1102. doi:[10.1007/s12559-023-10125-8](https://doi.org/10.1007/s12559-023-10125-8).
URL <https://doi.org/10.1007/s12559-023-10125-8>
- [294] Q. Ma, Y. Duan, Z. Yao, Meta-analysis of the impact of geospatial technologies on learning outcomes, *Education and Information Technologies* (2023). doi:[10.1007/s10639-023-11712-w](https://doi.org/10.1007/s10639-023-11712-w).
URL <https://doi.org/10.1007/s10639-023-11712-w>
- [295] D. H. Kim, M. K. Lee, D. Y. Choi, B. C. Song, Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild, in: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 529–535.
URL <https://dl.acm.org/doi/abs/10.1145/3136755.3143005>
- [296] A. Szymkowiak, P. Gaczek, K. Jeganathan, P. Kulawik, The impact of emotions on shopping behavior during an epidemic. what a business can do to protect customers, *Journal of Consumer Behaviour* 20 (1) (2021) 48–60. doi:[10.1002/cb.1853](https://doi.org/10.1002/cb.1853).
- [297] S. Pal, S. Mukhopadhyay, N. Suryadevara, Development and progress in sensors and technologies for human emotion recognition, *Sensors* 21 (16) (2021) 5554. doi:[10.3390/s21165554](https://doi.org/10.3390/s21165554).
- [298] R. Kosti, J. Alvarez, A. Recasens, A. Lapedriza, Context-based emotion recognition using emotic dataset, *IEEE Transactions on pattern analysis and machine intelligence* 42 (11) (2019) 2755–2766. doi:[10.1109/TPAMI.2019.2916866](https://doi.org/10.1109/TPAMI.2019.2916866).

- [299] A. Marmpena, Emotional body language synthesis for humanoid robots, Ph.D. thesis, University of Plymouth (2021). doi:10.24382/981.
- [300] I. Sarker, Ai-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems, SN Computer Science 3 (2) (2022) 158. doi:10.1007/s42979-022-01043-x.
- [301] A. Dziedzickis, A. Kaklauskas, V. Bucinskas, Human emotion recognition: Review of sensors and methods, Sensors 20 (3) (2020) 592. doi:10.3390/s20030592.
- [302] P. Baffour, H. Nunoo-Mensah, E. Keelson, B. Kommey, A survey on deep learning algorithms in facial emotion detection and recognition, Inform: Jurnal Ilmiah Bidang Teknologi Informasi Dan Komunikasi 7 (1) (2022) 24–32. doi:10.25139/inform.v7i1.4563.
- [303] P. Nandwani, R. Verma, A review on sentiment analysis and emotion detection from text, Social Network Analysis and Mining 11 (1) (2021) 81. doi:10.1007/s13278-021-00776-6.
- [304] S. Hamed, M. Ab Aziz, M. Yaakub, Fake news detection model on social media by leveraging sentiment analysis of news content and emotion analysis of users' comments, Sensors 23 (4) (2023) 1748. doi:10.3390/s23041748.
- [305] Y. Khurana, S. Gupta, R. Sathiyaraj, S. Raja, Robinnet: A multimodal speech emotion recognition system with speaker recognition for social interactions, IEEE Transactions on Computational Social Systems (2022). doi:10.1109/TCSS.2022.3228649.
- [306] M. S. Hossain, G. Muhammad, Emotion recognition using deep learning approach from audio-visual emotional big data, Information Fusion 49 (2019) 69–78. doi:10.1016/j.inffus.2018.09.008.
URL <https://www.sciencedirect.com/science/article/pii/S1566253517307066>
- [307] M. Karna, D. S. Juliet, R. Joy, Deep learning based Text Emotion Recognition for Chatbot applications, in: 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), 2020, pp. 988–993. doi:10.1109/ICOEI48184.2020.9142879.
URL <https://ieeexplore.ieee.org/document/9142879>
- [308] K. Chen, S. Gong, T. Xiang, C. C. Loy, Cumulative Attribute Space for Age and Crowd Density Estimation, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2467–2474, iSSN: 1063-6919. doi:10.1109/CVPR.2013.319.
URL <https://ieeexplore.ieee.org/document/6619163>
- [309] K. Chen, K. Jia, H. Huttunen, J. Matas, J.-K. Kämäräinen, Cumulative attribute space regression for head pose estimation and color constancy, Pattern Recognition 87 (2019) 29–37. doi:10.1016/j.patcog.2018.10.015.
URL <https://www.sciencedirect.com/science/article/pii/S0031320318303601>
- [310] W. Nie, Y. Yan, D. Song, K. Wang, Multi-modal feature fusion based on multi-layers LSTM for video emotion recognition, Multimedia Tools and Applications 80 (2021). doi:10.1007/s11042-020-08796-8.
- [311] H. Dolka, A. X. V. M, S. Juliet, Speech Emotion Recognition Using ANN on MFCC Features, in: 2021 3rd International Conference on Signal Processing and Communication (ICPSC), 2021, pp. 431–435. doi:10.1109/ICSPC51351.2021.9451810.
URL <https://ieeexplore.ieee.org/abstract/document/9451810>
- [312] M. Huddar, S. Sannakki, V. Rajpurohit, Attention-based multi-modal sentiment analysis and emotion detection in conversation using rnn, International Journal of Interactive Multimedia and Artificial Intelligence 7 (1) (2021) 44–51. doi:10.9781/ijimai.2020.07.004.
- [313] M. Schmitz, R. Ahmed, J. Cao, Bias and fairness on multimodal emotion detection algorithms, arXiv preprint arXiv:2205.08383 (2022). doi:10.48550/arXiv.2205.08383.

- [314] W. Mucha, M. Kampel, Depth and thermal images in face detection detailed comparison between image modalities, 2022 the 5th International Conference on Machine Vision and Applications (ICMVA) (2022) 16–21doi:10.1145/3523111.3523114.
- [315] M. Pagé Fortin, B. Chaib-draa, Multimodal multitask emotion recognition using images, texts, and tags, in: Proceedings of the ACM Workshop on Crossmodal Learning and Application, 2019, pp. 3–10. doi:10.1145/3326459.3329165.
- [316] A. Aslam, A. Sargano, Z. Habib, Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks, Applied Soft Computing (2023) 110494doi:10.1016/j.asoc.2023.110494.
- [317] Z. Liu, G. Huang, D. Chu, Y. Sun, Psrmer: Proactive services recommendation driven-by multimodal emotion recognition, in: 2023 IEEE International Conference on Web Services (ICWS), 2023, pp. 514–525. doi:10.1109/ICWS60048.2023.00070.
- [318] A. Mohammad, F. Siddiqui, M. Alam, S. Idrees, Tri-model classifiers for eeg based mental task classification: hybrid optimization assisted framework, BMC Bioinformatics 24 (1) (2023) 406. doi:10.1186/s12859-023-05544-1.
- [319] J. Tian, D. Hu, X. Shi, J. He, X. Li, Y. Gao, T. Toda, X. Xu, X. Hu, Semi-supervised multimodal emotion recognition with consensus decision-making and label correction, in: Proceedings of the 1st International Workshop on Multimodal and Responsible Affective Computing, 2023, pp. 67–73. doi:10.1145/3607865.3613182.
- [320] S. Chen, J. Tang, L. Zhu, W. Kong, A multi-stage dynamical fusion network for multimodal emotion recognition, Cognitive Neurodynamics 17 (3) (2023) 671–680. doi:10.1007/s11571-022-09851-w.
- [321] S. Patnaik, Speech emotion recognition by using complex MFCC and deep sequential model, Multimedia Tools and Applications 82 (2022). doi:10.1007/s11042-022-13725-y.
- [322] A. Joshi, A. Bhat, A. Jain, A. V. Singh, A. Modi, COGMEN: COntextualized GNN based multimodal emotion recognitioN (2022-05-05). arXiv:2205.02455[cs], doi:10.48550/arXiv.2205.02455.
URL <http://arxiv.org/abs/2205.02455>
- [323] Y. Cai, X. Li, J. Li, Emotion recognition using different sensors, emotion models, methods and datasets: A comprehensive review, Sensors 23 (5) (2023) 2455. doi:10.3390/s23052455.
- [324] P. Bhattacharya, R. Gupta, Y. Yang, Exploring the contextual factors affecting multimodal emotion recognition in videos, IEEE Transactions on Affective Computing (2021). doi:10.1109/TAFFC.2021.3071503.
- [325] K. Zhang, Y. Li, J. Wang, Z. Wang, X. Li, Feature fusion for multimodal emotion recognition based on deep canonical correlation analysis, IEEE Signal Processing Letters 28 (2021) 1898–1902. doi:10.1109/LSP.2021.3112314.
- [326] T. Shaikh, A. Khalane, R. Makwana, A. Ullah, Evaluating significant features in context-aware multimodal emotion recognition with xai methods, Authorea Preprints (2023). doi:10.22541/au.167407909.97031004/v1.
- [327] X. Zhang, M. Li, S. Lin, H. Xu, G. Xiao, Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild, IEEE Transactions on Circuits and Systems for Video Technology (2023). doi:10.1109/TCSVT.2023.3312858.
- [328] V. Nanduri, C. Sagiri, S. Manasa, R. Sanvithatesh, M. Ashwin, A review of multi-modal speech emotion recognition and various techniques used to solve emotion recognition on speech data, in: 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE, 2023, pp. 577–582. doi:10.1109/ICIRCA57980.2023.10220691.

- [329] X. Liu, Z. Xu, K. Huang, Multimodal emotion recognition based on cascaded multichannel and hierarchical fusion, *Computational Intelligence and Neuroscience* (2023). doi:10.1155/2023/9645611.
- [330] S. Sankala, S. M. Rafi B, S. R. M. K, Multi-feature integration for speaker embedding extraction, in: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7957–7961, iSSN: 2379-190X. doi:10.1109/ICASSP43922.2022.9746318.
URL <https://ieeexplore.ieee.org/document/9746318>
- [331] H. Filali, J. Riffi, C. Boulealam, M. A. Mahraz, H. Tairi, Multimodal Emotional Classification Based on Meaningful Learning, *Big Data and Cognitive Computing* 6 (3) (2022) 95, number: 3 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/bdcc6030095.
URL <https://www.mdpi.com/2504-2289/6/3/95>
- [332] A. Yousaf, M. Umer, S. Sadiq, D. S. Ullah, S. Mirjalili, V. Rupapara, M. Nappi, Emotion Recognition by Textual Tweets Classification Using Voting Classifier(LR-SGD), *IEEE Access* PP (2020) 1–1. doi:10.1109/ACCESS.2020.3047831.
- [333] K. D. N., A. Patil, Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks, in: *Interspeech 2020, ISCA, 2020*, pp. 4243–4247. doi:10.21437/Interspeech.2020-1190.
URL https://www.isca-speech.org/archive/interspeech_2020/n20_interspeech.html
- [334] Y. Cimtay, E. Ekmekcioglu, S. Caglar-Ozhan, Cross-subject multimodal emotion recognition based on hybrid fusion, *IEEE Access* 8 (2020) 168865–168878, conference Name: IEEE Access. doi:10.1109/ACCESS.2020.3023871.
URL <https://ieeexplore.ieee.org/document/9195813>
- [335] C. Guanghui, Z. Xiaoping, Multi-modal emotion recognition by fusing correlation features of speech-visual, *IEEE Signal Processing Letters* 28 (2021) 533–537, conference Name: IEEE Signal Processing Letters. doi:10.1109/LSP.2021.3055755.
URL <https://ieeexplore.ieee.org/document/9340264>
- [336] Y.-T. Lan, W. Liu, B.-L. Lu, Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism, in: *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020-07, pp. 1–6. doi:10.1109/IJCNN48605.2020.9207625.
URL <https://ieeexplore.ieee.org/document/9207625/>
- [337] H. Zhang, Expression-EEG based collaborative multimodal emotion recognition using deep AutoEncoder, *IEEE Access* 8 (2020) 164130–164143, conference Name: IEEE Access. doi:10.1109/ACCESS.2020.3021994.
URL <https://ieeexplore.ieee.org/document/9187342>
- [338] N. Ahmed, Z. A. Aghbari, S. Girija, A systematic survey on multimodal emotion recognition using learning algorithms, *Intelligent Systems with Applications* 17 (2023-02-01) 200171. doi:10.1016/j.iswa.2022.200171.
URL <https://www.sciencedirect.com/science/article/pii/S2667305322001089>
- [339] A. Solgi, A. Pourhaghi, R. Bahmani, H. Zarei, Improving SVR and ANFIS performance using wavelet transform and PCA algorithm for modeling and predicting biochemical oxygen demand (BOD), *Ecohydrology & Hydrobiology* 17 (2) (2017-04-01) 164–175. doi:10.1016/j.ecohyd.2017.02.002.
URL <https://www.sciencedirect.com/science/article/pii/S1642359316300672>
- [340] O. Vaidya, S. Gandhe, A. Sharma, A. Bhate, V. Bhosale, R. Mahale, Design and Development of Hand Gesture based Communication Device for Deaf and Mute People, in: *2020 IEEE Bombay Section Signature Conference (IBSSC)*, 2020, pp. 102–106. doi:10.1109/

IBSSC51096.2020.9332208.

URL <https://ieeexplore.ieee.org/document/9332208>

- [341] U. N. P. Release, International Day of Sign Languages, publisher: United Nations (2019).
URL <https://www.un.org/en/observances/sign-languages-day>
- [342] R. Alam, M. Munir, S. Ishrak, S. Hussain, M. Shalahuddin, M. N. Islam, A Machine Learning Based Sign Language Interpretation System for Communication with Deaf-mute People, Association for Computing Machinery, New York, NY, United States, 2021. doi: 10.1145/3471391.3471422.
- [343] R. San-Segundo, A. Pérez, D. Ortiz, L. D'Haro, M. Torres, F. Casacuberta, Evaluation of alternatives on speech to sign language translation, Vol. 1, International Speech Communication Association, 2007, journal Abbreviation: Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech Pages: 2532 Publication Title: Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. doi:10.21437/Interspeech.2007-672.
- [344] L. A. Goldstein, Speech-to-text apps for the deaf community, Hearing Like Me (June 2020).
URL <https://www.hearinglikeme.com/speech-to-text-apps-for-the-deaf-community/>
- [345] H. Daassi, J. Krahe, Universal Combined System: Speech Recognition, Emotion Recognition and Talking Head for Deaf and Hard of Hearing People, <https://www.researchgate.net/publication/237605884> (Jan. 2009). doi:10.3233/978-1-60750-042-1-503.
- [346] P. Millett, Accuracy of Speech-to-Text Captioning for Students Who are Deaf or Hard of Hearing, Journal of Educational, Pediatric & (Re)Habilitative Audiology 25 (2021) 1–13.
URL <https://www.edaud.org/journal/2021/1-article-21.pdf>
- [347] M. Shezi, A. Ade-Ibijola, Deaf Chat: A Speech-to-Text Communication Aid for Hearing Deficiency, Advances in Science, Technology and Engineering Systems Journal 5 (2020) 826–833. doi:10.25046/aj0505100.
- [348] F. Zhang, X.-C. Li, C. P. Lim, Q. Hua, C.-R. Dong, J.-H. Zhai, Deep emotional arousal network for multimodal sentiment analysis and emotion recognition, Information Fusion 88 (2022) 296–304. doi:10.1016/j.inffus.2022.07.006.
URL <https://www.sciencedirect.com/science/article/pii/S1566253522000653>
- [349] A. P. G, A. P. k, Design of an integrated learning approach to assist real-time deaf application using voice recognition system, Computers and Electrical Engineering 102 (2022) 108145. doi:10.1016/j.compeleceng.2022.108145.
URL <https://www.sciencedirect.com/science/article/pii/S0045790622003950>
- [350] K. Khan, D. N. Aslam, K. Abid, S. Munir, Robot Assist Sign Language Recognition for Hearing Impaired Persons Using Deep Learning, VAWKUM Transactions on Computer Sciences 11 (1) (2023) 245–267, number: 1. doi:10.21015/vtcs.v11i1.1491.
URL <https://www.vfast.org/journals/index.php/VTCS/article/view/1491>
- [351] J. J. S. AnithaElavarasi, Role of Machine Learning and Deep Learning in Assisting the Special Children's Learning Process, Journal of Algebraic Statistics 13 (2) (2022) 2327–2334, number: 2.
URL <https://publishoa.com/index.php/journal/article/view/430>
- [352] A. Shahid, S. Latif, J. Qadir, Generative Emotional AI for Speech Emotion Recognition: The Case for Synthetic Emotional Speech Augmentation, Applied Acoustics, 2023. doi:10.48550/arXiv.2301.03751.

- [353] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805 [cs] (May 2019). doi:10.48550/arXiv.1810.04805.
URL <http://arxiv.org/abs/1810.04805>
- [354] S. Liu, P.-T. Bremer, J. J. Thiagarajan, V. Srikumar, B. Wang, Y. Livnat, V. Pascucci, Visual Exploration of Semantic Relationships in Neural Word Embeddings, IEEE Transactions on Visualization and Computer Graphics PP (2017) 1–1. doi:10.1109/TVCG.2017.2745141.
- [355] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-P. Morency, S. Poria, Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis, in: Proceedings of the 2021 International Conference on Multimodal Interaction, 2021, pp. 6–15. arXiv:2107.13669.
URL <https://arxiv.org/abs/2107.13669>
- [356] S. Mai, Y. Zeng, H. Hu, Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations, IEEE Transactions on Multimedia 25 (2023) 4121–4134. doi:10.1109/tmm.2022.3171679.
URL <http://dx.doi.org/10.1109/TMM.2022.3171679>
- [357] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, E. Hoque, Integrating multimodal information in large pretrained transformers, in: Proceedings of the conference. Association for Computational Linguistics. Meeting, Vol. 2020, NIH Public Access, 2020, p. 2359. arXiv:1908.05787.
URL <https://arxiv.org/pdf/1908.05787.pdf>
- [358] S. Si, W. Ma, H. Gao, Y. Wu, T.-E. Lin, Y. Dai, H. Li, R. Yan, F. Huang, Y. Li, Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents (2024). arXiv:2305.13040.
URL <https://arxiv.org/pdf/2305.13040.pdf>

CURRICULUM VITAE

Hussein Farooq Tayeb ALSAADAWI

