A Systematic Review of Emotion Recognition in Children using Multimodal Data

Sangeetha S K B[1], Amiya Bhaumik[2], Raja Sarath Kumar Boddu[3]
[1]Postdoctoral Researcher,Lincoln University College, Malasiya; [2]Lincoln University College, Malasiya;
[3]Malla Reddy University, India
Email ID pdf.sangeetha@lincoln.edu.my

Abstract

Classifying emotion in children is an evolving research domain with various applications like education, healthcare, and psychological assessments. Unlike adults, children used to express highly dynamic emotional expressions, which makes emotion recognition even more complex. The proposed survey provides a review of recent developments in multimodal emotion classification which focus on the integration of facial expressions, speech signals, and physiological cues. The study explores different preprocessing techniques, feature extraction techniques, and classification models used for emotion recognition. The study also highlights key challenges in the variability of children's expressions, speech inconsistencies, physiological noise, and the necessity for effective multimodal fusion methods. In addition, research gaps in synchronization, standardization, and robustness of emotion classification models are also discussed. The review study provides valuable insights into possible research directions by analyzing the strengths and limitations of current methodologies to improve multimodal child emotion recognition systems.

Keywords: Multimodal Emotion Recognition, Child Emotion Detection, Facial Expression Analysis, Speech Signal Processing, Physiological Signal Preprocessing, Feature Fusion Techniques

1. Introduction

The complexity of accurate identification of emotional states in children has created a significant development in children's emotion detection over time. Traditionally, systems depended on basic algorithms to identify voice patterns and facial expressions based on visual and auditory cues. But, these systems needed to deal with issues such as variability in children's emotional expression and the challenges of discriminating among more subtle or culturally distinctive emotions. Children's facial expressions will not be as mature or conspicuous as those of adults, and their emotional responses are often more subtle, varying according to their stage of development and context. Due to this complexity, inferring emotions of children in traditional systems is less effective, which requires them to seek more advanced systems [1][2].

The most commonly employed modalities for emotion detection are visual, speech, and physiological cues. When facial expressions are subdued, hidden, or not typically considered to be indicators of emotion in adults, visual emotion detection is reduced in its capability of accurately discerning emotions. There has also been extensive use of speech-based systems that analyze prosodic features such as pitch,

tone, and rhythm. But, these systems have limitations such as cultural differences in the way individuals communicate, ambient noise, and the inability to properly extract emotions in children with possibly abnormal speech patterns [3][4]. While physiological signals like skin conductance, heart rate, Electroencephalogram(EEG) can yield informative data, they are often more difficult to obtain in real-time and it is difficult to correlate directly with observable emotional expressions [5].

With these disadvantages, a possible solution is provided through the integration of multiple modalities, such as visual, speech, and physiological signals. Multimodal emotion recognition systems can reduce the disadvantages of single-modality approaches by integrating numerous input sources. When facial expressions alone are ambiguous, voice or physiological signals can provide additional context to enhance accuracy. Additionally, multimodal systems are particularly beneficial for children with disorders such as Autism Spectrum Disorder (ASD), in which emotions are very hard to articulate or recognize through conventional means,because emotional signals of children could be subtle or developmental [6][7]. This multimodal recognition has proven a more promising approach in recognizing children's emotional complexity and understanding of their emotional lives [8][9].

In order to make children's emotion recognition systems more robust, adaptive, and effective under various circumstances and subject differences, improving emotion detection systems through the combination of these multiple modalities is much needed. In emotion recognition, increased accuracy and the potential for real-time use in the treatment, education, and health care of children, especially those with physiological conditions, are the outcomes of the combination between visual, auditory, and physiological inputs[10][11]. Developing systems to interact with children and to understand their emotional complexity and uniqueness, multimodal emotion recognition is more than a technological breakthrough.

Emotions are central to a child's cognitive, social, and psychological development, it is difficult to understand them. The emotional state of a child directly influences their behavior, learning, and overall well-being, it is difficult to detect it accurately. Children express emotions in subtle and complex ways that are hard for single modality methods such as words, facial expressions or physiological signals to capture. A multimodal approach provides a better understanding of a child's emotional state by combining multiple modalities such as body language, intonation of voice, facial expressions, and physiologic cues. The deep understanding is imperative for the identification of mental health challenges or emotional issues at an early stage so that timely and effective interventions can be initiated. In addition, fusion of various emotional signals in recognition systems is also important for enhancing accuracy so that environments supportive of emotions and responsive to the needs of children can be built.

The main contributions are

1. To provide review of recent developments in multimodal emotion classification which focus on the integration of facial expressions, speech signals, and physiological cues.

2. The study explores different preprocessing techniques ,feature extraction techniques, and fusion methods used for emotion recognition.

3. The study also highlights key challenges in synchronization, standardization, and robustness of emotion classification models and also provides valuable insights into possible research directions to improve multimodal child emotion recognition systems.

With this detailed introduction, Section 2 discusses background study, Section 3 studies the multiple data modalities for emotion classification, Section 4 outlines the preprocessing and feature extraction methods multimodalities, Section 5 explores different feature fusion strategies, Section 6 discusses possible research challenges and future research directions followed by conclusion in Section 7.

2. Related work

Due to social masking, multimodalities based emotion recognition is better than single modalities like speech or face analysis. The research reviews Machine Learning(ML) classifiers including Support Vector Machine(SVM) and K-Nearest Neighbor(KNN), feature extraction, and reduction. The study reviews emotions based connections among brain areas and EEG rhythms. The study also provides a comparison between deep learning and machine learning models. Research directions and future challenges in emotion recognitions are also discussed [1]. The research employed facial and vocal emotion tasks to analyze 57 non-ASD and 99 ASD adolescents. The groups had similar patterns of errors, and recognition ability was also impacted by Intelligent Quotient(IQ). Higher IQ adolescents are performed better on all emotional tests and ASD doesn't seem to have a core deficit in basic emotion perception[2].

Single-modalities like voice or face expression analysis have their limitations,so multimodal emotion recognition is increasingly gaining traction. For human-computer interaction, this study discusses data-driven multimodal emotion information fusion. Emotion data sets, feature extraction from speech, face expression, text, EEG, and fusion algorithms are also discussed. Real-time mental health monitoring applications are also discussed along with future research directions for multimodal emotion recognition [3].The study attempts to monitor the emotions of preschool children by examining facial expressions [4]. The study employs a deep learning approach to combine face, gesture, and context information with a lightweight network architecture. The study shows the geometric features to enhance contour detection and facial appearance. Emotion recognition using audio is also explored, and using hierarchical sampling method, class imbalance is resolved. In both controlled and outdoor environments ,the experimental results show the effectiveness of the proposed method.

[5] presents an ensemble-based, multimodal emotion recognition system to solve the issue of missing data in real-world applications. The study employs varieties of ensemble methods for handling missing modalities in the fusion level instead of ignoring incomplete input. Both traditional and advanced emotion-specific fusion methods are also explored. The CALLAS Expressivity Corpus contains facial, vocal, and gestural modalities, and is used to evaluate the method. The advantages and drawbacks of different fusion processes are also illustrated by comparison. [6] presents SVM and Multi Layer Perceptron(MLP) classifiers for testing a library of emotional speech in Russian for ages 8 to 12. The reliability of the database is also demonstrated by the automatic detection of four emotions: neutral, joy, sadness, and anger which surpasses perceptual testing. The results demonstrate SVM and MLP as benchmark baselines

for future deep learning models. The database used serves as a valuable resource for studying affective speech in child-computer interactions.

The study addresses concerns like face occlusion and blurry backgrounds for enhancing automatic emotion recognition in children with ASD in social interactions [7]. The proposed method uses body postures and face expressions for emotion recognition. Convolutional Neural Networks(CNNs) are used to extract facial features and temporal transformers. Graph convolutional networks and self-attention are employed to evaluate body postures. Multimodal fusion methods are also explored. The method has potential for therapeutic use in providing emotional feedback and is better than traditional methods using only face information when applied to a real-world child dataset. [8] presents the PRISMA approach to provide a systematic literature review and methods for detecting emotions in autistic children are also explored. The study uses physiological signals, speech, and facial expressions to examine basic emotions like fear, sadness, and happiness. Single modal and multimodal approaches are also discussed. The study presents early fusion being the most common for multimodal recognition and SVM being the widely used classifiers. The study also identifies the challenges including disturbances in emotion signals, labeling methods, and open datasets creation.

[9] studies problems like face occlusion and fuzzy backgrounds for enhancing automatic emotion detection in children with ASD during social interactions. The proposed method combines body postures with facial expressions for high precision. Facial features are extracted using CNNs and temporal transformers. The body postures are evaluated using graph convolutional networks and self-attention. Multimodal fusion methods are proposed to enhance recognition. The method presented for therapeutic use in giving emotional feedback and performing better than traditional methods using only face information when tested on a real-world child-clinician dataset. [10] uses the PRISMA approach to perform a systematic review for detecting emotions in children with autism. The study examines physiological signals, speech, and facial expressions to examine basic emotions such as fear, sadness, and happiness. Children with ASD exhibited high but diverse emotion identification scores and struggled to recognize emotions, especially neutral ones based on visual cues. Multimodal stimuli reduced some of these difficulties. The children's emotion recognition was correlated with developmental age, but only in the multimodal task in children with ASD. In children with ASD, language impairments were related to poorer auditory modality recognition. The findings highlight that when conducting research on emotion processing in ASD, developmental variables and comorbidities should be considered.

[11] focuses on explaining children's facial expressions to identify emotions to improve online learning interactions. The study obtained accuracy rates of 89.31% and 90.98%, respectively, using LIRIS and a new dataset of emotions from kids ages 7 to 10. The study used 3D landmark points to develop LIRIS-Mesh and Authors-Mesh versions to account for variations in children's face expressions. They improved the accuracy and interpretability of emotion recognition by using seven CNN models and explainable AI approaches such as Grad-CAM. The study examines emotion recognition, which includes identifying and analyzing emotional states including fear and happiness [12]. Since social media posts during the COVID-19 pandemic revealed the mental health of the people. The study compares popular emotion recognition datasets, discusses machine and deep learning classifiers for feature extraction, and provides an overview

of emotion acquisition tools with high recognition accuracy. To enhance emotion classification, it also examines various data fusion methods, emphasizing both their advantages and limitations.

[13] compares the capacity of children to recognize voice emotional expressions from speech prosody with visual emotional expressions from static faces. 313 kindergarten children (mean age = 51.01 months) took part. Children in the experiment viewed facial expressions or listened to spoken sentences in visual and aural blocks. The results showed that it was easier to recognize emotions from static faces compared to inferring them from prosody in speech alone. As individuals matured, the discrimination of "happy" emotion from facial expression and "sad" emotion from speech prosody increased. In the discrimination of "sad" facial expressions, girls significantly outperformed guys. The findings bear implications for teaching nonverbal communication. [14] introduces m_AutNet, a customized multimodal neural architecture to help kids with ASD recognize emotions better. The system combines a CNN to interpret speech expressions and a face feature extraction module to combine vocal and facial expressions. A generative adversarial network uses domain adaptation, for enhancing feature alignment for more accurate emotion classification. With an accuracy of 88.25%, the suggested approach performs better than traditional classifiers and has promise to improve emotion detection in kids with ASD.

[15] provides a detailed review of Multimodal Emotion Recognition (MER) systems developed between 2014 and 2024 for modalities like voice, body movements, facial expressions, physiological signals, linguistics, and recent ones like drawings. Along with the overview of the past of automated emotion recognition systems, the study classifies emotions, sentiments, moods, and feelings. The study analyzes 45 articles with the PRISMA guidelines and reports seven standards for examining MER methods. The study points out the human-centered perspective of the field while giving insight into existing MER technology, datasets, challenges, and directions for the future. Most existing datasets cannot record emotions in natural social interactions which is a challenge the K-EmoCon dataset tries to overcome. Multimodal measurements of video recordings, EEG, and physiological signals were collected using off-the-shelf equipment throughout 16 sessions of dual debates on social issues. The dataset is unique to provide emotion annotations from three perspectives that are external witnesses, argument counterparts, and self. Emotions are rated at every five seconds based on 18 category emotions and arousal-valence. The first publicly available dataset to study emotions in real social interactions is K-EmoCon [16].

[17] discusses the challenges and methods of automating emotional recognition, for applications such as identifying children's emotions before medical tests. The study also discusses the International Affective Picture System (IAPS) and modern biometric platforms such as iMotions can detect emotional responses in adults and children. From a mixed sample of children and adults, the study provides statistical analysis of the ratings to IAPS pictures, highlighting significant areas for future research and providing insights into how emotional reactions might differ by age. [18] discusses the role of Emotional Intelligence (EI) and Emotion Recognition (ER) in understanding and managing human emotions. The study discusses how physiological signals and non-invasive methods such as audio and video analysis can be employed to achieve emotion recognition. MER is popular with the help of deep learning advancements and the presence of emotion-laden video content. It also discusses unimodal and multimodal emotion recognition

methods, including feature fusion and joint representations. The study also lists their limitations and suggests directions for future work. It also introduces a multi-modal physiological emotion database that collects breathing, Electrocardiogram(ECG), galvanic skin response, and EEG data to study human emotions [19]. To minimize cultural effects, the database is built with a collection of 28 standardized emotion-provoking video clips which are assessed with psychological methods. The study utilizes classifiers like SVM and k-NN for emotion recognition, and multiple classification protocols and feature extraction methods. A novel attention- Long Short Term Memory(LSTM) model is proposed to enhance discriminative feature extraction. The database is made freely available for future emotion estimate research.

By integrating physiological, verbal, and visual signals, multimodal emotion recognition overcomes the limitations of single-modality methods and enhances the accuracy of emotion detection. While deep learning methods enhance fine-grained recognition, they have limitations such as noise, missing modalities, and insufficient training data. Besides pointing out the need for better data representation, richer feature extraction, and efficient multimodal fusion methods, this review highlights the importance of dynamic expression analysis to recognize modest micro-expressions. Robust frameworks addressing real-world challenges are required for effective computing developments for ensuring more accurate and contextually aware emotion detection systems [20]. For Human Robot Interactions (HRI), humanoid social robots need to be accepted such that their emotional expressions have to be unambiguous and convincing. The study explores the effects of contextual and cross-modal incongruence on humans' judgments of emotions expressed by robots in speech and gesture. Inconsistent emotional cues are puzzling for the observers, and this also decreases the recognition accuracy. The study presents recommendations on how to improve robotic emotion communication in HRI situations and highlights the importance of better multimodal emotional expression [21]. [22] evaluated two training courses for accuracy in nonverbal emotion recognition. One for facial micro-expressions and the other for multimodal expressions. Following three weeks of weekly training, the outcome showed that neither program had any transfer effects. Micro-expression training improved micro-expression recognition, while multimodal training was better.

Table 1. Proposed System Vs Existing Systems

Aspect	Existing Systems	Advantages of Proposed System
Emotion Detection Modalities (Ref. [3], [4], [15])	Typically use a single modality (e.g., facial expressions, speech prosody)	Integrates multiple modalities (e.g., facial expressions, speech, body language, physiological signals) for more accurate detection
Emotion Recognition Accuracy (Ref. [5], [10], [19])	Limited in recognizing complex or subtle emotions due to reliance on one modality	Improved accuracy by combining multiple emotional cues, capturing more complex emotional states

Adaptability (Ref. [7], [10], [12])	May struggle with varied emotional expressions, especially in children	Better suited for diverse emotional expressions by considering various signals and child-specific behavior
Real-world Application (Ref. [8], [13], [17])	May not perform well in uncontrolled environments or with real-time interaction	Designed for real-world scenarios, such as classrooms or clinical settings, handling noisy or incomplete data more effectively
Response to Complex Emotions (Ref. [9], [11], [18])	Struggles with recognizing complex emotions or mixed emotional states	Capable of detecting complex, mixed emotions and subtle emotional changes through multimodal data fusion
Development and Data Fusion (Ref. [5], [6], [16])	Often lacks sophisticated data fusion techniques, limiting its accuracy	Uses advanced fusion strategies for combining data from various sources, improving robustness and reliability
Context Awareness (Ref. [10], [17], [21])	Limited understanding of context and individual differences	More context-aware, considering child-specific factors (e.g., developmental stage, social context) for tailored emotion recognition
Target Group (Ref. [2], [14], [15])	Primarily focused on adults or general populations	Specifically designed for children, especially those with developmental challenges like ASD, improving accessibility and inclusivity
Emotion Categorization (Ref. [9], [12], [13])	Typically focused on basic emotions (e.g., happy, sad, angry)	Capable of recognizing a broader spectrum of emotions, including complex emotional states in children
Healthcare/Clinical Applications (Ref. [7], [8], [9])	Limited application in clinical or therapeutic settings	Enhanced potential for clinical interventions by providing real-time emotional feedback for therapists and caregivers

From Table 1, it is observed that while there are currently systems available to focus on detecting emotions in children, most of them utilize only one modality, such as speech or facial expressions. These approaches have limitations, particularly when it comes to detecting subtle or complex emotions. Children exhibit emotions harder to interpret through facial expressions, especially when they are not able to fully express themselves or when there are ambiguous emotional signs. Multimodal emotion detection systems make a much better and more reliable understanding of the emotional state of a child through the integration of numerous indications like body language, voice tone, facial expressions, and physiological indicators. These systems are in the stages of development, and they have problems such as data fusion, modalities missing, and adaptation to real-world environments. For real-time use in areas such as education, healthcare, or mental health, current systems may not be as advanced or readily available. Increased advances in multimodal emotion recognition are needed, especially in the creation of child-oriented models that are able to identify and respond to emotions in a range of environments with accuracy for ensuring personalized therapy to meet the needs of each child.

3. Multimodal Emotion Recognition

Children's spontaneous, unstable, and context-dependent reactions make their emotions difficult to detect. When compared to single-modality methods, a multimodal method enhances the consistency of emotion classification by combining speech, face and physiological information. The primary advantage of multimodal emotion recognition is that it solves the limitations of single modality by taking advantage of complimentary information from other modalities.

3.1 Significance of Multimodal Approaches in Child Emotion Classification

While children's feelings are often expressed in a multitude of sometimes ambiguous ways. Traditional emotion classification methods are based on a single source of information such as voice cues or facial expressions. Due to environmental contexts, social pressures, or developmental factors, single modalities of a child are not always congruent with what they say. Through the combination of multiple modalities, a multimodal framework enhances classification accuracy by detecting subtle emotional differences that an individual modality may miss. It enhances robustness by overcoming incomplete or erroneous data from single modalities and provides contextual awareness, allowing it to separate between similar emotions, such as sadness and frustration, that might have different physiological patterns but similar facial and speech characteristics. In real-world applications, like adaptive learning systems, psychiatric assessment, and assistive technology for children with special needs, this multimodal classification is a useful solution.

3.2 Commonly Used Modalities in Multimodal Emotion Classification

The key modalities used in child emotion recognition include facial expressions (visual modality), speech signals (audio modality), and physiological signals (biological modality). Each modality provides unique advantages and specific challenges in child emotion classification, as summarized in Table 2.

Table 2: Overview of Multimodal Emotion Recognition Modalities [23][24]

Modality	Signal Type	Advantages	Challenges
Facial Expressions	Visual (RGB/IR)	Directly observable, rich in emotion-related features	Affected by occlusions, cultural variations, and rapid changes in expressions
Speech Signals	Audio (Waveform)	Captures emotional intonations and variations	High variability in children's speech patterns, background noise sensitivity
Physiological Signals	Biological EDA, HRV, EEG	Objective measure of internal emotional states	Requires sensors, prone to motion artifacts and environmental influences

3.3 Facial Expressions – Importance in Emotion Recognition

A person's emotional state can be assumed from facial expressions, which are one of the most important aspects of human emotion. Especially in the development of verbal communication in early childhood, children primarily use facial expressions. Facial expression analysis necessitates several indispensable methods. Face detection within an image or video by using models such as Multi-Task Cascaded Convolutional Neural Network(MTCNN), You Only Look Once(YOLO), or Haar cascades is the first step, which is referred to as face detection. Facial landmark detection uses software such as Dlib or OpenFace to capture important facial landmarks, such as the eyes' corners and the edges of the mouth, which can also enable exact expression analysis. The extraction of features is a critical step to obtain high-level facial features. So deep learning models are often employed, which involve CNN-based structures like ResNet-50, VGG-Face, and EfficientNet. The OpenFace framework enables the detection of Facial Action Units (FAUs), which are micro-expressions linked to specific emotions. But detection of children's facial expressions is very challenging. Classification efforts are made to detect subtle changes in expression that children often exhibit. Intricacy is enhanced by expression variability, which is also affected by individual, cultural, and environmental factors. Precise detection is also affected by practical problems such as occlusions, varying illumination, head posture changes, and hand movements. Therefore, robust and adaptable analysis methods are essential in this field. Table 3 depicts the common facial expression methods.

Table 3: Common Facial Expression Recognition Techniques [25][26]

Technique	Description	Common Models

Face Detection	Locates faces in images	MTCNN, Haar cascades, YOLO
Landmark Detection	Identifies facial key points	Dlib, OpenFace
Feature Extraction	Extracts deep features from faces	ResNet-50, VGG-Face
Facial Action Unit Analysis	Identifies fine-grained muscle movements	OpenFace, Affectiva API

3.4 Speech Signals – Variability in Children's Speech Patterns

Speech also causes intonation, pitch, and rhythm variations, so it is a better source of emotional information. But the children's speech characteristics are less consistent because of variable phoneme articulation, so it is difficult to detect emotions from their speech than in adults. Children also often use shorter sentences with incomplete phrases, which also restricts the amount of data available for analysis. The tone and pitch emotional indicators of children are not more clear than those of adults because of their high-pitched voices are also challenging. In voice data preprocessing and feature extraction, background noise is removed using methods such as Wiener filtering and spectral subtraction, and non-speech segments are removed using WebRTC-based Voice Activity Detection (VAD). Feature extraction techniques include Mel-Frequency Cepstral Coefficients (MFCCs), which are widely used for speech emotion recognition, spectrogram analysis are used to visualize time-frequency variations in speech, and prosodic features that capture pitch, intensity, and duration metrics. Table 4 shows the key features for speech.

Table 4: Key Features for Speech-Based Emotion Recognition [27][28]

Feature Type	Description	Importance
MFCCs	Captures spectral properties of speech	Most effective for emotion classification
Spectrograms	Time-frequency representation of speech	Useful for deep learning models
Prosodic Features	Pitch, duration, and intensity variations	Enhances emotional expressiveness

3.5 Physiological Signals – Use of EDA, HRV, and EEG in Emotion Detection

A more objective measure of emotional arousal, physiological indicators are particularly valuable when facial and vocal expressions are ambiguous or suppressed. The physiological modalities are EEG which records patterns of brain activity associated with different emotions, Heart Rate Variability (HRV) tracks autonomic nervous system activity to provide an index of emotional control, and ElectroDermal Activity (EDA) quantifies skin conductance response to report levels of stress and arousal. The intrusiveness of sensors of EEG and EDA can be uncomfortable for children which impacts data quality. Movement-based noise can also corrupt signals and cause time synchronization problems. The physiological responses require alignment methods such as DTW, to solve issues in physiological emotion recognition. However, psychological signals are a more reliable basis for understanding children's emotions in multimodal emotion recognition. Table 5 depicts the psychological signals.

Table 5: Physiological Signals for Emotion Recognition [29][30]

Modality	Measured Signal	Emotional Relevance
EDA	Skin conductance response	Stress and arousal detection
HRV	Heart rate fluctuations	Anxiety and relaxation states
EEG	Brain activity	Cognitive and emotional processing

4. Preprocessing and Feature Extraction Techniques for Multimodal Data

To ensure that the extracted features are explicit, consistent, and coherent across multiple modalities, preprocessing is a crucial phase in multimodal emotion classification. Each modality physiological signals, voice signals, and facial expressions requires specific preprocessing methods for enhancing feature quality, reducing noise, and enabling better fusion.

4.1 Facial Expression Processing

Preprocessing is needed in order to handle occlusions, lighting variations, and face misalignment because facial expressions provide better visual cues for emotion recognition in children. The processing of facial images or videos involves face detection which is used to make sure only the face features are involved in emotion analysis by cropping the Region Of Interest (ROI). Feature extraction detects discriminative patterns for emotion recognition following face detection. Before feature extraction, image enhancement techniques are employed for enhancing the quality of facial images. While gamma correction rectifies brightness variation caused by inhomogeneous illumination, histogram equalization enhances contrast in face images. Denoising filters such as the Gaussian and median filters also help to remove noise from poor quality images, to provide more accurate and reliable feature extraction. Table 6 shows the face detection algorithms and Table 7 depicts feature extraction methods.

Table 6: Face Detection Algorithms [31]

Face Detection Algorithm	Description	Strengths	Limitations
MTCNN (Multi-task Cascaded CNN)	Deep learning-based face detection and alignment	High accuracy, detects multiple faces	Computationally expensive
Haar Cascades	Traditional method using hand-crafted features	Fast and lightweight	Sensitive to lighting variations
YOLO (You Only Look Once)	Real-time object detection including faces	Fast, works on video streams	May miss small faces
Dlib (HOG+SVM-based model)	Uses Histogram of Oriented Gradients (HOG) features	Robust to variations in pose	Not ideal for real-time applications

Table 7: Feature Extraction Methods [32]

Feature Extraction Method	Description	Common Models/Tools
Deep Learning-based	Extracts high-level features automatically	ResNet, VGG, EfficientNet
Facial Landmark-based	Identifies key facial points (eyes, mouth, eyebrows)	Dlib, OpenFace
Facial Action Units (AUs)	Recognizes muscle movements linked to emotions	OpenFace, Affectiva API

4.2 Speech Signal Processing

Pitch variations, rhythm changes, and intonation are all captured in speech signals and are essential for classifying emotions in children. Raw speech data includes noise, silent pauses, and unnecessary audio segments. VAD eliminates silent or unnecessary portions from speech recordings because they contain pauses, background noise, and non-verbal sounds. Table 8 shows the noise reduction methods , Table 9 depicts the feature extraction for speech and Table 10 displays voice activity detection algorithms.

Table 8: Noise Reduction Methods

Noise Reduction Method	Description	Common Tools
Wiener Filtering	Adaptive filtering method for reducing background noise	MATLAB, Librosa
Spectral Subtraction	Estimates noise spectrum and subtracts it from the signal	Audacity, Praat
Pre-emphasis Filtering	Boosts high frequencies to enhance speech clarity	Librosa, Kaldi

Table 9: Feature Extraction Methods [33][34]

Feature Type	Description	Common Extraction Methods
MFCC	Captures spectral features	Librosa, OpenSMILE
Spectrograms	Visual representation of frequency over time	STFT, Wavelet Transform
Prosodic Features	Measures pitch, energy, and rhythm variations	OpenSMILE, Kaldi

Table 10: Voice Activity Detection Algorithms

VAD Algorithm	Description	Common Use Cases
WebRTC VAD	Real-time VAD for noisy environments	Real-time applications
GMM-based VAD	Uses Gaussian Mixture Models to classify speech and silence	Emotion recognition datasets

Energy-based VAD	Determines speech based on energy thresholding	Simple offline processing

4.3 Physiological Signal Processing

The quantification of arousal and cognitive states for emotions can be derived from physiological signals such as EDA, HRV, and EEG. The signals should be properly preprocessed because they are often affected by motion artifacts, sensor noise, and environmental interference. The individual emotional variances can significantly affect the study and data normalization methods play a major role in physiological signal processing to ensure comparability between samples. The normalization methods include robust scaling, which rescales data based on the median and InterQuartile Range (IQR); Min-Max scaling, where data is scaled to 0 and 1; and Z-score standardization, where data is converted into a normal distribution with a mean of 0 and a standard deviation of 1. Table 11 shows the filtering methods and Table 12 shows the modalities of physiological signals.

Table 11: Filtering Methods

Filtering Method	Description	Application
Bandpass Filtering	Removes unwanted frequencies outside the signal range	EEG, EDA, HRV
Moving Average Smoothing	Reduces sudden fluctuations in physiological signals	Heart Rate Variability (HRV)
Wavelet Denoising	Decomposes signals into frequency bands to remove noise	EEG preprocessing

Table 12: Modalities of Physiological Signals [35][36]

Modality	Feature Type	Description
EDA (Electrodermal Activity)	Skin Conductance Level (SCL), Skin Conductance Response (SCR)	Measures emotional arousal
HRV (Heart Rate Variability)	Time-domain and frequency-domain features	Reflects stress, relaxation

EEG (Electroencephalography)	Power Spectral Density (PSD), Band Power (Alpha, Beta, Theta)	Analyzes cognitive and emotional states

In multimodal systems, speech, facial expressions, and physiological signals all operate at different sampling rates, so time synchronization is critical. Physiological signals like EEG are recorded at different rates (250-500 Hz), speech signals are recorded at different audio sample rates (16-44 kHz), and facial expression features are extracted at different video frame rates (30-60 fps). The validity of emotion classification in children can be reduced if the data streams are not synchronized. Synchronization methods involve linear interpolation, which resamples signals into a shared time base (e.g., alignment of EEG and HRV), cross-correlation, which calculates similarity between signals at various time shifts (e.g., alignment of facial and speech data), and dynamic time warping (DTW), which synchronizes time-series data with different speeds (e.g., speech and EEG synchronizing). To provide high-quality multimodal data for emotion identification in children, proper preprocessing pipelining of psychological signals is essential. Therefore, proper preprocessing methods are needed for each modality for aligning signals, eliminating noise, and extracting meaningful characteristics.

5. Feature Fusion Techniques

Feature fusion plays a n important role in multimodal child emotion classification because it integrates various data sources such as facial expressions, speech signals, and physiological signals for enhancing accuracy and robustness. The choice of fusion strategy significantly impacts the ability of models to capture emotional differences across multiple modalities.

5.1 Fusion Types

Fusion methods are classified into early fusion (feature-level integration) and late fusion (decision-level integration) as shown in Table 13 and its comparison in Table 14.

Table 13: Early Fusion vs. Late Fusion [37]

Fusion Type	Description	Advantages	Challenges
Early Fusion	Combines raw or preprocessed features from all modalities before classification.	Captures cross-modal interactions, improves feature richness.	Requires feature alignment, sensitive to missing data.
Late Fusion	Processes each modality independently and merges decisions at the classification stage.	More robust to missing data, modular approach.	Loses intermodal dependencies, may require ensemble learning.

Table 14. Comparison of Early and Late Fusion in Emotion Classification

Aspect	Early Fusion	Late Fusion
Information Preservation	Retains raw feature correlations	Limited interaction between modalities
Computational Complexity	High (requires large feature vectors)	Lower (simpler models per modality)
Flexibility	Less flexible (fixed feature extraction)	More adaptable to changing modalities
Handling Missing Data	Challenging (missing values disrupt training)	Better (each modality operates independently)

5.2 Attention Mechanisms for Fusion

Attention mechanisms dynamically emphasize important features when ignoring unimportant information to enhance emotion recognition performance and interpretability. Considering model accuracy enhancement, attention-based fusion provides insights into the multiple modalities that contribute to the final classification. Cross-attention effectively balances speech and face expressions by picking up on dependencies among a variety of modalities. Hierarchical attention is particularly beneficial for complex multimodal fusion models because it focuses on multiple levels of features, from low-level to high-level.

5.3 Challenges in Feature Fusion for Multimodal Data

Even though feature fusion in multimodal emotion recognition has advantages, it also has several drawbacks. Different modalities run at different temporal resolutions, e.g., TV at 30 frames per second compared to EEG at 500 Hz, which becomes an issue in aligning data. Cross-modal converters and Dynamic Time Warping (DTW) are two available methods to address this challenge. The multiple high-dimensional features can lead to redundant information, feature redundancy and high dimensionality. This issue can be resolved by methods like Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA), and feature selection. The main challenge is to deal with noisy or missing modalities because not all sensors accurately capture data, which can lead to problems like microphone noise or facial occlusions. These problems can be reduced by dropout-based multimodal training and imputation methods. In addition, computational complexity is a significant consideration, especially for deep learning-based fusion models which require a high amount of memory and processing. Efficient methods for managing these requirements are quantization, model pruning, and light-weight fusion systems. Because

feature fusion integrates multiple signals for overall emotion recognition, it is necessary for accurate multimodal based kid emotion classification. Whereas late fusion provides robustness against modalities lost, early fusion maintains more abundant feature interactions at the cost of careful alignment. But feature integration is enhanced by advanced techniques such as CCA, deep learning-based fusion, and attention mechanisms. The synchronization, redundancy, and computational requirements issues need to be addressed for optimal performance.

5.4 Comparative Study of Different Multimodal Feature Fusion Architectures for Child Emotion Classification

One of the key aspects of multimodal child emotion classification is feature fusion, which defines how well data from multiple modalities like speech, face, and physiology is integrated. While early fusion does not have feature interaction across modalities and has a medium computing cost, it is not highly robust to missing data. Even though late fusion does not depend on intermodal interconnections, it is computationally efficient and robust to missing data. Attention-based hybrid fusion balances out through the integration of medium computational requirements and cross-modal dependency. Transformer-based fusion provides better accuracy because of dynamic learning of relevance to individual modality, very high tolerance for missing data, and effective cross-modal interactions. But, it requires larger training data and an extremely high cost of computing. Late fusion is powerful but restricted in handling interdependencies and early fusion is robust for synchronized modalities but faces high-dimensional issues. Transformer-based fusion is optimal in accuracy and flexibility, but at great computational cost. Hybrid fusion provides an optimal balance out of all these architectures. Table 15 shows the comparison study of different fusion architectures.

Table 15. Comparison Study of Different Fusion Architectures

Model Type	Model	Strengths	Weaknesses	Computational Cost
Traditional ML	SVM	Good for small datasets, handles high-dimensional data	Lacks temporal modeling, limited scalability	Medium
	Random Forest (RF)	Robust to noisy data, prevents overfitting	Computationally expensive, not real-time	Medium

	Decision Trees (DT)	Easy to interpret, good for categorical data	Prone to overfitting, struggles with high-dimensional features	Low
Deep Learning	CNN	Strong spatial feature extraction	Limited sequential data handling	High
	LSTM	Effective for sequential data (speech, EEG)	High computational cost, needs large data	High
	Transformer-based Models	Captures long-range dependencies	Requires large datasets, high complexity	Very High
Hybrid/Ensemble	CNN + LSTM	Balances spatial and sequential learning	Resource-intensive	Very High
	Ensemble (RF + SVM)	Good generalization across datasets	High complexity, limited scalability	Medium
	Attention-based Fusion Model	Learns inter-modal relationships effectively	Needs large training datasets	Very High

6. Challenges and Research Gaps

Even though there has been significant improvement, still there are some significant hurdles in multimodal child emotion recognition. The main difficulty is the lack of large multimodal datasets for children. The diversity and size of existing datasets, like RAVDESS, IEMOCAP, and DEAP, are limited, and they are focusing on adults. Ethics and privacy regulations (e.g., COPPA and GDPR-K) also limit dataset collection. The emotional responses of children vary significantly by age so that generalization is difficult. The speech, facial expressions, and physiological cues always operate on different time scales and require precise synchronization. The cross-modal synchronization and feature combining problems are also considered as a great challenge. Early fusion and late fusion strategies of feature fusion which fail to find an equilibrium between eliminating redundancy and combining information. Along with, noise and absent

data among modalities can also reduce model consistency. The existing deep learning models often function as "black boxes" whose performance it is not easy for instructors and therapists to have confidence in. Interpretable AI models within psychological and therapy use are notably essential.

The interpretability of methods such as Local Interpretable Model-agnostic Explanations(LIME) and SHapley Additive exPlanations(SHAP), and attention mechanisms is also increased, but they are not fully optimized for multimodal child emotion analysis, and their lack of transparency makes them unsuitable for use in real-world applications such as adaptive learning and mental health monitoring. Future studies should concentrate on these crucial areas in order to overcome these issues. An automated, age-adaptive standardized preprocessing framework for children's multimodal data such as face, speech, and physiological signals are required. Through the utilization of unlabeled child emotion data, transformer topologies (e.g., BERT, ViTs, and multimodal transformers) can improve feature extraction and fusion by detecting long-range dependencies. The real-time child emotion recognition systems integrating real-time multimodal input should involve working on edge-friendly, lightweight AI models that are deployable on wearables, mobile phones, and Internet of Things platforms. Future research can significantly enhance emotional AI use in psychology, healthcare, and education by addressing these challenges through improved availability of datasets, feature fusion methods, interpretability, and model efficiency.

7. Conclusion

Multimodal emotion recognition for children is growing emerging as an important application-domain with significant impact in psychology, education, and medicine. The proposed study, along with preprocessing techniques , feature fusion approaches and comparison of classification frameworks, discusses the importance of multimodality based emotion classification, including speech, face expression, and physiological signal processing. Traditional machine learning algorithms such as SVM and Random Forest, deep learning methods such as CNNs and LSTMs, and Transformer-based models are demonstrating improved performance in emotion recognition. Apart from these advances, several challenges still remain, including the need for interpretable AI models, the lack of large multimodal datasets for children, and cross-modal synchronization problems. There should be standardized preprocessing frameworks, better self-supervised learning, and real-time adaptive recognition systems in order to avoid these limitations. Future research should focus on developing AI models that are scalable, privacy-preserving, and reliable so that they can be morally integrated into apps targeting children. Multimodal emotion detection can play a major role in early mental health assessment of children, personalized assistance, and child-oriented AI systems with the evolving data collection, fusion methods, and AI explainability.

References

1. Zhang, Jianhua, Zhong Yin, Peng Chen, and Stefano Nichele. "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review." *Information Fusion* 59 (2020): 103-126. https://doi.org/10.1016/j.inffus.2020.01.011
2. Jones, Catherine RG, Andrew Pickles, Milena Falcaro, Anita JS Marsden, Francesca Happé, Sophie K. Scott, Disa Sauter et al. "A multimodal approach to emotion recognition ability in autism

spectrum disorders." *Journal of Child Psychology and Psychiatry* 52, no. 3 (2011): 275-285. https://doi.org/10.1111/j.1469-7610.2010.02328.x

3. Jiang, Yingying, Wei Li, M. Shamim Hossain, Min Chen, Abdulhameed Alelaiwi, and Muneer Al-Hammadi. "A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition." *Information Fusion* 53 (2020): 209-221. https://doi.org/10.1016/j.inffus.2019.06.019
4. Yu, Guiping. "Emotion monitoring for preschool children based on face recognition and emotion recognition algorithms." *Complexity* 2021, no. 1 (2021): 6654455. https://doi.org/10.1155/2021/6654455
5. Wagner, Johannes, Elisabeth Andre, Florian Lingenfelser, and Jonghwa Kim. "Exploring fusion methods for multimodal emotion recognition with missing data." *IEEE Transactions on Affective Computing* 2, no. 4 (2011): 206-218. https://doi.org/10.1109/T-AFFC.2011.12
6. Matveev, Yuri, Anton Matveev, Olga Frolova, Elena Lyakso, and Nersisson Ruban. "Automatic speech emotion recognition of younger school age children." *Mathematics* 10, no. 14 (2022): 2373. https://doi.org/10.3390/math10142373
7. Liu, Jingjing, Zhiyong Wang, Wei Nie, Jia Zeng, Bingrui Zhou, Jingxin Deng, Huiping Li, Qiong Xu, Xiu Xu, and Honghai Liu. "Multimodal Emotion Recognition for Children with Autism Spectrum Disorder in Social Interaction." *International Journal of Human–Computer Interaction* 40, no. 8 (2024): 1921-1930. https://doi.org/10.1080/10447318.2023.2232194
8. Landowska, Agnieszka, Aleksandra Karpus, Teresa Zawadzka, Ben Robins, Duygun Erol Barkana, Hatice Kose, Tatjana Zorcec, and Nicholas Cummins. "Automatic emotion recognition in children with autism: a systematic literature review." *Sensors* 22, no. 4 (2022): 1649. https://doi.org/10.3390/s22041649
9. Cimtay, Yucel, Erhan Ekmekcioglu, and Seyma Caglar-Ozhan. "Cross-subject multimodal emotion recognition based on hybrid fusion." *IEEE Access* 8 (2020): 168865-168878. https://doi.org/10.1109/ACCESS.2020.3023871
10. Xavier, Jean, Violaine Vignaud, Rosa Ruggiero, Nicolas Bodeau, David Cohen, and Laurence Chaby. "A multidimensional approach to the study of emotion recognition in autism spectrum disorders." *Frontiers in psychology* 6 (2015): 1954. https://doi.org/10.3389/fpsyg.2015.01954
11. Rathod, Manish, Chirag Dalvi, Kulveen Kaur, Shruti Patil, Shilpa Gite, Pooja Kamat, Ketan Kotecha, Ajith Abraham, and Lubna Abdelkareim Gabralla. "Kids' emotion recognition using various deep-learning models with explainable ai." *Sensors* 22, no. 20 (2022): 8066. https://doi.org/10.3390/s22208066
12. Ahmed, Naveed, Zaher Al Aghbari, and Shini Girija. "A systematic survey on multimodal emotion recognition using learning algorithms." *Intelligent Systems with Applications* 17 (2023): 200171. https://doi.org/10.1016/j.iswa.2022.200171
13. Covic, Amra, Nicole von Steinbüchel, and Christiane Kiese-Himmel. "Emotion recognition in kindergarten children." *Folia Phoniatrica et Logopaedica* 72, no. 4 (2020): 273-281. https://doi.org/10.1159/000500589
14. Kurian, Asha, and Shikha Tripathi. "m_AutNet–A Framework for Personalized Multimodal Emotion Recognition in Autistic Children." *IEEE Access* (2024). https://doi.org/10.1109/ACCESS.2024.3403087

15. Kalateh, Sepideh, Luis A. Estrada-Jimenez, Sanaz Nikghadam Hojjati, and Jose Barata. "A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges." *IEEE Access* (2024). https://doi.org/10.1109/ACCESS.2024.3430850
16. Park, Cheul Young, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. "K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations." *Scientific Data* 7, no. 1 (2020): 293. https://doi.org/10.1038/s41597-020-00630-y
17. Flynn, Maria, Dimitris Effraimidis, Anastassia Angelopoulou, Epaminondas Kapetanios, David Williams, Jude Hemanth, and Tony Towell. "Assessing the effectiveness of automated emotion recognition in adults and children for clinical investigation." *Frontiers in human neuroscience* 14 (2020): 70. https://doi.org/10.3389/fnhum.2020.00070
18. Gladys, A. Aruna, and V. Vetriselvi. "Survey on multimodal approaches to emotion recognition." *Neurocomputing* (2023): 126693. https://doi.org/10.1016/j.neucom.2023.126693
19. Song, Tengfei, Wenming Zheng, Cheng Lu, Yuan Zong, Xilei Zhang, and Zhen Cui. "MPED: A multi-modal physiological emotion database for discrete emotion recognition." *IEEE Access* 7 (2019): 12177-12191. https://doi.org/10.1109/ACCESS.2019.2891579
20. Udahemuka, Gustave, Karim Djouani, and Anish M. Kurien. "Multimodal Emotion Recognition using visual, vocal and Physiological Signals: a review." *Applied Sciences* 14, no. 17 (2024): 8071. https://doi.org/10.3390/app14178071
21. Tsiourti, Christiana, Astrid Weiss, Katarzyna Wac, and Markus Vincze. "Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots." *International Journal of Social Robotics* 11 (2019): 555-573. https://doi.org/10.1007/s12369-019-00524-z
22. Döllinger, Lillian, Petri Laukka, Lennart Björn Högman, Tanja Bänziger, Irena Makower, Håkan Fischer, and Stephan Hau. "Training emotion recognition accuracy: results for multimodal expressions and facial micro expressions." *Frontiers in Psychology* 12 (2021): 708867. https://doi.org/10.3389/fpsyg.2021.708867
23. Zhang, S., Yang, Y., Chen, C., Zhang, X., Leng, Q., & Zhao, X. “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): 121692. https://doi.org/10.1016/j.eswa.2023.121692
24. Ramaswamy, M. P. A., & Palaniswamy, S. “Multimodal emotion recognition: A comprehensive review, trends, and challenges.” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *14*(6),(2024): e1563. https://doi.org/10.1002/widm.1563
25. Canal, F. Z., Müller, T. R., Matias, J. C., Scotton, G. G., de Sa Junior, A. R., Pozzebon, E., & Sobieranski, A. C. “ A survey on facial emotion recognition techniques: A state-of-the-art literature review.” *Information Sciences*,(2022): *582*, 593-617. https://doi.org/10.1016/j.ins.2021.10.005
26. Li, S., & Deng, W. “Deep facial expression recognition: A survey.” *IEEE transactions on affective computing*, *13*(3), (2020): 1195-1215.https://doi.org/10.1109/TAFFC.2020.2981446
27. Koolagudi, S. G., & Rao, K. S. “Emotion recognition from speech: a review.” *International journal of speech technology*, *15*, (2012): 99-117. https://doi.org/10.1007/s10772-011-9125-1

28. Swain, M., Routray, A., & Kabisatpathy, P. "Databases, features and classifiers for speech emotion recognition: a review." *International Journal of Speech Technology*,(2018): *21*, 93-120. https://doi.org/10.1007/s10772-018-9491-z
29. Egger, M., Ley, M., & Hanke, S. "Emotion recognition from physiological signal analysis: A review." *Electronic Notes in Theoretical Computer Science*, (2019): *343*, 35-55. https://doi.org/10.1016/j.entcs.2019.04.009
30. Domínguez-Jiménez, J. A., Campo-Landines, K. C., Martínez-Santos, J. C., Delahoz, E. J., & Contreras-Ortiz, S. H. "A machine learning model for emotion recognition from physiological signals." *Biomedical signal processing and control*, *55*, (2020): 101646. https://doi.org/10.1016/j.bspc.2019.101646
31. Kumar, A., Kaur, A., & Kumar, M. "Face detection techniques: a review." *Artificial Intelligence Review*,(2019): *52*, 927-948. https://doi.org/10.1007/s10462-018-9650-2
32. Wang, H., Hu, J., & Deng, W. "Face feature extraction: a complete review." *IEEE Access*, *6*, (2017): 6001-6039. https://doi.org/10.1109/ACCESS.2017.2784842
33. Aggarwal, A., Srivastava, A., Agarwal, A., Chahal, N., Singh, D., Alnuaim, A. A., ... & Lee, H. N. "Two-way feature extraction for speech emotion recognition using deep learning." *Sensors*, *22*(6),(2022): 2378. https://doi.org/10.3390/s22062378
34. Jahangir, R., Teh, Y. W., Hanif, F., & Mujtaba, G. "Deep learning approaches for speech emotion recognition: state of the art and research challenges." *Multimedia Tools and Applications*, *80*(16), (2021): 23745-23812. https://doi.org/10.1007/s11042-020-09874-7
35. Yan, M., Deng, Z., He, B., Zou, C., Wu, J., & Zhu, Z. "Emotion classification with multichannel physiological signals using hybrid features and adaptive decision fusion." *Biomedical Signal Processing and Control*, *71*, (2022): 103235. https://doi.org/10.1016/j.bspc.2021.103235
36. Ahmad, Z., & Khan, N. "A survey on physiological signal-based emotion recognition." *Bioengineering*, *9*(11), (2020): 688. https://doi.org/10.3390/bioengineering9110688
37. Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions." *Information Fusion*,(2023): *91*, 424-444. https://doi.org/10.1016/j.inffus.2022.09.025