Real-Time Emotional Health Detection using Fine-Tuned Transfer Networks with Multimodal Fusion

Aditi Sharma[1], Kapil Sharma[2], Akshi Kumar[3]*

[1] Department of Computer Science & Engineering, Delhi Technological University, New Delhi, India
[2] Department of Information Technology, Delhi Technological University, New Delhi, India
[3] Department of Information technology, Netaji Subhas University of Technology, New Delhi, India
**akshi.kumar@nsut.ac.in*

Abstract
Recognizing and regulating human emotion or a wave of riding emotions is a vital life skill as it can play an important role in how a person thinks, behaves and acts. Accurate real-time emotion detection can revolutionize the human-computer interaction industry and have the potential to provide a proactive approach to mental healthcare. Several untapped sources of data, including social media data (psycholinguistic markers), multimodal data (audio & video signals) combined with the sensor-based psychophysiological and brain signals help to comprehend the affective states and emotional experiences. In this work, we propose a model that utilizes three modalities i.e., visual (facial expression and body gestures), audio (speech) and text (spoken content), to classify emotion into discrete categories based on Ekman's model with an additional category for 'neutral' state. Transfer learning has been used with multi-stage fine-tuning for each modality instead of training on a single dataset to make the model generalizable. The use of multiple modalities allow integration of heterogeneous data from different sources effectively. The results of the three modalities are combined at the decision-level using weighted fusion technique. The proposed EmoHD model compares favorably to the state-of-the-art technique on two benchmark datasets MELD and IEMOCAP.

Keywords: Emotion recognition, Transfer learning, Multimodal, Healthcare Detection

Declarations

Funding: No Funding has been received.

Conflicts of interest/Competing interests: The authors certify that there is no conflict of interest in the subject matter discussed in this manuscript.

Availability of data and material: Publicly accessible data has been used by the authors.

Code availability: Can be available on request.

Authors' contributions: All the authors have equally contributed in the manuscript preparation.

Ethics approval: The work conducted is not plagiarized. No one has been harmed in this work.

Consent to participate: All the authors have given consent to submit the manuscript.

Consent for publication: Authors provide their consent for the publication.

1. Introduction

Emotions are the product of changes in the affective system brought about by sensory information stimulation [1]. The emotional state of a person or a group regulates social interactions which in turn affects quality of life. These are an integral part of human behaviour and are generally perceived as positive and negative emotions. While positive emotions improve the quality of life and overall well-being, negative emotions can adversely affect the health and reasoning capabilities in humans. Psychological issues like stress, anxiety and depression are often a result of amassed negative emotions for a prolonged period. Thus, it is imperative to detect emotional state as for overall wellness, early diagnosis and intervention can make a difference in the lives of individuals with psychological disorders. Moreover, in various use-cases such as tele-health and customer service, it is extremely important that the emotional state of the subject or customer is identified correctly so that an appropriate response can be given by a person (healthcare professional or business owner) or an automated system [1]. Automated emotion recognition would be an assistive

tool in the former case and an integral one in the latter. The levitating evidence regarding the significance of emotions in human-to-human communication has provided the basis for researchers in the data science communities to develop automatic emotion evaluation methods with a goal to achieve intelligent human-to-computer interaction. Typically, in a human-human investigational setting, the complete diagnostic picture of an individual's emotional functioning can be analyzed using various evaluation parameters, such as, appearance, psychomotor behavior, characteristics of speech, affect and mood, thought content and concentration among others [2]. The prominent application areas of affective emotion recognition include human-robot interaction, e-learning, summarizing videos and healthcare [4].

Affect is an observable reaction of an individual towards an event, which causes the change in emotional state of the person [3]. It can be expressed as a tone of voice, a smile, a frown, a laugh, a smirk, a tear, pressed lips, a crinkled forehead, a scrunched nose, furrowed eyebrows, an eye gaze, and changes in heart rate or blood pressure. Affective state can help convey if things are going well or if there is something amiss or out of sorts based on reaction towards certain events. Face-to-face communication with psychologists is one of the most trusted way to identify disorders or instability in emotions as the practitioners assess the mental and emotional well-being of the person by observing the facial expressions, hand gestures or change in tone of voice. Studies on automatic recognition of emotional expression have been studied either by collecting data from self-reported assessment-based questionnaires [8] or by using machine-based data which includes various bio-signals [7]. The emergence of smartphone and wearable devices have shown promise towards the remote assessment of psychological well-being which includes both emotional and mental health. But capturing and assessing modes of expressions or observed manifestations using digital IoT-based biomarkers is intrinsically challenging owing to the volume & variety of data generated. Further, for identifying the emotions in patients with psychological disorders, relying only on IoT-based wearable biomarkers is not viable solution, as it might be difficult to convince the patients with anger issues to use wearables. Most of the pertinent research uses different modalities as inputs to label and estimate emotions. Affect recognition is accurate when it combines different observations from users and has information about their context and circumstances, but identifying emotional affect based on one modality will never wrap all the variations of human feelings when going through a change in the affective emotional state.

Affective emotional state can be judged through various modalities - facial expression and body gestures (visual), speech (audio), spoken content (text), and physiological signals from sensors attached to the body. Most of the work conducted in the field is analyzing the human emotion state using unimodal data such as bio-signals, or only facial expressions. Deep learning architectures have achieved state-of-the-art results in computer vision with Convolutional Neural Networks (CNN), and in natural language processing (NLP) tasks with Transformer-based models. They are especially useful for feature extraction, being able to learn representations that cannot be modelled manually. This ability of deep learning architectures also enables transfer learning. By learning feature representations on large datasets, they can be used end-to-end or for feature extraction on another smaller dataset in the same or different domain. The pre-trained models can also be fine-tuned on a comparatively smaller dataset to adapt them to a particular domain. Deep learning models are expensive to train, but the concept of transfer learning allows the model to be trained once and used any number of times for classification or fine-tuning, which are significantly cheaper tasks. We use this complementary relationship to improve performance with sustainability as the foundation.

In this research, we propose a model for emotional health detection (EmoHD), which makes use of heterogeneous data from three modalities – visual (facial expressions), audio (tone) and text (linguistic), for classification of emotional state of the subject into one of seven categories - neutral, happiness, anger, sadness, fear, disgust and surprise. The EmoHD model is designed to emulate a real-world scenario and would be suitable for integrating for tasks such as smart home assistants, automated customer service, online education, Bullying detection, CCTV monitoring and telehealth, healthcare among others. In real-time audio-visual input can be taken from video surveillance and separated into video and audio streams for individual processing. The video component is sampled to obtain image frames, from which the subject's face is extracted for facial expression recognition. A fine-tuned ResNet50 pre-trained architecture is used for emotion classification. The audio signal is used for two components - speech emotion classification and automatic speech recognition (ASR). For classifying emotion based on audio, the audio clips are segmented into 960ms clips and converted to log mel spectrograms. A VGGish architecture pre-trained for audio classification is fine-tuned for this task. For automatic speech recognition, we use an out-of-the-box pre-trained model from nVIDIA's NeMo toolkit - QuartzNet which is trained for general human speech recognition. Transcription is compared with the available subtitles in the dataset using term frequency and cosine similarity. We find that the ASR is fairly accurate, hence the transcription is directly used instead of the subtitles for text emotion classification, in accordance with our objective of emulating a real-world scenario. For text classification, we fine-tune DistilBERT, a general- purpose language representation model which is condensed from the original BERT model resulting in reduced size and faster inference time while retaining almost all of the capabilities. For combining results from the three modalities, we choose a late fusion strategy which is justified both empirically and logically.

The primary contributions of the work are:

- A multimodal approach for emotion classification based on audio-visual input, modelled such that it can be adopted in smart services.
- Application of transfer learning for every modality component to achieve generalizability and reduce computation costs.
- Multi-stage fine-tuning of pre-trained models in each modality component - using a unimodal emotion dataset followed by the target dataset.
- The performance is evaluated on a portion of MELD, giving an F1-score of 61.27. Generalizability is evaluated by testing on the benchmark IEMOCAP dataset, giving an F1-score of 65.88.

2. Related Work

Studies on automatic recognition of emotional expression have been studied either by collecting data from self-reported assessment-based questionnaires [5] or by using machine-based data which includes various bio signals [6]. The prominent application areas include human-robot interaction [7, 8], e-learning [9-11], summarizing videos [12, 13] and healthcare [14, 15]. Most of the pertinent research uses different modalities as inputs to label and estimate emotions. These include studies which report the use of text-based analytics [17], facial expression coding [18] and acoustic-feature coding [19]. Few studies also report bimodal models which use audio-visual features [20- 22] or audio-textual features for emotion recognition [23, 24].

A recent research trend in the affective computing community includes studies that recognize emotions in multimodal content, i.e., using textual clues with visual and audio modality [25, 26]. In 2011, Soleymani et al. [27] put forward a model for emotion recognition using EEG signals and eye gaze data in response to watching video clips. In 2015, Tzirakis et al. [28] proposed a system for emotion detection in RECOLA dataset using CNN to extract features from the speech and a ResNet for the visual modality. In 2016, Ranganathan et al. [29] presented an emoFBVP database of multimodal (face, body gesture, voice and physiological signals) recordings and used convolutional deep belief network (CDBN). In the same year, Poria et al. [30] suggested the use of 3 CNNs to first generate high level features for text, audio and visual modalities individually and then use multiple kernel learning (MKL) to combine data. The proposed CRMKL model was evaluated on the IEMOCAP dataset. In 2017, Nguyen et al. [31] introduced a novel approach using 3-dimensional convolutional neural networks to model the spatio-temporal information with deep-belief networks (DBNs) for multimodal emotion recognition in the eNTERFACE multimodal emotion database. Later in 2019, Poria et al. [32] developed a Multimodal Emotion Lines Dataset, which included textual dialogues and their corresponding visual and audio components. In 2019, Mittal et al. [33] proposed M3ER, a multimodal emotion recognition algorithm which uses a data-driven multiplicative fusion technique for three modalities (face, speech, and text) with deep neural networks. The authors experimented on two benchmark datasets, IEMOCAP and CMU-MOSEI and showed improvement over prior works. More recently, in 2020, Delbrouck et al. [34] used Modulated Attention Transformer with Modulated Normalization Transformer and modulated fusion to combine linguistic and acoustic inputs. The authors evaluate their performances on the IEMOCAP, MOSI, MOSEI and MELD dataset.

Uddin et al. [35] created a new physiological dataset by integrating BioVid Emo dataset and BioVid pain dataset. They applied the advanced dataset to recognize emotions and affective states in pain assessment. The work presented in [36] proposed a minixception +LSTM model on the BioVid Emo dataset to determine the emotions in videos. Authors in [37] utilized the BioVid Emo dataset in order to detect negative emotions in individuals. Only two classes of emotions of the dataset, amusement and sadness were evaluated by the authors as indication of positive and negative emotions category. From wearable devices, only electrocardiogram (ECG) signals were recorded and a variety of classifiers such as SVM, KNN, DT, RF and gradient boosting decision tree were implemented to analyze the positive and negative emotions. Xie et al. [38] accessed ECG, EMG and SCL and executed Wavelet transform features and SVM on BioVid Emo dataset to attain an accuracy of 94.81%. None of the studies considered all the signals (bio and video both) for detecting the emotion state of humans.

Few recent studies on multimodal emotion recognition have reported the use of transfer learning. In 2017, a pre-training and fine-tuning (PT/FT) approach for transfer learning with deep neural networks for emotion recognition was used by Gideon et al. [39]. Ouyang et al. [40] proposed a model audio-visual emotion recognition using deep transfer learning and multiple temporal models. The results were evaluated on the WILD2017 dataset. Kumar et al. in 2021, has used summarization to reduce the size of data using Fuzzy-C mean [42], Tavallali et al. in has used synthetic nearest neighbour model in 2021 [41]. Dresvyanskiy et al. [43] report the use of transfer learning and various fusion techniques for emotion recognition in data with both audio and video modalities. Siriwardhana et al. [44] propose a "BERT-like" pre-trained self-supervised learning architecture to represent both speech and text modalities for multimodal speech emotion recognition.

3. Background concepts

This section gives a brief overview on the key concepts used in this research study.

3.1. Emotion Models

Emotions are an integral part of human behaviour and can generally be perceived as positive and negative emotions. While positive emotions improve the quality of life and overall well-being, negative emotions may adversely affect the health and reasoning capabilities in humans. Negative emotions are also an influential factor in causing mental health issues such as stress, anxiety and depression. Several untapped sources of data, including social media data, multimodal data (audio-video) combined with the sensor-based psychophysiological signals help to comprehend the affective states and emotional experiences.

Multiple emotional models have been presented by various researchers, broadly these models are categorized into two types: discrete emotion model and dimensional emotion model. Discrete models categorize the emotions into accurate emotions that a person experiences, such as joy, anger, frustration, sadness, whereas in dimensional models, emotion of a person is categorized into arousal, valence etc. Dimensional models are based on a hypothesis that the emotions are not independent, therefore, to exhibit their relationship they are placed into spatial space. Figure 1 highlights different Emotion Models.

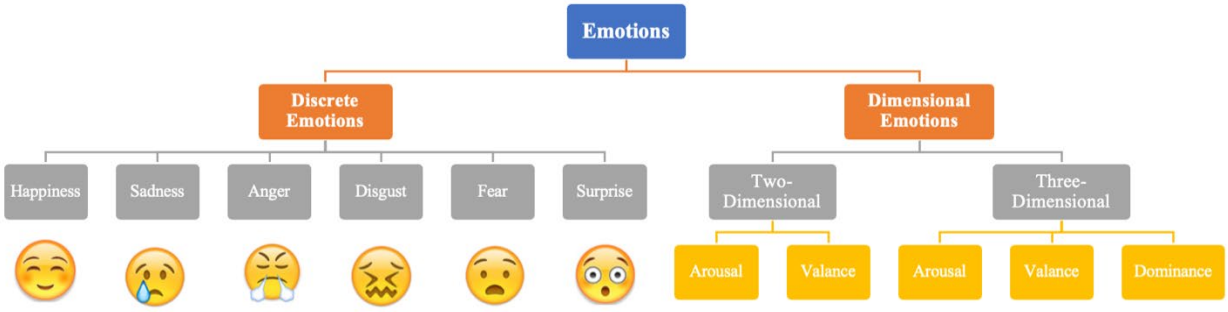

Fig. 1. Emotion Models

In the proposed work, we have used Ekman's model which is a discrete emotion-based approach having 6 different emotions: happiness, sadness, anger, disgust, surprise, and fear [16]. The models suggest that these fundamental six emotions originate from different part of the human neural network, activated from a reaction to an external stimulus. For our experimentation, we have included one extra emotional state as neutral, for the time when the subject is not experiencing any of the above six emotions.

3.2. Transfer Learning

The essence of transfer learning is extracting knowledge from one or more source tasks and applying it to the target task. In traditional machine learning models, each task has a different learning process, and the learned model is also heavily dependent on the size and quality of the dataset particular to that task. This methodology cannot be applied in situations where the labelled data is insufficient, such as text classification tasks for low-resource languages.

A general representation of the transfer learning model has been shown in figure 2, where the model is first pre-trained on a large unlabeled dataset and is then adapted to a supervised target task using the labelled data available [45]. The pre-trained model can also be made available for use to others who can fine-tune it for their target downstream tasks. Thus, while pre-training on a large dataset is computationally expensive, it only needs to be done once. Compared to it, downstream fine-tuning is much cheaper. The process of transfer learning in all three modalities are discussed next.

3.2.1. Image and Video

Transfer learning was first popularized in the field of computer vision, particularly for image classification. It can be attributed to the combination of two factors: the release of ImageNet [46], a large-scale dataset of more than 14 million images in over 20000 categories, and the advent of deep learning in computer vision which was made possible by a dataset of such scale. Since 2012, deep learning models started dominating ILSVRC, improving the results every year and eventually beating even the human benchmark. Convolutional architectures turned out to be most effective for

computer vision, the winners being AlexNet in 2012, ZFNet in 2013, VGG and GoogLeNet in 2014, and ResNet in 2015. Videos can be treated as a stack of images linked by temporal features. As with ImageNet, the release of large datasets for video clips such as Sport-1M, Youtube-8M and HowTo100M, along with their associated pre-trained models and extracted features have enabled researchers to apply them to downstream video understanding tasks by fine-tuning on smaller datasets.

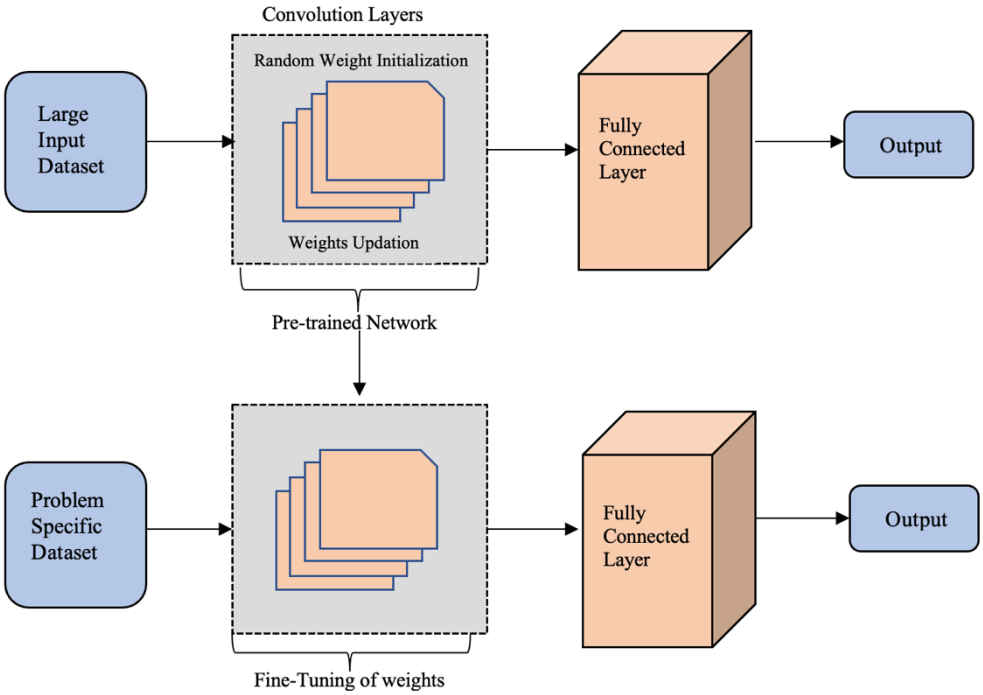

Fig. 2. Architecture of Transfer learning model

3.2.2 Audio

The representation of audio signals as waveforms inextricably links them to transfer learning methods applied to images i.e., CNN-based architectures. The development of the landmark audio dataset - AudioSet and its accompanying work on large scale audio classification resulted in pre-trained models such as VGGish and YAMNet, which can be applied to specific audio classification tasks. VGGish is a variant of the VGG model modified for log mel spectrogram audio inputs, it can be used either as a feature extractor or fine-tuned as a part of a larger model. SoundNet is a model which is able to recognize objects and scenes from sound, by training on millions of unlabeled videos.

3.2.3 Natural language

The early approaches for language processing were focused on representing words in a dense vector space, called the word embeddings. Traditional word embeddings such as Word2vec, GLoVe and fastText learn a single representation for a word, irrespective of context. The concept of context was added to embeddings in ELMo, making it possible for a word to have multiple representations depending on its use in a sentence.

Building upon the foundation of word embeddings, and with the advancement in Transformer network, language model pre-training has achieved landmark achievements in the field of NLP. BERT (Bidirectional Encoder Representations from Transformers) is pre-trained on two unsupervised prediction tasks - Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). For MLM, some of the tokens from the input text are masked randomly and the objective is to predict the original word of the masked word based on its context. Unlike statistical language models which are oriented left-to-right, MLM allows a fusion of left and right context. The NSP task captures the relationship between consecutive sentences. The training data consists of sentence-pairs, with half of them having correct pairings while the other half do not. Following the success of the training scheme and architecture of BERT, several models have adapted from its strengths and continue to break the state-of-the-art for NLP tasks using more data and compute power, such as RoBERTa, GPT-2, XLNet, T5 and GPT-3 among others.

4. Datasets

Studies on automatic recognition of emotional expression have been studied either by collecting data from self-reported assessment-based questionnaires or by using IOT-based data which includes various bio signals. Most of the pertinent research is conducted on different datasets such as ISEAR, DailyDialog, MELD, SMILE, BioVidEmo, And IEMOCAP datasets. For our study we have used two datasets, Multimodal Emotion Lines Dataset (MELD), and Interactive Emotional Motion Capture (IEMOCAP) which use audio-visual and textual features for emotion recognition.

Most of the work conducted in the field is analyzing the human emotion state using bio-signals, or only facial expressions, we have used multimodal data having 3 modalities identifying the changes in visual signals (facial expressions) along with audio signals (tone) and the text analytics (linguistic). We have used VGGish model to pre-train the model, it contains VGGFace2 for face identification and extraction; its audio component for the sound classification. For textual component, we use ASR on both the datasets to extract the transcripts, this textual component is then pre-trained on DistilBERT. These pre-trained networks are fine-tunned using MELD and IEMOCAP for each modality separately except for Visual modality, where the fine-tunning is performed twice, first on CIFE dataset, and then finally on the selected datasets. The proposed model is validated on two datasets MELD and IEMOCAP.

4.1. MELD

To fine-tune and evaluate our model, we use a portion of the MELD dataset, which is an extension of the EmotionLines dataset [49]. The latter only has text modality while the former includes video and audio to accompany the text. MELD contains 1433 dialogues taken from the 'Friends' TV series, where each dialogue encompasses one or more participants (actors) and emotions. An utterance is a part of the dialogue where only one participant is speaking and has one corresponding emotion. There are a total of 13708 utterances in MELD, with an average duration of 3.59 seconds. The utterances are categorized under our desired emotion states.

For our experiments, we select a subset of MELD (henceforth referred to as 'MELD-sub') by leaving out the utterances where multiple participants are speaking, or multiple participants are facing the camera while one of them is speaking. The selected utterances have either a single participant, or only the speaker having their face completely visible on the camera.

4.2 IEMOCAP

Interactive emotional dyadic motion capture database was collected by SAIL laboratory at University of South California. Busso et. al. [50], they hired 10 actors to participate in a study. Total of 5 sessions were conducted, each having 2 actors having markers on their face, head, and hands to record their facial expressions as well as their hand movements. The study was conducted for a period of 12 hours. The sessions contain both scripted and spontaneous communication between the two actors. The 12-hour recording has been manually segmented into 10039 utterances, belonging to 9 different emotions- happiness, anger, surprise, sadness, fear, excitement, frustration, neutral, and others. For our study we have merged the frustration into anger and utterances of others has been dropped, converting the data with 7 emotions only. Table 1 highlights the emotion-wise distribution of the datasets.

Table 1. Emotion-wise data distribution

Emotion	**CIFE**	**MELD**	**IEMOCAP**
Anger	1785	1607	1103
Disgust	266	361	471
Fear	761	358	589
Happiness	3636	2308	648
Neutral	644	6436	1708
Sadness	2485	1002	1084
Surprise	997	1636	988

4.3. CIFE

Candid images for facial expression (CIFE) is a dataset created by Li et al. [56] to construct an improved facial expression model for analyzing real time facial expression tasks. The CIFE dataset is produced through social media

and the Web. Web crawling methods are employed to obtain natural expressions in the seven chosen categories of emotions. These categories are happy, anger, disgust, sad, surprise, fear and neutral. Utilizing related phrases of these expressions, a huge amount of pictures are accumulated corresponding to the seven classes of expressions. There were 14756 pictures in total for these seven expressions in which pictures of anger, disgust, fear, happiness, neutral, sadness and surprise, and some pictures were added manually to the dataset corresponding to the classes where data was unbalanced. Viola face detector was availed to uncover images of faces with these seven expressions [56]. CIFE dataset is a freely available public dataset.

5. The Proposed Multimodal Emotional Health Detection Model-EmoHD

To identify the affective human emotion under real-time video surveillance, we have proposed a model using three modalities. The video surveillance received as input is separated into two streams of information - visual and audio. This makes it robust to data quality in either modality, in situations where the subject is visible in the camera, but the audio is unclear or vice-versa. The content of the subject's utterance is also evaluated to judge their emotional state. To enable real-time evaluation, we use the audio to create a transcription using ASR instead of directly using the text associated with it in the dataset. If the ASR is not accurate, it would affect the text analysis downstream. Therefore, we track its accuracy by comparing it with the dataset text for each utterance.

A weighted fusion of the three i.e., visual, audio and text components, gives the emotional state of the subject. We select a decision-level fusion after considering that keeping the three components parallel until decision ensures that they are unaffected by the quality of data in another modality. Figure 3 represents the proposed model.

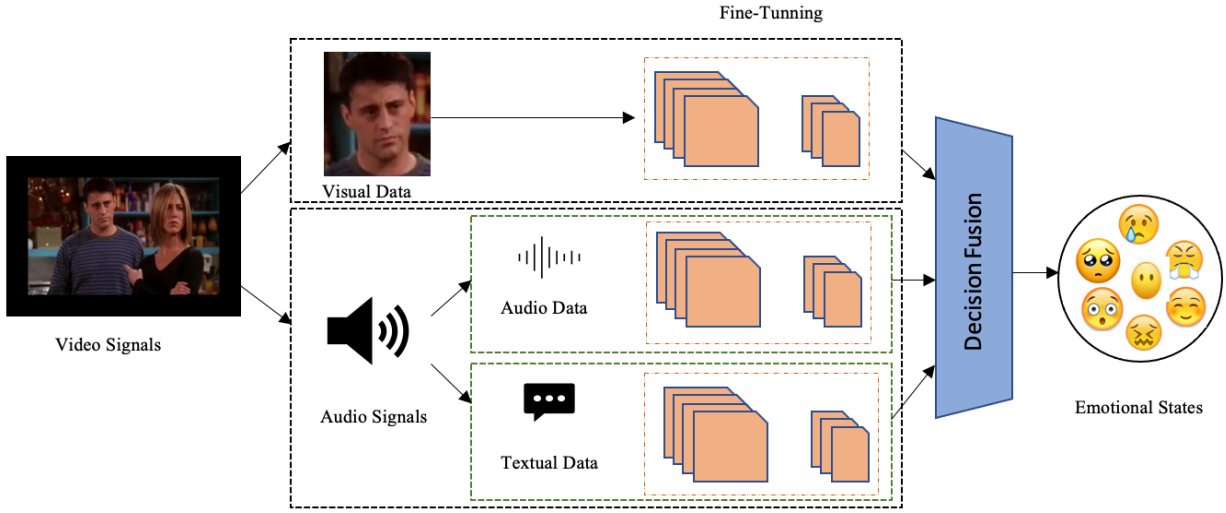

Fig. 3. Proposed Model

5.1. Visual Modality

A video can be considered to be a spatio-temporally connected stack of images. Since our goal is real-time evaluation and our data also consists of short utterances, we choose to focus on facial expressions in sampled images from the video clips. We do not consider temporal features which would be effective in evaluating longer videos.

5.1.1. Preprocessing

The video clips from MELD-sub and IEMOCAP are sampled at 1 frame per second. Since the utterances are short (3-4 seconds on average), we avoid picking a peak frame for analysis. Instead, we adopt a majority voting strategy after individual emotion classification of the frames tagged to a video clip. As the sampled frame depicts a scene, but for emotion analysis, only facial expressions are required, multi-task cascaded convolution network (MTCNN) is used to extract faces. Frames with no faces detected are discarded. The extracted face images are then used to fine-tune the pre-trained model. The sampled frames from MELD-sub before and after preprocessing are shown in figure 4.

Fig. 4. Face extraction using MTCNN

5.1.2. Fine-tuning

We start with a ResNet50 model loaded with weights pre-trained on the VGGFace2 dataset as our base model. VGGFace2 consists of face images with significant variations in pose, illumination, ethnicity and age of the subject. This acts as a suitable precursor to the data we will use downstream for emotion classification. ResNet50 is a convolution-based architecture which achieved state-of-the-art results on ImageNet and several other image classification tasks.

ResNet, the winner of ImageNet Large Scale Visual Recognition Challenge, helps to reduce the error rate even on increasing the number of layers [47]. Basic architecture of residual network follows the convolution architecture, containing 1-D convolution layer along with the max pooling layer and ReLU activation layer. The major distinction between two are the skip connections, i.e. in basic deep learning architectures, consecutive hidden layers are connected to each other, but in ResNet50, the ReLU function on every alternative layer is performed after taking the output of i^{th} layer along with the output of $(i+2)^{th}$ layer, and is provided as an input to $(i+3)^{th}$ layer after performing the activation function, known as skip connections or residual connections. ResNet50 preserves the knowledge gained during training and can speed up the new model with more hidden layers. This ResNet50 model has been implemented in Keras for emotion detection from face recognition by pre-training the model on VGGFace2 dataset for face identification and then using CIFE to map the faces with different emotions, and at last this pre-trained model detects the affective emotion state of an individual from MELD-sub and IEMOCAP's visual signals.

CNN-architectures learn general features in the lower layers and more task-specific features in the higher layers [48]. We will utilize this property to fine-tune the base model by unfreezing some layers from the top and retraining the network on the CIFE dataset and converge the architecture towards our target task of emotion classification. The CIFE dataset also contains images with varying illumination, age and ethnicity, making it suitable to act as a bridge between VGGFace2 and both the datasets. The fully connected output layer of the base model, softmax classifier for 8631 categories is removed. The top 10 layers (average pooling layer and 9 convolutional layers) of the model are unfrozen and a fully connected layer with 5 nodes is added at the top for our target emotion categories for the second round of pre-training on CIFE, to minimize categorical cross-entropy loss using the Adam optimizer.

5.1.3 Feature extraction and classification

After fine-tuning on CIFE, we freeze all the layers of the model and remove the top layer. The MELD-sub and IEMOCAP images are passed through the fine-tuned model to obtain vectors representing the extracted features. We use these feature vectors as input to the multi layer perceptron (MLP) network for emotion classification. The configuration of the MLP network is shown in figure 5.

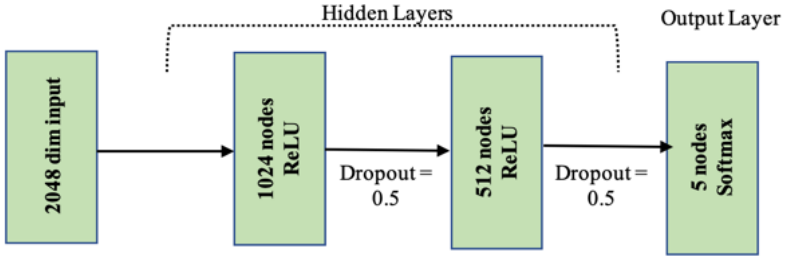

Fig. 5. MLP Network

5.2. Audio Modality

The field of audio classification has seen significant work using handcrafted features as well as deep learning methods. RNNs and LSTMs are effective in modelling the temporal nature of audio signals. CNNs are also capable of achieving

state-of-the-art results through the representation of audio signals as spectrograms. To avoid training to fit a dataset or a particular domain, we adopt a transfer learning approach. We extract the audio from video clips using the *MoviePy* library in *wav* format and *librosa* library is used for audio preprocessing.

5.2.1 Preprocessing

The audio clips are divided into 1 second frames to match the image sampling. They are resampled to 16 kHz and converted to a spectrogram using short-time Fourier transform. The window size and window hop are set to 25ms and 10ms respectively. A mel spectrogram is then computed by mapping the spectrogram into 64 bins within a range of 125 to 7500 Hz. Finally, the log mel spectrogram is computed with an offset of 0.01.

5.2.2 Fine-tuning and classification

We utilize the VGGish model for the audio component. VGGish is pre-trained on a large YouTube dataset and provides a 128-dimension embedding. Since the model is trained for general audio categorization, we fine-tune it for the emotion classification task. The preprocessing is done in accordance with the audio features required as inputs to VGGish. The pre-trained weights are frozen, except for the fully connected layers, which are trainable. This Pre-trained model is then fine-tunned with MELD-sub and IEMOCAP datasets by taking 65% of both the datasets with seven classes. The final layer is modified to have 7 nodes for each emotion category. Additionally, the activation function for the final layer is changed from sigmoid to softmax since each audio sample has only one associated emotion label. The log mel-spectrogram features are passed in as input. Since the weights of convolutional layers are frozen, the top 3 fully-connected layers converge the model towards emotion classification.

5.3. Text Modality

We analyze the spoken content by converting extracted audio clips to text using ASR. This analysis is important because there may be cases where the person is not expressive enough or speaks in monotone. While these are important factors in determining the emotional state of the subject, they would not be reliable in such cases. Another factor to consider is that facial features of a person might be similar for different emotions, for instance, furrowed eyebrows to show anger as well as surprise. In such a case audio features and spoken content will be the determining factors.

5.3.1. Audio transcription

NVIDIA's NeMo toolkit is used for transcription. The model architecture used for ASR is QuartzNet which is comparatively smaller than other existing models. In particular, the model we use is QuartzNet15x5NR-En which has been pre-trained on the LibriSpeech corpus and fine-tuned with Room Impulse Responses (RIR) and noise augmentation to make it robust to noise. The pre-trained model is used out-of-the-box since it has state-of-the-art performance on general speech. The audio transcription is done for both the dataset. We extract the output in two forms - lowercase alphabetic tokens for similarity check, and complete sentences formed by applying the DistilBERT-punctuation scheme for our input to the masked language model.

5.3.2 Similarity check

For Real-time applications, where the video-surveillance will be provided as input, and the transcripts will not be available, the ASR will provide the adequate help in detecting the emotions of an individual accurately. To gauge the performance of our ASR module, we are interested only in word-level similarity and not the context or meaning. Ideally, we want the transcription to match the subtitles word to word. For MELD-sub we perform the similarity check with the EmotionLines dataset [49], and for IEMOCAP, its textual component which is provided along with the dataset. We also used Term Frequency (TF) and Cosine Similarity (CS) for checking their alignment. The ASR module has shown the average accuracy of 93.07% for both the datasets.

5.3.3 Fine-tuning and classification

Transformer-based architectures and BERT-inspired training schemes dominate the state-of-the-art in NLP tasks. With each jump in accuracy, the models keep getting larger with the number (in millions) of parameters being - Google's BERT-base (110) and BERT-large (340), Facebook's RoBERTa (355), OpenAI's GPT-2 (1500) and nVIDIA's MegatronLM (8300) among others. However, with increasing size the models keep becoming less suitable for small devices such as smartphones, which are the backbone of smart services. We utilize DistilBERT, a model condensed from BERT-base using a student - teacher training method. It is reported to have 60% faster inference time at 97% of the performance of BERT-base while having only 66 million parameters, which is a 40% reduction in size.

We use HuggingFace's transformers library to fine-tune a pre-trained DistilBERT model, specifically Distilbert-base-uncased for complete sentence generation. These transcriptions generated from the ASR module are tokenized and padded to align with the maximum-length utterance. The fine-tunned model on Distilbert-base-uncased is then again fine-tunned, this time for mapping emotions to the sentences, it is fine-tunned on 65% of MELD-sub and IEMOCAP datasets separately. A 'CLS' token (Special classification token) is added at the beginning for the classification task. The output of the final transformer layer for the CLS token will be used as the input features that will be fed into an MLP classifier.

5.4 Weighted Decision Fusion

In general, multimodal data fusion can be done in three ways, early fusion, late fusion and joint fusion [51]. To enable real-time evaluation, we select a decision-level fusion after evaluating the three components parallel ensuring that they are unaffected by the quality of data in another modality. This makes it robust to data quality in either modality, in situations where the subject is visible in the camera, but the audio is unclear or vice-versa. The content of the subject's utterance is also evaluated to judge their emotional state. To process the results of the 3 parallel models (Visual, Audio, and Text Models) the weighted decision fusion technique has been used. Every model has been fitted for varied modality, each having results dependent on subjects not just the cause. To identify the impact of each modality the model has been trained separately for the three signals. A thorough review of impact of different fusion studies has been shown by [52]. In the proposed model, we have used the weighted late fusion framework provided by Tsanousa et al. [53]. The framework assigns weights on the basis of detection ratio rather than F-scores. Detection Ratio (DR) is shown in equation 1.

$$DR= TP/(TP+TN+FP+FN) \quad (1)$$

where, TP represent true positive, TN represents True Negative, FP represents False Positive, and FN is False Negative. DT is calculated for each class, as the model has been executed for 7 Discrete Emotions, the number of classes are 7. The weight, W of each output class is calculated as:

$$W= 1-DR \quad (2)$$

The weight of each class is then multiplied with the probability vector, P belonging to each model to find the predictive score of the class.

$$S=W*P \quad (3)$$

After calculating the weight of each class for individual model, the score of the model is calculated by adding the scores of each class. The final decision is opted trough the maximum function, i.e., the model having highest predictive score for the test case is chosen as the output level. As the proposed architecture contains 3 models for 3 modalities, the final output class is provided as:

$$\text{Output Class}= \text{Max } (S_{Visual}, S_{Audio}, S_{Text}) \quad (4)$$

This late weighted fusion strategy helps to choose the output class from the model best suitable for the output class.

6. Implementation and Results

The EmoHD model is evaluated on two separate datasets- MELD and IEMOCAP. The three individual models are first fine-tuned using ResNet, VGGish and BERT for visual, audio and text modalities respectively. Now these fine-tuned models are trained using 75% of the two datasets in two separate execution. Rest 25% of the data is used for evaluating the performance of the model. In the EmoHD model, the three modalities are evaluated separately, and at last the decision-level fusion is used to detect the emotional state of a person. To understand the impact of each modality, we also evaluate the performance of models separately as well as combining two modalities as well. The various models are compared using the F1-score. The pseudo-code of the proposed EmoHD model is given next:

Pseudo-Code: **EmoHD Model**

Input: MELD-sub and IEMOCAP
1. Create Transcript from Audio Signals using ASR.
2. Pre- Train three models for 3 modalities (Simultaneously):
    ResNet50 for Visual
    VGGish for Audio
    DistilBERT for Text
3. Fine-Tune the Visual component on CIFE dataset.
4. Fine-Tune Pre-trained text on Distilbert-base-uncased for complete sentence generation
5. Fine-Tune Models on 65% of the respective Dataset (simultaneously).
6. Do Decision Fusion on outputs of Fine-Tune Models

7. Output: Emotion
8. Evaluate the Model

Initially the model is tested for each modality. In the first model (Visual), after fetching the facial expressions using MTCNN, the facial expressions of each frame are provided as an input to fine-tuned transferred network using ResNet50. The model provided an F1-score of 54.93 for IEMOCAP and 51.96 for MELD dataset. Both the datasets were also evaluated for the audio model (VGGish) and the textual model (BERT) and report an F1-score of 53.39 and 54.19 for IEMOCAP and 50.85 and 51.48 for the MELD dataset respectively. As video signals are further broken down into visual and audio signals, so same input is required for both, therefore the model is evaluated for these modalities, providing a better F1-score for both the datasets as compared to when tested on a single modality. Similarly as textual data is fetched form audio signals using ASR, the two modalities are also combined to understand the impact of two modalities rather than one. But since textual data cannot be directly generated from the video signals, the model is directly evaluated on the three modalities to detect the final class using weighted late fusion. The performance of each implementation is shown in Table 2 and 3 for the IEMOCAP and the MELD datasets respectively. The performance of the models is shown for each emotion as well. As observed from the table 2 and 3 the model performs better for the 'happiness' emotion followed by 'sadness' emotion. This variation in accuracy of detection of each emotion has happened because of variations in the training datasets as well as certain emotions are expressed differently by different individuals. To have an effective real-time emotion detection, we need to fine tune the models on even larger datasets.

Table 2. F1-Score for varied emotions for IEMOCAP

Modalities / Emotions	**Visual**	**Audio**	**Text**	**Visual +Audio**	**Audio + Text**	**Visual + Audio + Text**
Anger	55.18	52.1	54.62	58.91	58.14	65.07
Disgust	52.61	51.66	47.3	54.03	54.1	58.41
Fear	57.06	53.08	51.62	59.73	56.32	64.81
Happiness	58.18	55.34	61.48	64.05	61.28	71.23
Neutral	54.29	56.65	52.53	59.24	52.1	65.51
Sadness	52.4	49.58	60.02	56.11	62.7	69.78
Surprise	54.83	55.34	51.8	60.72	53.03	66.41
Total	**54.93**	**53.39**	**54.19**	**58.97**	**56.81**	**65.88**

Table 3. F1- Score for MELD dataset

Modalities / Emotions	**Visual**	**Audio**	**Text**	**Visual +Audio**	**Audio + Text**	**Visual + Audio + Text**
Anger	53.12	49.23	52.41	56.18	53.25	59.95
Disgust	49.14	47.62	43.11	48.91	49.21	54.82
Fear	57.81	50.21	48.93	55.69	54.31	59.31
Happiness	52.91	52.47	57.29	57.83	56.21	65.36
Neutral	51.23	53.78	49.71	55.78	49.78	61.46
Sadness	48.19	46.89	57.21	52.34	54.78	66.71
Surprise	51.36	55.78	51.72	56.53	56.45	61.29
Total	**51.96**	**50.85**	**51.48**	**54.75**	**53.42**	**61.27**

The proposed EmoHD model which combines three modalities, namely visual, audio and textual, surpasses the results with late weighted fusion in comparison to separate modalities or subset of modalities. But at the same time, it can also be observed from the figure 6 and 7, that although the model and execution environment is same, but the performance varies widely for two datasets. The variation in the performance is impacted by a lot of factors, the most promptly, each individual expresses the emotions differently, the actors for both the datasets were different. Along with this, the datasets available for fine-tunning models were of a limited size, the achieved accuracy has been because of the use of pre-trained models and fine-tunning the model twice for visual modality, as can be observed from table 2 and 3, the visual modality has performed better as compared to audio and text modality individually, because the visual component was earlier fine-tunned on CIFE and then on MELD-sub and IEMOCAP.

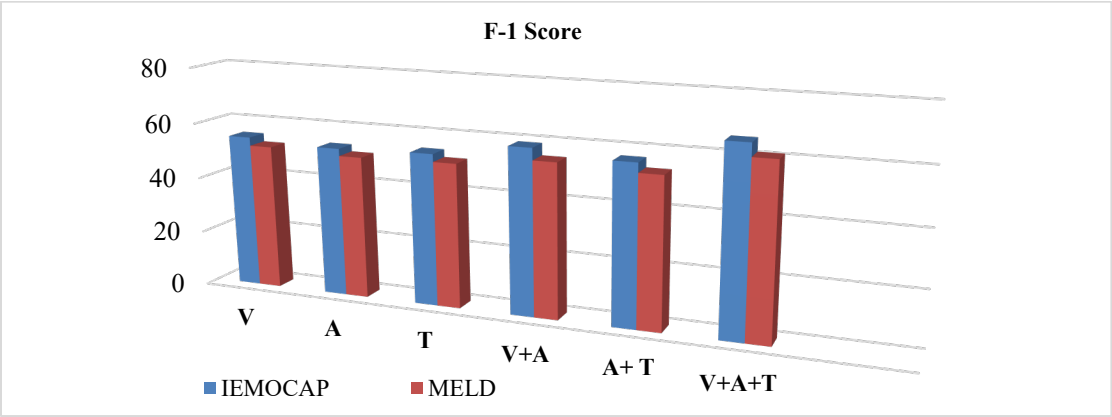

Fig. 6. Performance of various models on MELD and IEMOCAP

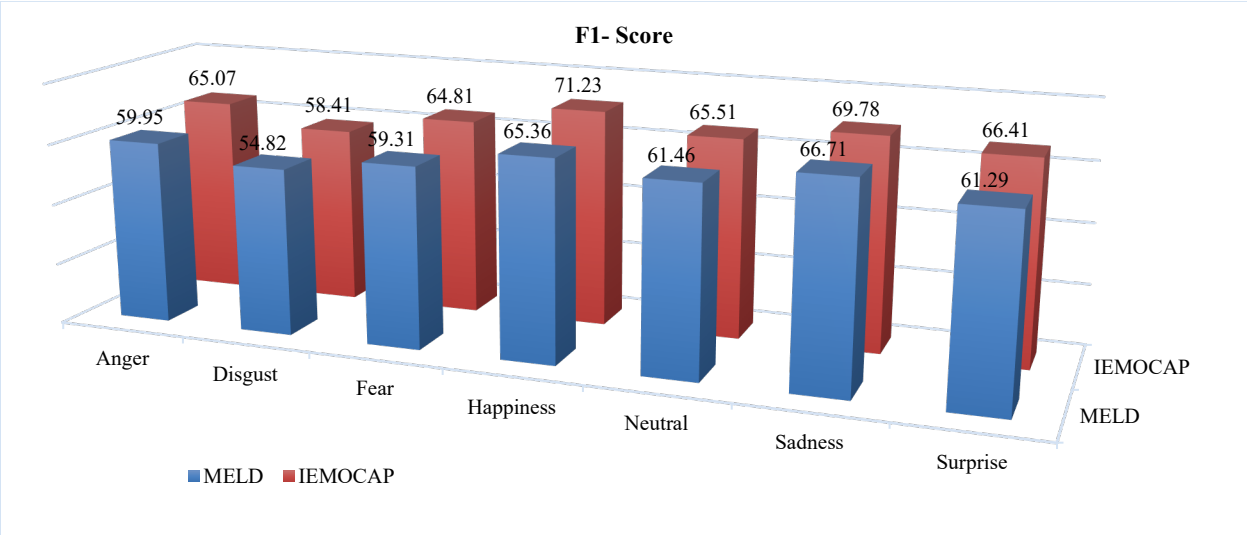

Fig. 7. Proposed Model for Discrete Emotions

The proposed deep transfer learning weighted late-fusion model has identified the discrete emotion more precisely in comparison to the existing models. The performance of the proposed EmoHD model is compared with existing state-of-the-art [54, 55]. Table 4, shows the comparison of EmoHD model for both the datasets, with the results provided by Acheampong et al. for MELD [54] and Hazarika et al. for IEMOCAP [55]. As observable from Table 4, authors in [54] and [55] have performed empirical analysis on MELD and IEMOCAP datasets respectively. The deep convolution model shows better results on MELD, whereas modified recurrent neural network performs better on IEMOCAP. The proposed model has performed subpar than SOTA, as it takes individual modalities into account, and have used the existing data to pre-train the models which eliminates the impact of small dataset on the accuracy. Although the variation in results on the two datasets over same model indicates the accuracy of the model will depend upon the data (subjects). For accurate detection, a pre-trained model should be fine-tunned per user specific for accurate estimation of emotional state of an individual.

Table 4. Comparison of Models

Models	IEMOCAP	MELD
CNN	48.1	55.02
c-LSTM	54.9	56.44
c-LSTM + Att	56.1	-
DialogueRNN	59.8	57.03
TL-ERC	58.85	-
ConGCN	-	59.40
Proposed Model	**65.88**	**61.27**

As indicated from results, the weighted late fusion model can act as a promising model for merging the three modalities of visual, audio and text for accessing the affective emotional state of an individual in real-time using video surveillance.

7. Conclusion

A perceived situation generates affective responses in the form of emotion reactions and bodily state consequently prompting behavioural actions. Electrophysiological signals, behavioural signals (e.g., posture), speech signals, social interactions and psycholinguistic features in social data can be used as observable traits for detecting human emotional states. This work used the capabilities of deep learning, transfer learning and late data fusion to detect emotions in multimodal dataset. A multi-stage fine-tuning of pre-trained model was used for visual, audio and textual signals individually. Decision fusion was subsequently used to give decision on the identified emotion categories of happiness, surprise, neutral, fear, anger, sadness and disgust on two benchmark database MELD and IEMOCAP. The model has performed better than the state of the art. The proposed model can be used for real-time video surveillance of patients, that can automate the detection of their unstable emotions, and can help the doctors in identifying the cause of the spike in their emotions and can decide their treatment plan accordingly.

References

1. Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, *23*(10), 1175-1191.
2. Zhang, S., Zhang, S., Huang, T., Gao, W., & Tian, Q. (2017). Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(10), 3030-3043.
3. Gunes, H., & Pantic, M. (2010). Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, *1*(1), 68-99.
4. Kumar, A., Sharma, K., & Sharma, A. (2021). Hierarchical deep neural network for mental stress state detection using IoT based biomarkers. *Pattern Recognition Letters*, *145*, 81-87.
5. García-Magariño, I., Chittaro, L., & Plaza, I. (2018). Bodily sensation maps: exploring a new direction for detecting emotions from user self-reported data. *International Journal of Human-Computer Studies*, *113*, 32-47.
6. Zhang, L., Walter, S., Ma, X., Werner, P., Al-Hamadi, A., Traue, H. C., & Gruss, S. (2016, December). "BioVid Emo DB": A multimodal database for emotion analyses validated by subjective ratings. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1-6). IEEE.
7. Rabiei, M., & Gasparetto, A. (2014, October). A system for feature classification of emotions based on speech analysis; applications to human-robot interaction. In *2014 Second RSI/ISM International Conference on Robotics and Mechatronics (ICRoM)* (pp. 795-800). IEEE.
8. Szabóová, M., Sarnovský, M., Maslej Krešňáková, V., & Machová, K. (2020). Emotion Analysis in Human–Robot Interaction. *Electronics*, *9*(11), 1761.
9. Bahreini, K., Nadolski, R., & Westera, W. (2016). Towards multimodal emotion recognition in e-learning environments. *Interactive Learning Environments*, *24*(3), 590-605.
10. Ashwin, T. S., Jose, J., Raghu, G., & Reddy, G. R. M. (2015, December). An e-learning system with multifacial emotion recognition using supervised machine learning. In *2015 IEEE seventh international conference on technology for education (T4E)* (pp. 23-26). IEEE.
11. Ayvaz, U., Gürüler, H., & Devrim, M. O. (2017). Use of facial emotion recognition in e-learning systems. *Інформаційні технології і засоби навчання*, (60, вип. 4), 95-104.
12. Zeng, H., Shu, X., Wang, Y., Wang, Y., Zhang, L., Pong, T. C., & Qu, H. (2020). EmotionCues: Emotion-Oriented Visual Summarization of Classroom Videos. *IEEE Transactions on Visualization and Computer Graphics*.
13. Tu, G., Fu, Y., Li, B., Gao, J., Jiang, Y. G., & Xue, X. (2019). A Multi-Task Neural Approach for Emotion Attribution, Classification, and Summarization. *IEEE Transactions on Multimedia*, *22*(1), 148-159.
14. Hossain, M. S., & Muhammad, G. (2017). Emotion-aware connected healthcare big data towards 5G. *IEEE Internet of Things Journal*, *5*(4), 2399-2406.
15. Weitz, K., Hassan, T., Schmid, U., & Garbas, J. (2018, November). Towards explaining deep learning networks to distinguish facial expressions of pain and emotions. In *Forum Bildverarbeitung* (pp. 197-208).
16. Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, *98*(45-60), 16.
17. Saravia, E., Liu, H. C. T., Huang, Y. H., Wu, J., & Chen, Y. S. (2018). Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3687-3697).
18. P. Ekman and W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Stanford University, Palo Alto, 1977.
19. D. Datcu and L. Rothkrantz, "Semantic audio-visual data fusion for automatic emotion recognition," Euromedia'2008, 2008.
20. L. C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in Information, Communications and Signal Processing, 1997. ICICS., Proceedings of 1997 International Conference on, vol. 1. IEEE, 1997, pp. 397–401

21. D. Datcu and L. J. Rothkrantz, “Emotion recognition using bimodal data fusion,” in Proceedings of the 12th International Conference on Computer Systems and Technologies. ACM, 2011, pp. 122–128.
22. B. Schuller, “Recognizing affect from linguistic information in 3d continuous space,” Affective Computing, IEEE Transactions on, vol. 2, no. 4, pp. 192–205, 2011.
23. A. Metallinou, S. Lee, and S. Narayanan, “Audio-visual emotion recognition using gaussian mixture models for face and voice,” in Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on. IEEE, 2008, pp. 250–257
24. F. Eyben, M. Wollmer, A. Graves, B. Schuller, E. Douglas-Cowie, and ¨ R. Cowie, “On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues,” Journal on Multimodal User Interfaces, vol. 3, no. 1-2, pp. 7–19, 2010.
25. V. Rosas, R. Mihalcea, and L.-P. Morency, “Multimodal sentiment analysis of spanish online videos,” IEEE Intelligent Systems, vol. 28, no. 3, pp. 0038–45, 2013P. Ekman and W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Stanford University, Palo Alto, 1977.
26. V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, “Speech language & multimedia technol., raytheon bbn technol., cambridge, ma, usa,” in Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific. IEEE, 2012, pp. 1–4.
27. Soleymani, M., Pantic, M., & Pun, T. (2011). Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing*, *3*(2), 211-223.
28. Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, *11*(8), 1301-1309.
29. Ranganathan, H., Chakraborty, S., & Panchanathan, S. (2016, March). Multimodal emotion recognition using deep learning architectures. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1-9). IEEE.
30. Poria, S., Chaturvedi, I., Cambria, E., & Hussain, A. (2016, December). Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)* (pp. 439-448). IEEE.
31. Nguyen, D., Nguyen, K., Sridharan, S., Ghasemi, A., Dean, D., & Fookes, C. (2017, March). Deep spatio-temporal features for multimodal emotion recognition. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1215-1223). IEEE.
32. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.
33. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues. In *AAAI* (pp. 1359-1367).
34. Delbrouck, J. B., Tits, N., & Dupont, S. (2020). Modulated Fusion using Transformer for Linguistic-Acoustic Emotion Recognition. *arXiv preprint arXiv:2010.02057*.
35. Hagar, A. F., Abbas, H. M., & Khalil, M. I. (2019, December). Emotion Recognition In Videos For Low-Memory Systems Using Deep-Learning. In *2019 14th International Conference on Computer Engineering and Systems (ICCES)* (pp. 16-21). IEEE.
36. Hagar, A. F., Abbas, H. M., & Khalil, M. I. (2019, December). Emotion Recognition In Videos For Low-Memory Systems Using Deep-Learning. In *2019 14th International Conference on Computer Engineering and Systems (ICCES)* (pp. 16-21). IEEE.
37. Iskhakova, A., Wolf, D., & Meshcheryakov, R. (2020, October). Automated Destructive Behavior State Detection on the 1D CNN-Based Voice Analysis. In *International Conference on Speech and Computer* (pp. 184-193). Springer, Cham.
38. Xie, J., Xu, X., & Shu, L. (2018, May). WT feature based emotion recognition from multi-channel physiological signals with decision fusion. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)* (pp. 1-6). IEEE.
39. Gideon, J., Khorram, S., Aldeneh, Z., Dimitriadis, D., & Provost, E. M. (2017). Progressive neural networks for transfer learning in emotion recognition. *arXiv preprint arXiv:1706.03256*.
40. Ouyang, X., Kawaai, S., Goh, E. G. H., Shen, S., Ding, W., Ming, H., & Huang, D. Y. (2017, November). Audio-visual emotion recognition using deep transfer learning and multiple temporal models. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (pp. 577-582).
41. P Tavallali, et al., An EM-Based Optimization of Synthetic Reduced Nearest Neighbor Model towards Multiple Modalities Representation with Human Interpretability, Multimedia Tools and Applications, 2021.
42. Kumar, A., Sharma, K., & Sharma, A. (2021). Genetically optimized Fuzzy C-means data clustering of IoMT-based biomarkers for fast affective state recognition in intelligent edge analytics. *Applied Soft Computing*, 107525.
43. Dresvyanskiy, D., Ryumina, E., Kaya, H., Markitantov, M., Karpov, A., & Minker, W. (2020). An Audio-Video Deep and Transfer Learning Framework for Multimodal Emotion Recognition in the wild. *arXiv preprint arXiv:2010.03692*.
44. Siriwardhana, S., Reis, A., Weerasekera, R., & Nanayakkara, S. (2020). Jointly Fine-Tuning" BERT-like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition. *arXiv preprint arXiv:2008.06682*.
45. Abbas, A., Abdelsamea, M. M., & Gaber, M. M. (2020). Detrac: Transfer learning of class decomposed medical images in convolutional neural networks. *IEEE Access*, *8*, 74901-74913.
46. Huh, M., Agrawal, P., & Efros, A. A. (2016). What makes ImageNet good for transfer learning?. *arXiv preprint arXiv:1608.08614*.
47. Wu, Z., Shen, C., & Van Den Hengel, A. (2019). Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, *90*, 119-133.

48. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98-125.
49. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.
50. Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*(4), 335-359.
51. Kumar, A., Sharma, A., & Arora, A. (2019, March). Anxious depression prediction in real-time social data. In International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India.
52. Hossain, M. S., & Muhammad, G. (2019). Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, *49*, 69-78.
53. Li, W., Tsangouri, C., Abtahi, F., & Zhu, Z. (2018). A recursive framework for expression recognition: from web images to deep models to game dataset. *Machine Vision and Applications*, *29*(3), 489-502.
54. Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review*, 1-41.
55. Hazarika, D., Poria, S., Zimmermann, R., & Mihalcea, R. (2021). Conversational transfer learning for emotion recognition. *Information Fusion*, *65*, 1-12.
56. Li, W., Abtahi, F., & Zhu, Z. (2015, November). A deep feature based multi-kernel learning approach for video emotion recognition. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 483-490).