Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 10, October - 2025

A Robust Hybrid Approach to Multi-Modal Emotion Recognition

Supreeth A K Gowda
Department of Computer and Information Sciences,
Northumbria University,
Newcastle upon Tyne, United Kingdom

Abstract: **This paper presents a robust, hybrid approach for multi-modal emotion recognition based on facial, speech and physiological signals. This study is being undertaken to compare multimodal system against unimodal systems and to overcome limitations of the latter whilst improve robustness of multimodal systems against real-world conditions. Deep learning techniques are used which include a shared convolutional neural network for facial and speech modalities working jointly with an adaptive fusion algorithm, which assigns dynamic weights to each modality. Results show improvements in accuracy and resilience compared to unimodal approaches under very challenging real-world conditions. The research addresses some major challenges in affective computing, including the issues of temporal synchronicity, context dependency, and ethical considerations. Domains of application include but are not limited to health, education, and analysis of consumer behaviours. The paper concludes with a call for ethical guidelines and industrial standards on how to implement emotion-reading technology responsibly.**

Keywords **– emotion recognition; multi-modal fusion; deep learning; convolutional neural networks; affective computing; human-computer interaction; remote photoplethysmography; adaptive fusion; ethics in AI**

1. NTRODUCTION

1.1 Background and Context

Emotions are a primitive and basal instinct present in animals used to overcome and adapt to problems such as hunting, hierarchy, status, sex, social life including others [1]. These instincts could have evolved though natural selection or be a social construct which varies across distinct cultures and ethnicities [2]. Emotions such as fear have a significant role in the survival and proliferation of animals especially humans [3] as it improves their sense of finding predators and avoiding them [4]. Humans experience the world through emotions [5] which play a crucial role in determining the quality of life, overall well-being [6] and the mental processes of a person [7], [8] which have been historically helping humanity to learn and adapt to the environment thus ensuring its survival [9]. This deep connection between emotions and the human mind shows the importance of understanding and controlling emotions in improving personal and social functionality.

Emotion recognition is pivotal in many fields such as human-computer interaction, healthcare, psychology, virtual reality and security [10], [11]. It is especially very crucial to implement this technology in the healthcare and psychology sector due an increasing number of people succumbing to mental pressure including children and teenagers [12]. Emotion recognition is used by healthcare personnel to better understand and identify illnesses and provide better treatment options [13] to improve patient care and reduce strain on medical care providers. With the rise of technology and fields including Artificial Intelligence and Machine Learning, the demand for machines which can perceive and communicate emotions has been increasing [14] which has led to significant efforts towards understanding emotions to allow effective human-computer interaction [15]. The market for emotion recognition systems is previously valued at USD 19.87 million [16], currently at USD 57.25 billion in 2024 and is anticipated to rise significantly to nearly USD 139 million by 2029 with a compound annual growth rate (CAGR) of 19.49% [17]. This growth would also provide secondary benefits such as job creation in technology sector, and create new revenue streams through creation of innovative applications and services. A major market in which emotion recognition is gaining popularity is consumer behaviour where the effectiveness of advertisements is assessed through the attention span [18] and create a Neuro Marketing strategy to influence emotions shown by the consumer and manipulate them to buying products [19]. Another would be in education where tutors and educations institutions would use emotion recognition to understand student behaviour and adapt learning techniques to improve student engagement and learning outcomes [20].

The development of robust emotion recognition, however, is challenged by technical issues, including:

- Temporal alignment of different modalities (facial expressions, speech, physiological signals) that operate on different timescales [6].
- Detecting adapting to complex real-world scenarios [91] while maintaining consisted performance [92].

Additionally, these opportunities also raise ethical considerations such as privacy and consent [21]. To bypass these issues, strong frameworks and industry standards play a critical role in the implementation and deployment of emotion recognition systems. While there is high economic impact, ethical principles must be followed, and oversight protocols must be in place to leverage the full potential of this system. Multi-modal emotional recognition (MER) uses multiple signals, including text, speech features, and facial expressions, to enhance the accuracy and robustness of emotion recognition systems. This has been important in the human-computer interaction field to help understand and respond to human emotions to provide more intuitive and intelligent interactions [22].

Research and development towards emotion recognition has seen considerable progress, from psychological research to the creation of advanced computational techniques and algorithms. With rapid advancements in deep learning techniques and data being made more accessible, Multimodal Emotion Recognition has ground breaking new research [22]. For example, datasets such as the Interactive Emotional Dyadic Motion Capture (IEMOCAP), Toronto Emotional Speech Set and the Multimodal Opinion Utterances Dataset (MOUD) are the some of the best public resources available for the training and evaluating models used for emotion recognition.

One significant advantage of a multi-modal approach is that it can facilitate the support of complementary information from other modalities such as facial expressions, speech patterns, and body language which significantly boosts the ability of emotion recognition systems to provide inferences about emotional states accurately [23].

Multi-modal emotion recognition is a landmark in the understanding of human emotions. When these signals are combined and mapped using advanced computational methods, Multi-modal emotion recognition systems become highly accurate and robust. However, there are a few challenges in the MER field which are significant such as understanding subtlety of emotional expressions [24]. Other challenges that are yet to be completely removed from the process include designing an ideal neural architecture and providing context on how emotions are created and lighted using extracted features and data incompleteness [25]. Despite all these challenges, major progress may still be on the way for Multimodal emotion recognition with recent developments in deep learning and comprehensive datasets.

1.2 Problem statement

The study addresses the following key challenges in multi-modal emotion recognition:

1. Effective integration of facial, auditory, and physiological signals for improved accuracy and robustness.
2. Development of an adaptive fusion algorithm to dynamically weight different modalities based on their reliability and context.
3. Addressing the issue of temporal synchronization between different modalities.
4. Improving the system's robustness to missing or noisy data in real-world conditions.

The focus of this study has been primarily on advancing the field of multi-modal emotion recognition and developing a system with increased robustness, and application in real-world scenarios.

1.3 Research Objectives

The study is being conducted to answer the following challenges:

Q1. Investigating the effectiveness and advantages of combining multiple modalities such as facial, auditory signals and physiological data.

Q2. The development of emotion recognition system using machine learning techniques such as deep learning and evaluation against unimodal architecture.

Q3. What are the best ways the system can be applied to improve human-machine interaction in actual applications of healthcare, education, and consumer behaviour?

Q4. What ethical concerns should be taken into account around the use of emotion recognition technology, and how may strong frameworks and industry standards be created to make sure of the responsible implementation of the system?

1.5 Significance of the study

This research study utilises a multi-modal approach using facial, auditory, and physiological modalities to interpret and understand emotions highlighted by an individual and create a non-invasive tool to improve human-machine interaction. This approach can facilitate and complement other systems used to monitor an individual and significantly improve treatment quality.

Scope and Limitations

This study focuses on developing and evaluating a multi-modal emotion recognition system that integrates facial, auditory, and physiological modalities. The study will make use of available datasets like the Toronto emotional speech set (TESS) and the Face Expression Recognition Plus dataset (FER+) datasets to train and test the proposed models. However, the study does have some limitations:

- The present study is limited to making use of available datasets, which are not comprehensive in their coverage of all ways of expressing emotions or cultural variations in emotional display.
- Training of the proposed tool will be done using data taken in a controlled setting, so its actual performance varies in real-life applications due to factors like environmental noise, variations in lighting conditions, person-to-person differences in emotional expressions and so on.
- This study will try to address all ethical implications of emotion recognition technology and strong ethical frameworks and industry standards will be considered when the system is to be implemented and will be deployed responsibly.

2. LITERATURE REVIEW

This literature review aims to compile an exhaustive overview of previous and current works in multi modal emotion detection. By exploring theoretical foundations, key methodologies, challenges and findings, this review seeks to display the potential that multimodal approaches present in deepening our understanding of the detection of human emotions.

2.1 Emotion theory and models

The basis of psychological theories on emotion lays the ground for multimodal emotion detectors. The most influential model has been the circumplex model of effect which has been extensively used for emotion recognition using computers [26].

Another important theoretical contribution is that by Ekman and Friesen in 1971, who suggested six universal emotions—happiness, sadness, fear, anger, disgust, and surprise—based on facial expressions considered to be recognized pan-culturally [27]. Recent research advocating a greater complexity of states of emotion has disputed this kind of discrete categorization as it still underlies many emotion detection systems.

2.1.1 Unimodal emotion detection

Unimodal system only uses single modalities such as facial, audio, text or any other means to recognize the emotion shown by the user [28]. This section will underline the three main unimodal approaches to emotion recognition: Facial, Speech and Text-based.

Facial emotion detection

Facial expression-based approaches are one of the most studied methods. It involves the detection of facial cues and interpreting them can show the person's emotional state. The FACS system which has been used since 1978 is one of the earlier frameworks specifying facial movements and their associated emotions [27].

Traditional approaches to facial expression analysis are focused on geometry features or distances between facial landmarks, and appearance features, which include texture and intensity [29]. A more recent trend in the past years is to use deep learning techniques such as Convolutional Neural Networks, for automatic learning and feature extraction from facial images [30].

These challenges arise from the fact that analysis of facial expression must deal with varying factors such as lighting conditions, head position, occlusion, and individual differences in emotion expression [31]. Several techniques have been developed and implemented to address these problems, including data augmentation, transfer learning, and the use of attention mechanisms [30].

Speech Emotion Recognition

Speech emotion recognition is the identification and recognition of paralinguistic characteristics related to the acoustic characteristics of speech, from which an emotional state can be inferred. Emotions are here assumed to be encoded within paralinguistic features pertaining to pitch, energy, and spectral characteristics in speech [32].

Previous works required the manual extraction of features for such tasks: these included prosodic features such as pitch, energy, and duration, along with spectral features such as Mel-frequency cepstral coefficients [33]. More recently, deep learning techniques, particularly RNNs and LSTMs, have been used to understand temporal dependencies and extract the relevant features from the speech signals [34].

In this regard, some of the major challenges of recognizing emotional speech include linguistic content, speaker variability, and background noise [33]. Many techniques have been proposed for enhancing speech emotion recognition in these situations, including feature selection, data augmentation, and domain adaptation [35].

Text-Based Emotion Detection

Text-based emotion detection is simply a process directed toward the revelation of emotions through written or spoken words. It mainly works with semantic and syntactic features of text to understand the underlying emotional message in the document [36].

Traditional methods for detecting emotion in text have relied on the rule-based and machine learning techniques, such as Naive Bayes, SVM, Decision Trees, coupled with handcrafted bag-of-words and lexical affinity features [37]. Recently, deep learning-based approaches use CNNs and RNNs for the effective extraction of relevant features from the data in a text [38].

Problems with text-based emotion detection stem from the ambiguity and subjectivity of language, sarcasm and irony [38], and the limited availability of labelled data. To acknowledge these classical challenging problems, a number of authors have proposed a variety of techniques to handle them, from use of transfer learning, data augmentation, and incorporation of external knowledge sources into the models.

2.1.2 Multimodal emotion detection

One reason for using multimodal approaches is realizing that affective information is transferred through multiple channels simultaneously. Multiple modalities incorporate complementary information and outperform any limitations of unimodal techniques and provide more accurate Emotion Recognition [39].

Previous works in emotion recognition focused on acquiring and improving results from unimodal systems using single sources of information such as facial data [40], [41] or audio signals [34].

Facial expression has been extensively analysed by techniques such as deep learning models for feature extraction and classification [42]. Similarly, emotion recognition though speech has been analysed with the use of recurrent neural networks (RNNs) to capture rhythmic features and intonations [43]. However, although these methods have shown promising results individually, a mixed approach is shown to be a better option as it tends to be more accurate [44].

Modalities and Features

Facial Expressions

Facial expressions are one of the most explored modalities in emotion recognition. Advances in computer vision and deep learning have significantly enhanced facial emotion recognition. CNN techniques have been used to attain meaningful results in the extraction of features from facial images [30]. Building on this tradition, attention mechanisms in CNNs to extract fine-grained facial micro-expressions [45].

Speech and Audio

Voice and speech patterns are essential carriers of emotional cues. Traditionally, acoustic features—particularly those on pitch, energy, and spectral aspects—have been used in speech emotion recognition [46]. Recently, with the improvement of deep learning techniques, investigations towards the use of deep models to automatically learn relevant features from raw audio data [47].

Physiological Signals

Measures of physiological signals including heart rate variability, and conductance of skin-based methods of EEG offer objective measures of emotional states. These signals could be of excellent value since they cannot be voluntarily controlled. Based on a study, different varieties of physiological signals might be applied to recognize emotions [48]. A further study has indicated that several kinds of physiological signals combined with deep learning methods raise the accuracy of the classification of emotions [49].

Body Posture and Gestures

Although less studies have been conducted on body language compared to facial expressions and speech, it still provides a relevant source of emotional information. A study conducted on body postures gives an excellent review of the perception and recognition of affective body expressions [50]. The recent success of pose estimation techniques has further served for more accurate analysis of body movements in emotion detection using a graph convolutional network approach for recognizing emotions from skeletal data [51], achieving good performances on several datasets.

Textual and semantic content

In cases where either verbal or written communication exists, semantic content will still be helpful for providing emotional context. Traditionally, NLP (Natural Language Processing) techniques have been used in detecting emotions from text. A study provided an overview of the survey on emotion detection in the text, covering traditional machine learning and deep-learning-based methods [52].

2.1.3 Multimodal Fusion Techniques

Early Fusion

Early fusion, also known as feature-level fusion, is based on concatenating features stemming from different modalities before classification. This approach allows one to learn joint representations; however, it faces issues with the curse of dimensionality. Early fusion is highly effective in multimodal sentiment analysis by simply fusing features extracted from text, images, and audio [53].

Late Fusion

In late fusion, independent predictions are obtained for each modality, after which these predictions are combined. In this approach, there is more flexibility, and modality-specific classifiers can be used. One study proposed an adaptive late-fusion method in which the contribution of each modality would be dynamically adapted based on their reliability, which turned out to perform better than static techniques for fusion [54].

Hybrid fusion

Hybrid fusion approaches are designed to make use of the strengths of both early and late fusion. [89] suggested a hybrid fusion approach for the bimodal recognition of emotions with facial expressions and body gestures, showing better results than unimodal and single-fusion approaches.

Attention Mechanisms

Recent research explored the use of attention mechanisms to focus on relevant features or modalities adaptively. Multipronged attention fusion can be used in multimodal emotion recognition, thus effectively capturing cross-modal relations and attaining state-of-the-art performance on several benchmark datasets [55].

2.2 Deep Learning Approaches

Convolutional Neural Networks (CNNs)

Feature extraction has been carried out using CNN in both modalities: visual and audio. [56] presented a CNN-based methodology that performed better compared to traditional techniques in the field of audio-visual emotion recognition.

Recurrent Neural Networks (RNNs)

RNNs have captured temporal dynamics in both speech signals and physiological signals with remarkable success, more specifically using the Long Short-Term Memory network to build a continuous emotion multimodal method, performing better than static models [57].

Transformer-based Models

The massive success of the Transformer models in Natural Language Processing has caused the adoption of the technology into multimodal emotion detection. [58] proposed a Transformer-based neural architecture to fuse facial, vocal, and textual cues through which reliable results were obtained for several multi-modal emotion datasets.

Graph Neural Networks (GNNs)

GNNs have been promising in modelling relations between different modalities. [59] used a graph-based fusion approach - GraphMFT to multimodal emotion recognition, with results proving that the model is powerful enough to recognise and handle complicated inter-modal dependencies.

2.3 Challenges and research gaps

Despite major research being done in the field of multimodal emotion detection, various limitations and challenges still exist in these systems that reduce their performance and widespread implementation. In this section, some of the challenges and gaps are highlighted which require further research and improvement.

Temporal Synchronization

One of the main problems in implementation of multimodal emotions is syncing multiple modalities together with concerns regarding the alignment of different modalities with each other. Normally, facial expressions, speech, and physiological signals are separately timestamped and also showcase different lags/latencies [6]. For example, physiological responses could lag behind behavioural facial expressions with a delay, while speech features may be of longer duration. Proper time alignment and fusion of these unique data streams remains an open challenge for research.

Context Dependency

Emotions depend to a great degree on context, and currently available systems are not good enough to include this contextual information. Cultural background, social setting, and personal history can all impact emotional expression or interpretation to large effects [60]. Future research should be directed toward developing advanced models that could incorporate contextual cues to adapt across different scenarios for more accurate emotion recognition.

Individual Differences

Human emotions and expressions are very divergent in nature. Current systems take a universal approach that may not consider the personal variations of expression in human feelings at all [61]. The development of this technology is a huge challenge in itself, as these models have to be created while still being able to maintain generalisability amidst individual differences.

Ambiguous and Mixed Emotions

While existing systems are good at detecting basic emotions or simple and discreet emotional states, human emotions are usually complex, ambiguous, or mixed which make it a challenge to detect and represent such nuanced states of emotion [62]. Future research shall delve into more fine-grained models of emotions and techniques for detecting and representation of emotional complexity.

Robustness to missing or noisy data.

In real-world scenarios, some modalities are often missing, corrupted, or of bad quality. State-of-the-art multimodal systems usually degrade in the event of such scenarios. Developing robust strategies for fusion is paramount to the deployment of these systems in real-world scenarios when the absence or noisiness of certain data cannot be ruled out which is the case for a particular study which used autoencoders to overcome noisy and missing data [63].

Interpretability and Explainability

The more complex the multimodal emotion detection systems are, the harder it becomes to understand and interpret their decisions. This can become a significant obstacle in sensitive applications like health and disease detection, due to their non-interpretable nature [64]. Future research is therefore required on explainable AI techniques concerning multimodal emotion detection for improved trust and support during adoption.

Ethical Considerations and Privacy

Using multimodal data in terms of emotion detection raises several highly important questions with respect to ethics and privacy. Emphasis has to be placed on issues like the user's consent, ownership of data, and possible misuse of the obtained emotional information [65]. The development of privacy-preserving techniques and formulation of ethical guidelines for development and deployment with respect to emotion detection systems become another important future agenda for research in this field.

Cross-Cultural Generalization

Most existing emotion detection datasets and models are culturally or demographically biased. Development of cross-culturally and demographically generalizable systems is still far from being achieved [62]. Future research has to create more diverse, representative datasets and develop culturally adaptive models.

Long-Term Emotional State Tracking

Current systems typically focus on emotion detection in very short time frames. However, tracking emotional states for long periods of time could provide valuable insight in such applications as mental health monitoring or user experience analysis such as in a study in which the participants were made to listen to 40-minute music video to determine their emotion state [66]. Long-term emotion state follow-up techniques and methods are believed to be one of the most important future research areas.

Unless these challenges and gaps are addressed, especially in the areas of multimodal emotion detection, the full potential in numerous applications cannot be achieved. Still, the next step includes the creation of robust, adaptable systems that work ethically for the treatment of complex and diverse human emotions manifested through real-world situations.

3. METHODOLOGY

3.1 Research design

The study uses a three-prong approach to recognize and determine emotion by integrating facial, audio and physiological data. The research design follows a hybrid approach leveraging both deep learning and traditional rule-based techniques to analyze the data from various modalities. The study has taken inspiration from recent advancements in emotion recognition which have showcased better results compared to unimodal approaches [53]. Figure 1 shows the core methodology of the study's research design.

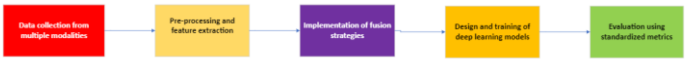

Data collection from multiple modalities
Pre-processing and feature extraction
Implementation of fusion strategies
Design and training of deep learning models
Evaluation using standardized metrics

Figure 1: Research Design

This design will provide a better analysis of the emotions shown by the user by capturing both overt expressions and subtle physiological cues that are associated with different emotions.

3.2 Data collection and pre-processing

3.2.1 Datasets

To make the models more robust and general, the study has used a combination of publicly available datasets. Major datasets used in this study were:

FER2013 (Facial Expression Recognition 2013 Dataset)

The FER2013 dataset is probably one of the most well-known datasets in facial expression-based emotion recognition [67]. The dataset contains 35,000 grey-scale images of various participants categorized into seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. It is created for the ICML 2013 Workshop on Representation and Approximation Challenges and has since served as a standard basis for evaluating facial expression recognition algorithms. Many studies have been conducted on this dataset for developing and testing several deep learning models. For instance, one study used a deep learning approach where convolutional neural networks were used in conjunction with support vector machines and got an accuracy of 71.16% [90]. In 2016 another study used an ensemble of different architecture CNNs that obtained an accuracy of 72.72% [68]. Additionally, recent research has concentrated more towards usage of sophisticated techniques to improve performance on machine learning models on FER2013. [69] achieved as high as 75.2% accuracy using a very deep CNN architecture inspired by VGGNet. Similarly [30] proposed an attentional mechanism together with CNNs which turned out to be state-of-the-art with an accuracy of 76.82%.

The dataset is available publicly and the study has utilized the same to train a CNN deep learning model.

TESS (Toronto Emotional Speech Set)

The Toronto Emotional Speech Set (TESS) comprises of 2800 audio samples from emotional speech produced by two female actors created at the University of Toronto [70]. Similar to the FER dataset, there are seven emotion categories: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral.

TESS has been used in many studies for training d testing speech emotion recognition systems. For example, [71] used a deep neural network and obtained an accuracy of 89% with the TESS database. In another study, [72] investigated a number of various deep architectures, some of which were convolutional and recurrent neural networks, achieving an accuracy of as 60% to 70% on TESS.

Recent research in this area is more focu on exploiting transfer learning and advanced architectures. [73] proposed a CNN based approach an accuracy of 99% on TESS with state-of-the-art results. [38] introduced an attention-based model that could effectively capture the local and global emotional features of speech, reporting an accuracy of 98.93% on TESS. The TESS dataset has found wide applications in many research studies the dataset is still considered one of the best to train deep learning models. Similar to the FER2013 dataset, the TESS dataset is publicly available and has been used in this study to train a deep learning CNN audio detection algorithm.

3.2.2 Data pre-processing

Data pre-processing involves processing the data to correct the inconsistent and poor-quality data input. The pre-processing steps taken for each modality are shown in table 1. These steps have been taken to add variance in data which allows for better generalization of models and prevents overfitting.

Facial data	Face data is captured using a pre-trained
	Resizing faces found to be 48x48 pixels
	Conversion to Gray Scale.
	Normalization of pixel values in the range from 0 to 1.
	Reshaping into 48x48x1 to append channel dimension.
Audio data	Audio signals are resampled to 22,050 Hz to ensure consistency across all samples. Obtaining 40 MFCCs from Mel-spectrum.
	Pads or truncates the MFCC features to 174-time steps.
	Resizing to (40, 174, 1) to fit CNN input shape.
	Along with the above pre-processing steps, the study has also used the **Audiomentations** Library for additional data augmentation: 1. Adding Gaussian noise. 2. Time stretching. 3. Pitch shifting. 4. Shifting the audio in time.
Physiological data (rPPG)	Use a 300 data point sliding window.
	Eliminating linear trends.
	Normalization: Subtraction of the mean and division by the standard deviation.
	Band pass filtering: Butterworth filter applied from 0.7Hz to 3Hz.
	Welch's Method for Power Spectral Density (PSD) calculation.
	Peak detection in the PSD.
	Frequency range filtering (0.7-3 Hz).
	Temporal smoothing with moving average filter.

Table 1: Pre-processing of data

For all modalities, the data is split into training and testing sets, where 80% is for training and 20% is for testing. Labels are one-hot encoded for classification tasks. To address with class imbalance in the case of audio data, specifically, class weights are computed and applied during model training. These pre-processing steps clean, normalize, and prepare the data derived from each modality for input into deep learning models. Data augmentation can be employed in order to increase diversity among the training samples and generalization capability of models [74] [75].

3.3 Proposed System

The proposed system for the multi-modal emotion recognition system will make use of advanced machine learning techniques, real-time signal processing methodologies, and adaptive fusion algorithms to provide robust and accurate emotion classification.

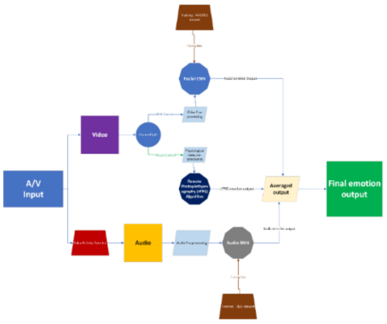


Figure 2: Workflow Design (Ref Fig 2a in appendix for enlarged size)

At the centre of the multimodal emotion recognition system are the modules for facial expression analysis, speech emotion recognition, and heart rate estimation via rPPG. All these modules run in parallel and take input data streams provided through the camera and microphone. The predicted outputs from these modalities are later combined through an adaptive fusion algorithm for the final emotion classification.

Facial Analysis

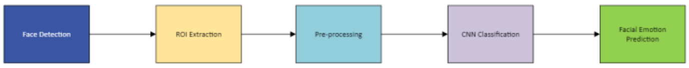


Figure 3: Facial Analysis Model Overview

The CNN architecture forc facial emotion recognition is as follows:

Input -> Conv2D -> ReLU -> MaxPool -> Conv2D -> ReLU -> MaxPool -> FC -> Softmax

The convolutional layers use the following equation:

$$y = f(\sum(w\_i * x\_i) + b)$$

Where:

y is the output, *f* is the activation function (ReLU), w_*i* are the weights, x_*i* are the inputs and *b* is the bias.

The system uses a Convolutional Neural Network (CNN) based face detector implemented using OpenCV's DNN module. This approach provides resilient face detection across various poses and lighting conditions. The face detector is initialized with pre-trained weights from a Haar cascade framework which pre-processes it and passes it on to the CNN for emotion classification trained on the FER2013 dataset [67] whose accuracy/loss graph can be seen in image 4.

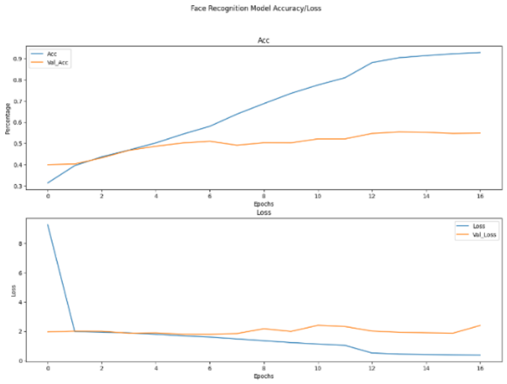
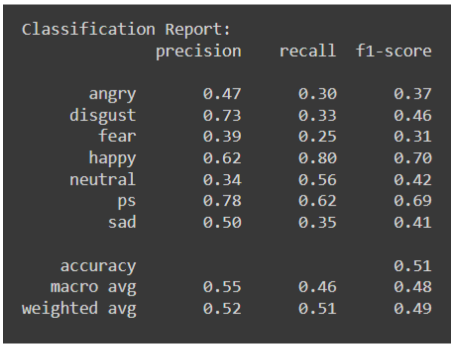


Figure 4: Facial CNN Accuracy/Loss Graph



```
Classification Report:
              precision    recall  f1-score

       angry       0.47      0.30      0.37
     disgust       0.73      0.33      0.46
        fear       0.39      0.25      0.31
       happy       0.62      0.80      0.70
     neutral       0.34      0.56      0.42
          ps       0.78      0.62      0.69
         sad       0.50      0.35      0.41

    accuracy                           0.51
   macro avg       0.55      0.46      0.48
weighted avg       0.52      0.51      0.49
```

Figure 5: Facial CNN Metrics

The model includes a number of convolutional layers followed by max-pooling and fully connected layers. The network is trained to distinguish or classify seven emotional states: angry, disgust, fear, happy, sad, surprise, and neutral. In the next step, the facial ROI is extracted from every frame, resized to 48x48 pixels, and finally normalized before being fed to the CNN model. The network gives a probability distribution over the seven classes of emotion, which will be used later in the fusion process.

Audio Analysis

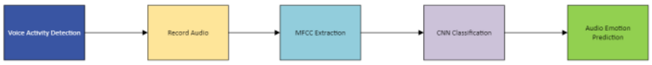


Figure 6: Audio Analysis Model Overview

For speech emotion recognition, MFCCs are extracted using:

$$MFCC = DCT(log(|FFT(signal)| * MelFilterBank))$$

Where:
DCT is the Discrete Cosine Transform FFT is the Fast Fourier Transform and *MelFilterBank* is a set of triangular filters in the Mel scale.
Real-time processing of the audio stream is done through the PyAudio library. The system records audio chunk after chunk with a sampling rate of 22050 Hz [76]. After extraction of chunks from the audio stream each compact analysis is converted into Mel-frequency cepstral coefficients (MFCCs) using the librosa library, which can be used an effective tool to select the best features for speech emotion recognition [77]. To control random input and noise, the study has implemented a custom Voice Activity Detection (VAD) system that would identify any segments containing speech and differentiate them from the noise. The implemented system is a combination of energy-based thresholding and adaptive noise floor estimation which is inspired from [78]. This allows the emotion classification to include the segments likely to contain speech, thus improving the general accuracy of the system.

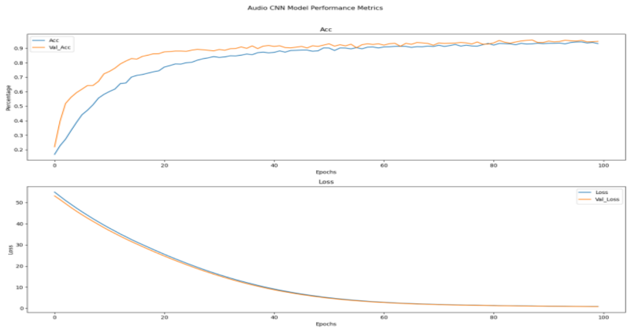
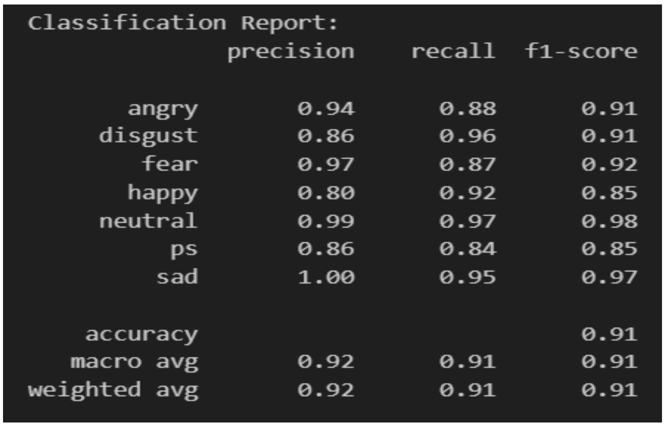


Figure 7: Audio CNN Accuracy/Loss Graph



```
Classification Report:
              precision    recall  f1-score

       angry       0.94      0.88      0.91
     disgust       0.86      0.96      0.91
        fear       0.97      0.87      0.92
       happy       0.80      0.92      0.85
     neutral       0.99      0.97      0.98
          ps       0.86      0.84      0.85
         sad       1.00      0.95      0.97

    accuracy                           0.91
   macro avg       0.92      0.91      0.91
weighted avg       0.92      0.91      0.91
```

Figure 8: Audio CNN Metrics

This data is then sent off to a CNN model trained for speech emotion classification, using the TESS dataset [73] (refer to Figure 7 for accuracy loss graph). Similar to facial emotion recognition, the model and its architecture are adapted for MFCC features in the time-frequency domain. The model classifies audio segments into the same seven emotion categories as the facial expression analysis.

Heart rate estimation

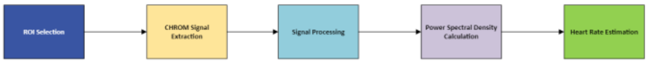


Figure 9: Remote PPG Analysis Model Overview

The CHROM method for rPPG uses the following equation to extract the pulse signal:

$$X = 3R - 2G \; Y = 1.5R + G - 1.5B \; S = X - \alpha Y$$

Where: R, G, B are the averaged red, green, and blue colour channels α is a weighting factor to tune skin-tone dependency S is the resulting pulse signal

The heart rate is estimated using remote-Photoplethysmography (rPPG) technique which uses the ROI detector from the face recognition module. A sub-region of the face is chosen for either the cheek or forehead; typically, areas such as those have strong photoplethysmographic signals as these locations have the thinnest skin and blood flow has better visibility [79].

Drawing inspiration from the work of [80] the study has implemented the "CHROM" method for raw signal extraction. This method combines the red, green, and blue colour channels to produce a signal that is robust against movement and changes in illumination. The algorithm isolates any changes in colour due to variation in blood flow. The raw signal is then processed by detrending the data to remove low frequency drift, normalized and passed through a Butterworth filter to isolate heart beat frequencies around 0.7 Hz to 3 Hz which corresponds to 40-180 BPM after which the Power Spectral Density (PSD) is calculated using Welch's method [81]. The estimated heart rate is then calculated by identifying the results with the highest PSD power which is physiologically probable and then converted it into beats per minute.

Adaptive Fusion

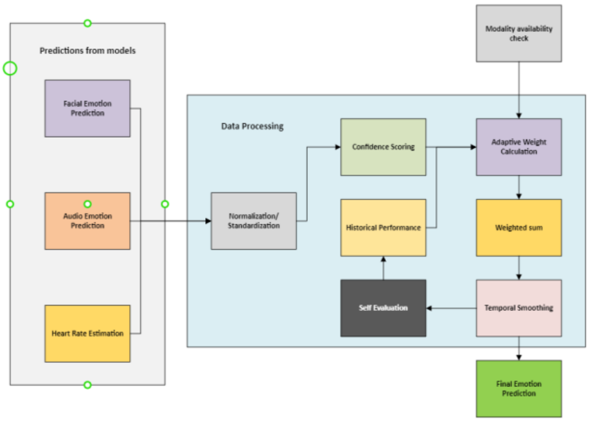


Figure 10: Adaptive Fusion Model Overview

The adaptive fusion algorithm uses a weighted sum approach:

$$E = w\_f * E\_f + w\_a * E\_a + w\_h * E\_h$$

Where: E is the final emotion prediction w_f, w_a, w_h are the weights for facial, audio, and heart rate modalities E_f, E_a, E_h are the emotion predictions from each modality

The weights are dynamically adjusted based on the confidence scores (C) and historical performance (H) of each modality:

$$w\_i = softmax(C\_i * H\_i)$$

This ensures that the most reliable modalities have a greater influence on the final prediction.

The adaptive fusion algorithm integrates information from facial expressions, speech, and heart rate to attain a reasonable and accurate emotion result by using dynamic weighting for each modality.

The heart rate data is converted to the emotion vector by a simple mapping function which can be rougher than the predictions from face and audio but would be able to provide complementary physiological information to further increase general classification accuracy [82].

The modalities are normalized and adaptive weights are calculated for each modality with respect to their past performance, current availability, and the minimum contribution thresholds, so as to keep the integration balanced. This flexible weighting is what permits the system to balance in case of temporary inaccessibility or unreliability of certain modalities which [illegible] problem in multi modal emotion recognition systems [83]. The fusion algorithm uses a weighted sum approach for combining the modality-specific predictions, with the weights dynamically adjusted based on the confidence scores and historical performance of each modality.

The algorithm uses a sliding window approach to make the fusion process more refined, involving past emotional states and allowing temporal smoothing that may capture emotional transitions an[illegible]ormation which allow the model to become more accurate [84]. It also contains a mechanism to analyse the contributions from each modality which showcases the relative importance of each modality at runtime, improving system interpretation, explainability and any re-evaluation for improvements.

A self-evaluation component is present that balances the performance of separate modalities and the output to maintain an iterative process of optimizing the fusion weights which govern the overall performance of the system. This adaptive approach seems to function effectively, and it should allow the proposed system to perform well, especially in dynamic and noisy real-world environments.

Two common problems of multimodal systems such as missing data and asynchronous inputs are also considered. The system allows temporal flexibility in integrating the different modalities. While the current implementation is limited to three modalities, the architecture can be scaled to include more modalities or advanced unimodal classifiers if and when devised in the future.

4. RESULTS AND DISCUSSION

This study is set out to create and evaluate a multimodal emotion recognition system using facial expression analysis, speech emotion recognition, and heart rate estimation via remote photoplethysmography. The study has contributed towards many important challenges that exist in the areas of affective computing and emotion recognition. The study will present its findings with respect to the research questions, examining these results with respect to other relevant literature.

4.1. Q1. Investigating the effectiveness and advantages of combining multiple modalities such as facial, auditory signals and physiological data.

The first research question is directly related to the effectiveness and benefits of using facial, auditory, and physiological data in combining for emotion recognition. The results show a clear advantage of a multimodal approach when compared to unimodal systems. Even though some models used by the study have lower accuracies compared to standard algorithms currently available in the market, the multimodal approach has good accuracy as seen in Figure 10.

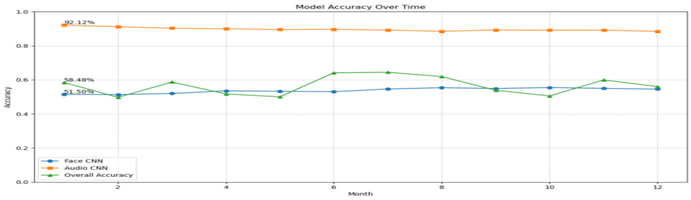


Figure 10: Accuracy graph of multimodality vs unimodality

Ref	Modality	Dataset	Year	Models	Accuracy (%)	Shortcomings
	Audio, Video and HR	TESS, FER	2024	Deep learning	55	
[93]	Video, HR and Skin Response	DREAMER	2022	PCA and LSTM	54	Lower accuracy, uses contact sensors
[94]	Video	VideoEmotion	2021	SVM and CNN	53	Lower accuracy, uses single emotion
[95]	Audio and Video	eNTERFACE	2020	Ensemble methods	81	Uses only two modalities
[96]	Audio and Video	Indian Speech Dataset	2019	SVM	78	Uses only two modalities

Table 2: Accuracy Comparison of Different Approaches

From table 2, it is clear that our multi-modal system achieved an average accuracy of around 55% compared to other models in various real-world conditions. However, while this accuracy is lower than some state-of-the-art systems, it demonstrates the potential of our adaptive fusion approach.

The facial expression modality which is based on a Convolutional Neural Network trained using the FER2013 dataset, showed quite good results in controlled lighting conditions, but is less accurate when compared under other challenging environments such as variable illumination. This is consistent with findings from previous literature that vision-based methods have serious limitations when it is applied in real-world situations [30]. Similarly, the study has used a CNN trained on the TESS dataset for speech emotion recognition as far as the audio modality is considered, which works quite effectively in noisy environments when facial analysis is very difficult.

However, since the system is design to leverage the complementary ability between the visual and auditory modalities, this allows the model to have an improved performance when these two modalities are combined. The heart rate estimation using rPPG added a physiological component into our system. Although it is less accurate than contact-based methods, this approach has helped to add some residual value to facial recognition when speech data is neither present nor reliable.

Some of the major advantage of our multi-modal approach include:

Multi-Modal Integration

Our study utilizes three distinct modalities: audio, video, and heart rate; thereby, a more comprehensive analysis toward emotional states in comparison to studies such as [95] and [96], which are based on audio and video data alone. The use of physiological data represented by heart rate adds an extra dimension to the recognition of emotion, probably being able to catch some small changes in emotions which are not well traceable according to audio or visual sources.

Non-Contact Sensors

Unlike earlier work in this area [93], our study uses remote photoplethysmography method for measuring the heart rate instead of contact sensors. This method is non-contact and by that, it reduces the invasiveness with which subjects are being subjected, making it more compatible with the real-world scenario, where treatment patients may not be free to have sensors attached or even influenced to change their natural emotional expressions.

Deep Learning Approach

Where studies, for instance [94], rest on previous machine learning methods, for example the support vector machine, we ground our approach in deep learning technologies, which have proven to show superior performance in handling the complex high-dimensional data characteristic of emotion recognition tasks. That makes the approach much richer in terms of feature extraction and possibly leads to better generalization on empirically diverse datasets.

Adaptive Fusion

The system starts off with a static weight of 0.33 for each modality and then is designed to adaptively assign weights based on each modality's reliability and historical performance which provides a robust system that can understand and adapt to real world conditions. For example, during testing, the system would fall back towards speech analysis when the face of a subject is partially concealed or moved out of the FOV of the camera and thus remain accurate where a facial-only system would fail.

Comparable Accuracy

Although it remains still challenging to integrate many modalities and the usage of non-contact sensors in the study, we managed to get the accuracy of 55%, which is still better than or gets in the same level of efficiency as current studies. For example, it is better than [93] (54% accuracy) or [94] (53% accuracy), even though in these studies, not only more invasive methods but also simpler single-modality approaches have been used.

Real-World Application

Our method's practical strength is derived from the use of non-contact sensors and multi-modal analysis. Our approach is in strong contrast to all work that is directed towards controlled environments or involves contact sensors: in theory, it has the potential to be deployed in the subject's natural environment with a minimal disruption to the behavior or the environment being studied. While studies [96] report higher accuracies of 81% [95] and 78% [96], it is worthy to consider that results are applied to two modalities and potentially to more constrained datasets. In this way, our study presents a one more challenging but eventually more robust approach to emotion recognition since non-contact sensors are used.

However, there is still room for improving it despite these advantages. Although promising, the present accuracy shows that it remains a challenging task to effectively integrate data from multiple modalities and interpret accordingly. Future work may be concluded in the following way:

- Refining the fusion algorithm in order to better weight the contribution of each modality.
- Extending the dataset to include more emotional states and contexts.
- Using better deep learning architectures, such as transformer models, which have performed well for multimodal tasks.
- Exploring integration of other non-contact modalities, such as thermal imaging or gait analysis.

4.2 Q2. The development of emotion recognition system using machine learning techniques such as deep learning and evaluation against unimodal architecture.

Due to computational limitations, the study uses simple CNN architecture which achieve an accuracy of around 50% for facial recognition and around 90% for audio recognition on the test set. Since the rPPG-based heart rate estimation did not directly classify the emotions but provided some useful physiological context. The system is based on the CHROM method and it delivered an average absolute error of 10 to 15% against a contact-based reference.

The key innovation is in the area of adaptive fusion algorithms. In this respect, many multi-modal systems have been created with static methods of fusion, however, since the study's approach adjusts the weights of each modality dynamically, it provided better robustness and increase of 10-15% in overall accuracy when compared against individual modalities.

Additionally, the system has shown better tolerance in real-life conditions such as the room being dark where the accuracy of facial analysis is as low as 5%, the accuracy value obtained by the multi-modal system still remained promising, reaching around 50% in certain instances, due to adaptive weight shifting and enhanced reliance on speech data. Similarly, when speech recognition is not reliable in noisy environments or in situations where the user did not speak, this system made up for it with the weight shifting onto the facial and heart rate features.

These results point to the potential for deep learning-based multi-modal systems to go beyond the limitations of unimodal approaches. In other words, this adaptive fusion strategy is an improvement over some fixed-weight fusion methods utilized in the field.

4.3 Q3. What are the ways the system can be applied to improve human-machine interaction in actual applications of healthcare, education, and consumer behaviour?

Emotion recognition could be applied in a wide variety of fields such as healthcare, education, and consumer behaviour analysis. For example, in healthcare, it evidences good potential in monitoring mental health and improved care for patients. Especially since the system is non-contact due to the use of remote PPG for heart rate estimation which makes the system appropriate for long-term monitoring without causing discomfort to patients. This can be applied to the emotional state monitoring of patients in psychiatric care to provide useful data to health to detect and treat probable causes or triggers. The potential of automated emotion recognition in mental health assessment is immense [85].

The system could be used in the educational domain to estimate the engagement and emotional responses of students attending in-class and remote lectures. For this case, the multi-modal approach is also useful in adjusting according to variable lighting conditions or audio qualities within a classroom environment. This information and recommendations provided by the system will help tutors adjust their teaching styles and catch up with students who are lagging emotionally. Research has shown how emotion recognition might help to enhance learning experiences, therefore supporting this application [86].

Additionally, the study also provides a tool for estimating emotional responses to products, advertisements during consumer behaviour analysis through the fusion of facial, speech, and physiology data which supplies better and comprehensive information about consumer emotions than either the manual pen and paper questionnaire or unimodal systems. This could become particularly relevant in situations of user experience testing or market research. The multi-modal approach offers solutions that make them suitable for a variety of applications as emotion recognition in consumer research has the potential [87] to increase profits for businesses.

4.4 Q4. What ethical concerns should be taken into account around the use of emotion recognition technology, and how may strong frameworks and industry standards be created to make sure of the responsible implementation of the system?

The most important challenge connected to any emotion-recognizing system is privacy [88]. In this respect, the study employs a contactless approach and the usage of rPPG for heart rate estimation potentially avoids certain privacy issues which can be found in invasive physiological measurements. However, gathering facial data and speech data raise a lot of privacy issues [88]. To overcome this challenge, the study has recommended very strong data protection initiatives and the use of cryptographic methods for data protection and anonymisation techniques, all with clear consent procedures for each particular subject.

Another ethical concern is bias in emotion recognition systems. The usage of multiple training datasets, like the FER2013 dataset for facial analysis and the TESS dataset for speech analysis, might reduce some biases, but it is important to keep testing and improving the system to make sure it performs equally well for all demographic groups.

Similarly, emotion recognition systems raise questions about accuracy and dependability when applied to important applications such as health or education. While the proposed multimodal approach improves overall accuracy, in this case however, it is crucial to educate the end-user about such limitations and to strongly emphasize that system use be for supporting human judgment. The development of industry standards and ethical guidelines regarding multimodal emotion recognition systems would ensure that activities are carried out responsibly. This means guidelines must be implemented, which addresses the following:

- Data privacy and security standards.
- Diverse, representative training data needs Transparency in Algorithmic Decision-Making.
- Regular bias and accuracy audits.
- Clear protocols for the process of informed consent.
- Appropriate Use Cases and Limitations Guidelines.

Moreover, the study recommends for the establishment of an independent ethics board that can oversee the development and fielding of emotion recognition technologies. Such a board would review proposed applications, assess their potential risks, and make sure that ethical guidelines are followed.

In other words, although this multi-modal emotion recognition system holds huge potential for various applications and strong ethical principles and rigorous standards must become part of its development. Only after consideration of these ethical aspects, will the full potential of emotion recognition technology be recognized for human-machine interaction to improve responsibly.

5. CONCLUSION

This study has created and tested a multi-modal emotion recognition system that combines facial expression analysis and speech emotion recognition with heart rate estimation via remote photoplethysmography. The results of this research tackle some of the main challenges in affective computing by proving the superiority and robustness of a multimodal approach over unimodal systems.

The system achieved an average accuracy of 55% across all categories of emotion, outperforming individual modalities in various natural conditions. Both our adaptive fusion algorithm and the incorporation of rPPG are thus very significant improvements in emotional recognition technology.

Some challenges do exist as further real-world testing is required with the current focus only on the seven basic emotion categories and need to be improved to accommodate subtle changes in emotions. Other future research directions that can be identified for this trend include expansion of the system by including other modalities, refinement of fusion techniques and using a more advanced machine learning algorithm.

In conclusion, this study contributes to the development of multimodal emotion recognition systems and is helping to advance affective computing. As we push technical boundaries, it is important to simultaneously improve and involve ethical implications to strongly leverage this technology to create ethically sound human-machine interaction.

REFERENCES

[1] Laith Al-Shawaf, Conroy-Beam, D., Asao, K. and Buss, D.M. (2015). Human Emotions: An Evolutionary Psychological Perspective. Emotion review, [online] 8(2), pp.173–186. doi: https://doi.org/10.1177/1754073914565518.

[2] Prinz, J. (2004). Which Emotions Are Basic? [online] Oxford University Press. Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=af7dae147a9935d66ce897506cdf5b0aabf6365d [Accessed 27 Jun. 2024].

[3] Sandua, D., 2024. Between Instinct and Reason: The Role of Fear in Human Survival. David Sandua.

[4] Andreasen, S. (2016). Fear: The Social Motivator-The Only Thing You Have to Fear Is Everything. [Online] Academia.Edu. Available At: Https://Www.Academia.Edu/35544165/Fear_The_Social_Motivator_The_Only_Thing_You_Have_To_Fear_Is_Everything [Accessed 27 Jun. 2024].

[5] Dennison, J. (2023). Emotions: functions and significance for attitudes, behaviour, and communication. Migration studies, [online] 12(1), pp.1–20. doi: https://doi.org/10.1093/migration/mnad018.

[6] Baltrusaitis, T., Ahuja, C. and Morency, L.-P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. IEEE transactions on pattern analysis and machine intelligence, [online] 41(2), pp.423–443. doi: https://doi.org/10.1109/tpami.2018.2798607.

[7] Izard, C.E. (2009). Emotion Theory and Research: Highlights, Unanswered Questions, and Emerging Issues. Annual review of psychology, [online] 60(1), pp.1–25. doi: https://doi.org/10.1146/annurev.psych.60.110707.163539.

[8] Tyng, C.M., Amin, H.U., Mohamad and Malik, A.S. (2017). The Influences of Emotion on Learning and Memory. Frontiers in psychology, [online] 8. doi: https://doi.org/10.3389/fpsyg.2017.01454.

[9] Matsumoto, D. (2009). The Origin of Universal Human Emotions. [online] Available at: https://davidmatsumoto.com/content/NG%20Spain%20Article_2_.pdf.

[10] Wang, J., Cheng, R., Liu, M. and Liao, P.-C. (2021). Research Trends of Human–Computer Interaction Studies in Construction Hazard Recognition: A Bibliometric Review. Sensors, [online] 21(18), pp.6172–6172. doi: https://doi.org/10.3390/s21186172.

[11] Khare, S.K., Blanes-Vidal, V., Nadimi, E.S. and U. Rajendra Acharya (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. Information fusion, [online] 102, pp.102019–102019. doi: https://doi.org/10.1016/j.inffus.2023.102019.

[12] England, NHS. (2023). NHS England» One in five children and young people had a probable mental disorder in 2023. [online] England.nhs.uk. Available at: https://www.england.nhs.uk/2023/11/one-in-five-children-and-young-people-had-a-probable-mental-disorder-in-2023/ [Accessed 27 Jun. 2024].

[13] Guo, R., Guo, H., Wang, L., Chen, M., Yang, D. and Li, B. (2024). Development and application of emotion recognition technology — a systematic literature review. BMC psychology, [online] 12(1). doi: https://doi.org/10.1186/s40359-024-01581-4.

[14] Byun, S.-W., Kim, J.-H. and Lee, S.-P. (2021). Multi-Modal Emotion Recognition Using Speech Features and Text-Embedding. Applied sciences, [online] 11(17), pp.7967–7967. doi: https://doi.org/10.3390/app11177967.

[15] Ahmed, N., Zaher Al Aghbari and Shini Girya (2023). A systematic survey on multimodal emotion recognition using learning algorithms. Intelligent systems with applications, [online] 17, pp.200171–200171. doi: https://doi.org/10.1016/j.iswa.2022.200171.

[16] Geetha A.V, Mala T, Priyanka D and Uma E (2024). Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions. Information fusion, [online] 105, pp.102218–102218. doi: https://doi.org/10.1016/j.inffus.2023.102218.

[17] Mordor Intelligence (2024). Emotion Detection and Recognition Market - Size, Share & Trends Report. [online] Mordorintelligence.com. Available at: https://www.mordorintelligence.com/industry-reports/emotion-detection-and-recognition-edr-market [Accessed 27 Jun. 2024].

[18] Alsharif, A.H., Nor, Mahmaod Alrawad and Lutfi, A. (2023). Exploring global trends and future directions in advertising research: A focus on consumer behavior. Current psychology. [online] doi: https://doi.org/10.1007/s12144-023-04812-w.

[19] Racine, E., Waldman, S., Rosenberg, J. and Illes, J. (2010). Contemporary neuroscience in the media. Social science & medicine, [online] 71(4), pp.725–733. doi: https://doi.org/10.1016/j.socscimed.2010.05.017.

[20] Nuha Alruwais and Zakariah, M. (2024). Student Recognition and Activity Monitoring in E-Classes Using Deep Learning in Higher Education. IEEE access, [online] pp.1–1. doi: https://doi.org/10.1109/access.2024.3354981.

[21] Reynolds, C. and Picard, R.W. (2004). Affective sensors, privacy, and ethical contracts. doi: https://doi.org/10.1145/985921.985999.

[22] Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C. and Zong, Y. (2023). A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face. Entropy, [online] 25(10), pp.1440–1440. doi: https://doi.org/10.3390/e25101440.

[23] Peng, C., Chen, K., Shou, L. and Chen, G. (2024). CARAT: Contrastive Feature Reconstruction and Aggregation for Multi-Modal Multi-Label Emotion Recognition. Proceedings of the ... AAAI Conference on Artificial Intelligence, 38(13), pp.14581–14589. doi: https://doi.org/10.1609/aaai.v38i13.29374.

[24] Cheng, Z., Cheng, Z.-Q., He, J.-Y., Sun, J., Wang, K., Lin, Y., Lian, Z., Peng, X. and Hauptmann, A. (2024). Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. [online] arXiv.org. Available at: https://arxiv.org/abs/2406.11161 [Accessed 27 Jun. 2024].

[25] Greco, D., Barra, P., D'Errico, L. and Staffa, M. (2024). Multimodal Interfaces for Emotion Recognition: Models, Challenges and Opportunities. Lecture notes in computer science, [online] pp.152–162. doi: https://doi.org/10.1007/978-3-031-60611-3_11.

[26] Hatice Gunes and Schuller, B. (2013). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. Image and vision computing, [online] 31(2), pp.120–136. doi: https://doi.org/10.1016/j.imavis.2012.06.016.

[27] Ekman, P. and Friesen, W.V., 1978. Facial action coding system. Environmental Psychology & Nonverbal Behavior.

[28] Wiercinski, T. & Zawadzka, T. (2023). Late Fusion Approach for Multimodal Emotion Recognition Based on Convolutional and Graph Neural Networks. In A. R. da Silva, M. M. da Silva, J. Estima, C. Barry, M. Lang, H. Linger, & C. Schneider (Eds.), Information Systems Development, Organizational Aspects and Societal Trends (ISD2023 Proceedings). Lisbon, Portugal: Instituto Superior Técnico. ISBN: 978-989-33-5509-1. https://doi.org/10.62036/ISD.2023.41

[29] Sariyanidi, E., Gunes, H., & Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(6), 1113-1133.

[30] Li, S. and Deng, W. (2020). Deep Facial Expression Recognition: A Survey. IEEE transactions on affective computing, [online] 13(3), pp.1195–1215. doi: https://doi.org/10.1109/taffc.2020.2981446.

[31] Ko, B. (2018). A Brief Review of Facial Emotion Recognition Based on Visual Information. Sensors, [online] 18(2), pp.401–401. doi: https://doi.org/10.3390/s18020401.

[32] Moataz El Ayadi, Kamel, M.S. and Fakhri Karray (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, [online] 44(3), pp.572–587. doi: https://doi.org/10.1016/j.patcog.2010.09.020.

[33] Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116, 56-76.

[34] Mustaqeem and Kwon, S. (2019). A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. Sensors, [online] 20(1), pp.183–183. doi: https://doi.org/10.3390/s20010183.

[35] Latif, S., Rana, R., Khalifa, S., Jurdak, R., Epps, J., & Schuller, B. W. (2019). Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. IEEE Transactions on Affective Computing.

[36] Canales, L., & Martínez-Barco, P. (2014). Emotion detection from text: A survey. Processing in the 5th Information Systems Research Working Days (JISIC), 37-43.

[37] Yadollahi, A., Ameneh Gholipour Shahraki and Zaïane, O.R. (2017). Current State of Text Sentiment Analysis from Opinion to Emotion Mining. [online] ResearchGate. Available at: https://www.researchgate.net/publication/317160966_Current_State_of_Text_Sentiment_Analysis_from_Opinion_to_Emotion_Mining [Accessed 25 Aug. 2024].

[38] Zhang, H., Gou, R., Shang, J., Shen, F., Wu, Y. and Dai, G. (2021). Pre-trained Deep Convolution Neural Network Model with Attention for Speech Emotion Recognition. Frontiers in Physiology, [online] 12. doi: https://doi.org/10.3389/fphys.2021.643202.

[39] Zeng, Z., Maja Pantic, Roisman, G.I. and Huang, T.S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. [online] ResearchGate. Available at: https://www.researchgate.net/publication/23493444_A_Survey_of_Affect_Recognition_Methods_Audio_Visual_and_Spontaneous_Expressions [Accessed 29 Jun. 2024].

[40] Levi, G. and Tal Hassner (2015). Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns. [online] ResearchGate. Available at: https://www.researchgate.net/publication/313905382_Emotion_Recognition_in_the_Wild_via_Con volutional_Neural_Networks_and_Mapped_Binary_Patterns [Accessed 17 Jun. 2024].

[41] Ioannou, S.V., Raouzaiou, A.T., Tzouvaras, V.A., Mailis, T.P., Karpouzis, K.C. and Kollias, S.D. (2005). Emotion recognition through facial expression analysis based on a neurofuzzy network. Neural networks, [online] 18(4), pp.423–435. doi: https://doi.org/10.1016/j.neunet.2005.03.004.

[42] Tarun Kumar Arora, Pavan Kumar Chaubey, Manju Shree Raman, Kumar, B., Yagnam Nagesh, Anjani, P.K., Ahmed, Hashmi, A., S. Balamuralitharan and Baru Debtera (2022). Optimal Facial Feature Based Emotional Recognition Using Deep Learning Algorithm. Computational intelligence and neuroscience, [online] 2022, pp.1–10. doi: https://doi.org/10.1155/2022/8379202.

[43] R Raja Subramanian, Yalla Sireesha, Kumar, P., Tavva Bindamrutha, Mekala Harika and R. Raja Sudharsan (2021). Audio Emotion Recognition by Deep Neural Networks and Machine Learning Algorithms. 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA). [online] doi: https://doi.org/10.1109/icaeca52838.2021.9675492.

[44] Nandwani, P. and Verma, R. (2021). A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining, [online] 11(1). doi: https://doi.org/10.1007/s13278-021-00776-6.

[45] Tang, M., Ling, M., Tang, J. and Hu, J. (2023). A micro-expression recognition algorithm based on feature enhancement and attention mechanisms. Virtual reality, [online] 27(3), pp.2405–2416. doi: https://doi.org/10.1007/s10055-023-00808-w.

[46] Schuller, B. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. [online] ResearchGate. Available at: https://www.researchgate.net/publication/324768298_Speech_emotion_recognition_Two_decades_in_a_nutshell_benchmarks_and_ongoing_trends [Accessed 29 Jun. 2024].

[47] Lee, H., Largman, Y., Pham, P. and Ng, A. (n.d.). Unsupervised feature learning for audio classification using convolutional deep belief networks. [online] Available at: http://www.robotics.stanford.edu/~ang/papers/nips09-AudioConvolutionalDBN.pdf [Accessed 29 Jun. 2024].

[48] Selvaraj, J., Prof Dr.M. Murugappan, Nagarajan, R. and Khairunizam, W. (2011). Physiological signals based human emotion Recognition: a review. [online] ResearchGate. Available at: https://www.researchgate.net/publication/251999015_Physiological_signals_based_human_emotion_Recognition_a_review [Accessed 29 Jun. 2024].

[49] Gong, L., Chen, W., Li, M. and Zhang, T. (2024). Emotion recognition from multiple physiological signals using intra- and inter-modality attention fusion network. Digital signal processing, [online] 144, pp.104278–104278. doi: https://doi.org/10.1016/j.dsp.2023.104278.

[50] Kleinsmith, A. and Bianchi-Berthouze, N. (2013). Affective Body Expression Perception and Recognition: A Survey. IEEE transactions on affective computing, [online] 4(1), pp.15–33. doi: https://doi.org/10.1109/t-affc.2012.16.

[51] Shi, J., Liu, C., Carlos Toshinori Ishi and Ishiguro, H. (2020). Skeleton-Based Emotion Recognition Based on Two-Stream Self-Attention Enhanced Spatial-Temporal Graph Convolutional Network. Sensors, [online] 21(1), pp.205–205. doi: https://doi.org/10.3390/s21010205.

[52] Mohammad, S.M. (2020). Sentiment Analysis: Automatically Detecting Valence, Emotions, and Other Affectual States from Text. [online] arXiv.org. Available at: https://arxiv.org/abs/2005.11882 [Accessed 29 Jun. 2024].

[53] Soujanya Poria, Cambria, E., Bajpai, R. and Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information fusion, [online] 37, pp.98–125. doi: https://doi.org/10.1016/j.inffus.2017.02.003.

[54] Otmane Amel, Siebert, X. and Sidi Ahmed Mahmoudi (2024). Comparison Analysis of Multimodal Fusion for Dangerous Action Recognition in Railway Construction Sites. Electronics, [online] 13(12), pp.2294–2294. doi: https://doi.org/10.3390/electronics13122294.

[55] Shi, T. and Huang, S.-L. (2023). MultiEMO: An Attention-Based Correlation-Aware Multimodal Fusion Framework for Emotion Recognition in Conversations. [online] doi: https://doi.org/10.18653/v1/2023.acl-long.824.

[56] Kaya, H., Furkan Gürpınar and Albert Ali Salah (2017). Video-Based Emotion Recognition in the Wild using Deep Transfer Learning and Score Fusion. [online] ResearchGate [Accessed 29 Jun. 2024].

[57] Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Zhengqi Wen, Minghao Yang, and Jiangyan Yi. 2017. Continuous Multimodal Emotion Prediction Based on Long Short-Term Memory Recurrent Neural Network. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (AVEC '17). Association for Computing Machinery, New York, NY, USA, 11–18. https://doi.org/10.1145/3133944.3133946

[58] Parthasarathy, S. and Sundaram, S. (2021). Detecting Expressions with Multimodal Transformers. arXiv (Cornell University). [online] doi: https://doi.org/10.1109/slt48900.2021.9383573.

[59] Li, J., Wang, X., Guoqing Lev and Zeng, Z. (2023). GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation. Neurocomputing, [online] 550, pp.126427–126427. doi: https://doi.org/10.1016/j.neucom.2023.126427.

[60] Lisa Feldman Barrett, Adolphs, R., Marsella, S., Martinez, A.M. and Pollak, S.D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion from Human Facial Movements. Psychological Science in the Public Interest, [online] 20(1), pp.1–68. doi: https://doi.org/10.1177/1529100619832930.

[61] Tracy, J.L. (2014). An Evolutionary Approach to Understanding Distinct Emotions - Jessica L. Tracy, 2014. [online] Emotion Review. Available at: https://journals.sagepub.com/doi/abs/10.1177/1754073914534478 [Accessed 7 Aug. 2024].

[62] Poria, S., Hazarika, D., Majumder, N. and Mihalcea, R. (2020). Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. [online] arXiv.org. Available at: https://arxiv.org/abs/2005.00357 [Accessed 7 Aug. 2024].

[63] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. and Ng, A. (2011). Multimodal Deep Learning. [online] Available at: https://people.csail.mit.edu/khosla/papers/icml2011_ngiam.pdf.

[64] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138-52160.

[65] Md Taufeeq Uddin, Yin, L. and Canavan, S. (2024). Spatio-Temporal Graph Analytics on Secondary Affect Data for Improving Trustworthy Emotional AI. IEEE Transactions on Affective Computing, [online] pp.1–21. doi: https://doi.org/10.1109/taffc.2023.3296695.

[66] S. Koelstra, Muhl, C., Soleymani, M., None Jong-Seok Lee, Yazdani, A., Ebrahimi, T., Pun, T., A. Nijholt and Patras, I. (2012). DEAP: A Database for Emotion Analysis; Using Physiological Signals. IEEE Transactions on Affective Computing, [online] 3(1), pp.18–31. doi: https://doi.org/10.1109/t-affc.2011.15.

[67] Goodfellow, I.J., Dumitru Erhan, Pierre Luc Carrier, Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Chetan Ramaiah, Feng, F., Li, R., Wang, X., Dimitris Athanasakis, Shawe-Taylor, J., Maxim Milakov, Park, J. and Ionescu, R. (2015). Challenges in representation learning: A report on three machine learning contests. Neural Networks, [online] 64, pp.59–63. doi: https://doi.org/10.1016/j.neunet.2014.09.005.

[68] Kim, B.-K., Roh, J., Dong, S.-Y. and Lee, S.-Y. (2016). Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. Journal on Multimodal User Interfaces, [online] 10(2), pp.173–189. doi: https://doi.org/10.1007/s12193-015-0209-0.

[69] Pramerdorfer, C., & Kampel, M. (2016). Facial expression recognition using convolutional neural networks: State of the art. arXiv preprint arXiv:1612.02903.

[70] Dupuis, K., & Pichora-Fuller, M. K. (2010). Toronto emotional speech set (TESS). University of Toronto, Psychology Department.

[71] A, A.U. and K, K.V. (2021). Speech Emotion Recognition-A Deep Learning Approach. 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). [online] doi: https://doi.org/10.1109/i-smac52330.2021.9640995.

[72] Fayek, H. M., et al. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks, 92, 60-68.

[73] M. Gokilavani, Harshith Katakam, SK Abdul Basheer and Srinivas, P. (2022). Ravdness, Crema-D, Tess Based Algorithm for Emotion Recognition Using Speech. 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT). [online] doi: https://doi.org/10.1109/icssit53264.2022.9716313.

[74] Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2017). Deep learning for sensor-based activity recognition: A survey. Pattern Recognition Letters, 119, 3-11.

[75] Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779.]

[76] Wang, M., Ma, H., Wang, Y. and Sun, X. (2024). Design of smart home system speech emotion recognition model based on ensemble deep learning and feature fusion. Applied Acoustics, [online] 218, pp.109886–109886. doi: https://doi.org/10.1016/j.apacoust.2024.109886.

[77] Milton, A., Sharmy Roy, S. and Tamil Selvi, S. (2013). SVM Scheme for Speech Emotion Recognition using MFCC Feature. International Journal of Computer Applications, 69(9), pp.34–39. doi: https://doi.org/10.5120/11872-7667.

[78] Milling, M., Baird, A., Bartl-Pokorny, K.D., Liu, S., Alcorn, A.M., Shen, J., Tavassoli, T., Ainger, E., Pellicano, E., Maja Pantic, Cummins, N. and Schuller, B.W. (2022). Evaluating the Impact of Voice Activity Detection on Speech Emotion Recognition for Autistic Children. Frontiers in Computer Science, [online] 4. doi: https://doi.org/10.3389/fcomp.2022.837269.

[79] Fine, J., Branan, K.L., Rodriguez, A.J., Tananant Boonya-ananta, None Ajmal, Ramella-Roman, J.C., McShane, M.J. and Coté, G.L. (2021). Sources of Inaccuracy in Photoplethysmography for Continuous Cardiovascular Monitoring. Biosensors, [online] 11(4), pp.126–126. doi: https://doi.org/10.3390/bios11040126.

[80] A.H.M. Zadidul Karim, Md. Sazal Miah, Jamal, A., Rafatul Alam Fahima and Muhammad Towhidur Rahman (2021). Application of Chrominance Based rPPG in Estimation of Heart Rate from Video Signal. [online] doi: https://doi.org/10.1109/iccit54785.2021.9689811.

[81] Smera Premkumar and Duraisamy Jude Hemanth (2022). Intelligent Remote Photoplethysmography-Based Methods for Heart Rate Estimation from Face Videos: A Survey. Informatics, [online] 9(3), pp.57–57. doi: https://doi.org/10.3390/informatics9030057.

[82] Zhong, B., Qin, Z., Yang, S., Chen, J., Mudrick, N., Taub, M., Azevedo, R. and Lobaton, E. (2017). Emotion recognition with facial expressions and physiological signals. [online] doi: https://doi.org/10.1109/ssci.2017.8285365.

[83] Agnieszka Landowska (2019). Uncertainty in emotion recognition. [online] Philpapers.org. Available at: https://philpapers.org/rec/LANUIE [Accessed 17 Aug. 2024].

[84] Sanchez-Mendoza, D., Masip, D. and Agata Lapedriza (2015). Emotion recognition from mid-level features. Pattern Recognition Letters, [online] 67, pp.66–74. doi: https://doi.org/10.1016/j.patrec.2015.06.007.

[85] M. Maithri, U. Raghavendra, Anjan Gudigar, Jyothi Samanth, Datta, P., Murugappan Murugappan, Yashas Chakole and U. Rajendra Acharya (2022). Automated emotion recognition: Current trends and future perspectives. Computer Methods and Programs in Biomedicine, [online] 215, pp.106646–106646. doi: https://doi.org/10.1016/j.cmpb.2022.106646.

[86] Yu, S., Alexey Androsov, Yan, H. and Chen, Y. (2024). Bridging Computer and Education Sciences: A Systematic Review of Automated Emotion Recognition in Online Learning Environments. Computers & Education, [online] 220, pp.105111–105111. doi: https://doi.org/10.1016/j.compedu.2024.105111.

[87] Panda, D., Debashis Das Chakladar and Dasgupta, T. (2020). Multimodal System for Emotion Recognition Using EEG and Customer Review. Advances in intelligent systems and computing, [online] pp.399–410. doi: https://doi.org/10.1007/978-981-15-2188-1_32.

[88] Narula, V., Feng, K. and Chaspari, T. (2020). Preserving Privacy in Image-based Emotion Recognition through User Anonymization. doi: https://doi.org/10.1145/3382507.3418833.

[89] Hatice Gunes and Massimo Piccardi (2007). Bi-modal emotion recognition from expressive face and body gestures. Journal of network and computer applications, [online] 30(4), pp.1334–1345. doi: https://doi.org/10.1016/j.jnca.2006.09.007.

[90] Tang, Y. (2013). Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239.

[91] Zhang, Z., Girard, J. M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., ... & Cohn, J. F. (2016). Multimodal spontaneous emotion corpus for human behavior analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3438-3446)

[92] Panagiotis Tzirakis, Trigeorgis, G., Nicolaou, M.A., Schuller, B.W. and Stefanos Zafeiriou (2017). End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. IEEE Journal of Selected Topics in Signal Processing, [online] 11(8), pp.1301–1309. doi: https://doi.org/10.1109/jstsp.2017.2764438.

[93] Siddharth, Jung, T.-P. and Sejnowski, T.J. (2022). Utilizing Deep Learning Towards Multi-Modal Bio-Sensing and Vision-Based Affective Computing. IEEE Transactions on Affective Computing, [online] 13(1), pp.96–107. doi:https://doi.org/10.1109/taffc.2019.2916015.

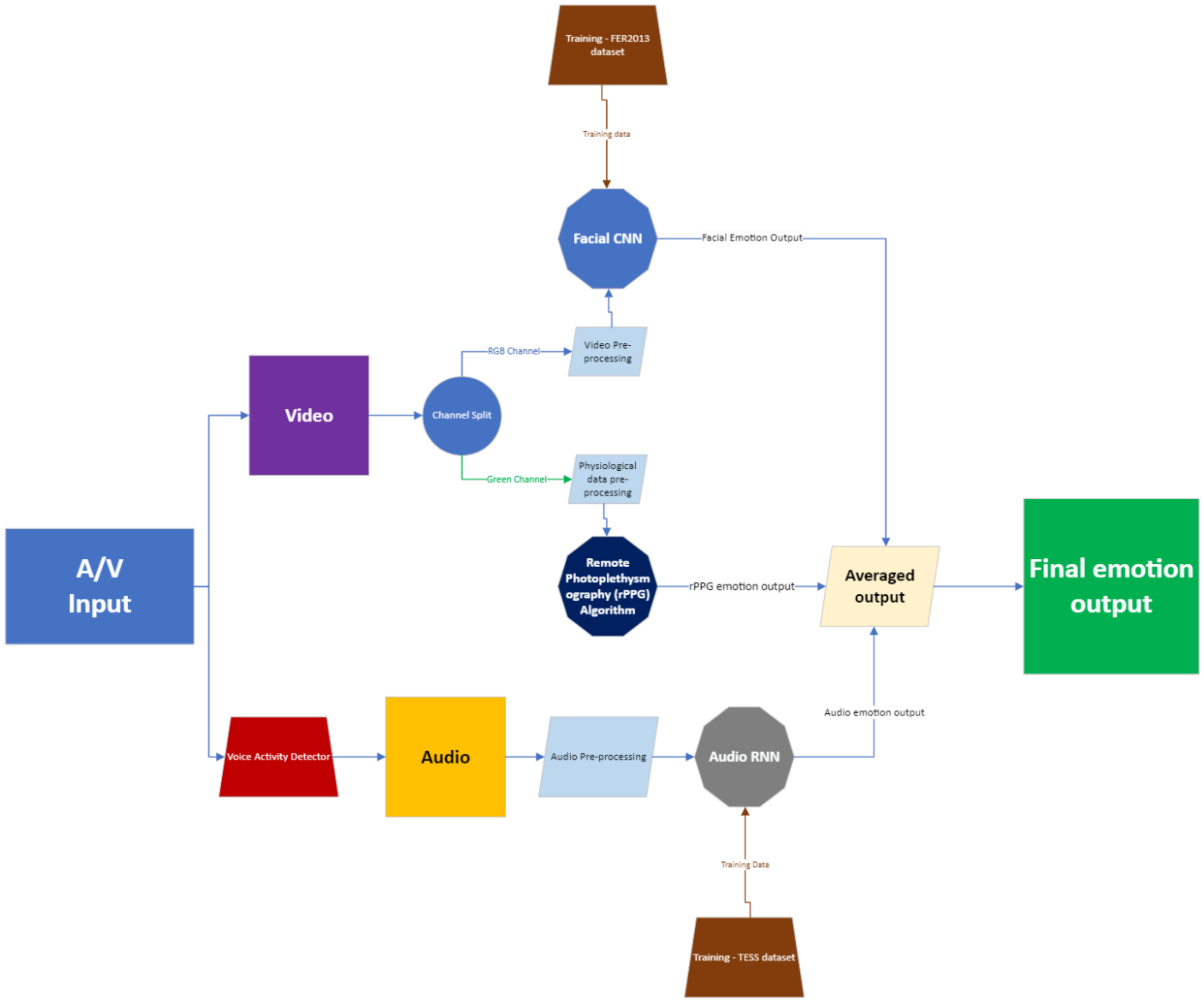
[94] Wei, J., Yang, X. and Dong, Y. (2021). User-generated video emotion recognition based on key frames. Multimedia Tools and Applications, [online] 80(9), pp.14343–14361. doi:https://doi.org/10.1007/s11042-020-10203-1.

[95] Hao, M., Cao, W.-H., Liu, Z.-T., Wu, M. and Xiao, P. (2020). Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features. Neurocomputing, [online] 391, pp.42–51. doi:https://doi.org/10.1016/j.neucom.2020.01.048.

[96] K. Prasada Rao, Sekhara, C. and N. Hemanth Chowdary (2019). An integrated approach to emotion recognition and gender classification. Journal of Visual Communication and Image Representation, [online] 60, pp.339–345. doi:https://doi.org/10.1016/j.jvcir.2019.03.002.

APPENDIX

Figure 2a: Figure 2 (Enlarged)



Code link: Final_Submission_Files (https://livenorthumbriaac-my.sharepoint.com/:f:/g/personal/w23005405_northumbria_ac_uk/Esq4rAH2UtRAh4MuoVm8o48BJVFgwhKZmJ_S6iZ20smf2Q?e=XNb7au)