Review on Multimodal Fusion Techniques for Human Emotion Recognition

Ruhina Karani[1]
Research Scholar, School of Computer Engineering and Technology, MITWPU, Pune, India

Dr. Sharmishta Desai[2]
Associate Professor, School of Computer Engineering and Technology, MITWPU, Pune, India

***Abstract*—Emotions play an essential role in human life for planning and decision making. Emotion identification and recognition is a widely explored field in the area of artificial intelligence and affective computing as a means of empathizing with humans and thereby improving human machine interaction. Though audio visual cues are vital for recognizing human emotions, they are sometimes insufficient in identifying emotions of people who are good at hiding emotions or people suffering from Alexithymia. Considering other dimensions like Electroencephalogram (EEG) or text, along with audio visual cues can aid in improving the results in such situations. Taking advantage of the complementarity of multiple modalities normally helps capture emotions more accurately compared to single modality. However, to achieve precise and accurate results, correct fusion of these multimodal signals is solicited. This study provides a detailed review of different multimodal fusion techniques that can be used for emotion recognition. This paper proposes in-depth study of feature-level fusion, decision-level fusion and hybrid fusion techniques for identifying human emotions based on multimodal inputs and compare the results. The study concentrates on three different modalities i.e., facial images, audio and text for experimentation; at least one of which differs in temporal characteristics. The result suggests that hybrid fusion works best in combining multiple modalities which differ in time synchronicity.**

***Keywords*—Feature-level fusion; decision-level fusion; hybrid fusion; artificial intelligence; EEG**

I. Introduction

Recently, emotion recognition has been explored extensively in the areas of artificial intelligence, affective computing and human computer interaction. Emotion is a psychological events generated by a person's tendency toward need, which can be broadly classified into Physiological arousal and subjective experience [1, 2]. Physiological arousal refers to the physiological responses of the human body, which can be measured by electrical signals like electrocardiogram (ECG) and electroencephalogram (EEG), whereas subjective experience is a phenomenon, which relates to individual's feelings about different emotional states. It is expressed through facial expression, audio, gestures, etc. [1, 3, 4].

Audio visual features are generally treated as vital cues for emotion recognition, but they are prone to deception, if preformed deliberately. Physiological signals, on the other hand, cannot be deceived easily as they are based on internal physiological responses. Combining physiological responses with subjective experiences leads to more accurate emotion recognition.

Although emotion recognition can be achieved through a uni-modal approach, it may not work well for certain conditions of subjective experience where the people are excellent at hiding their emotions or the input data is very noisy. Also cross culture approaches for expressiveness act as barriers to identifying emotions correctly. Multimodal approach, on the contrary, can achieve significant accuracy by combining inputs from multiple modalities [5]. However, to get precise and accurate results, correct fusion of these multimodal signals is required.

A major concern for fusing multiple modalities is deciding the level at which multimodal fusion should occur and how to achieve the fusion [6, 7]. Since different modalities differ significantly in temporal characteristics, synchronization among them plays an important role in multimodal fusion [7]. Multimodal fusion techniques are broadly classified into decision-level fusion, feature-level fusion and hybrid fusion. Decision-level fusion and feature-level fusion are the most regularly used techniques for multimodal fusion in emotion recognition.

The existing literature on review of fusing multiple modalities is either based on signals which are synchronous in time or same type of signals (e.g. Fusion of different 2D images). B. Huang et al. proposed a review of medical image fusion techniques for Computed Tomography Scan (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET) and Single-photon Emission Computed Tomography (SPECT) images which describes spatial domain fusion and transform domain fusion [8]. Chen Xiao Yu et al. published a trend of machine learning based on fusion which focused on Ensemble Learning, Transfer Learning and Federated Learning for technology and data fusion [9].

This paper reviews different multimodal fusion techniques for decision level, feature level and hybrid fusion which can be used for human emotion recognition. The study also provides comparative analysis of these techniques. The proposed work focuses on the modalities which are synchronous as well as asynchronous in time to achieve emotion recognition. For experimentation, inputs from multiple modalities like facial images, audio and text are considered.

The paper is organized as follows: Section II describes decision-level fusion and the various techniques used to

achieve it for emotion recognition. Section III is devoted to discussing feature-level fusion and the various techniques used for achieving feature-level fusion. In Section IV, Hybrid approach for multimodal fusion is discussed in detail. Section V provides experimental analysis of different fusion methods. Section VI focuses on the discussion on analyzing and comparing different fusion methods. Section VII concludes the paper.

II. Decision-Level Multimodal Fusion

Since emotion recognition works with multiple modalities, which differ from each other in various aspects, at times it becomes challenging to extract features from different and coupled modalities. Also, changes in the time synchrony of different modalities affect the dimensions which are used for emotion recognition [10]. To deal with this issue, decision-level multimodal fusion takes each modality independently for classification.

Decision-Level fusion technique first classifies the data of each modality individually and the result is combined at a later stage to achieve fusion. This fusion technique is also called late fusion as combining results occurs at a very late stage after classification. Fig. 1 shows a general framework of decision-level multimodal fusion considering audio, visual, text and EEG input for emotion recognition.

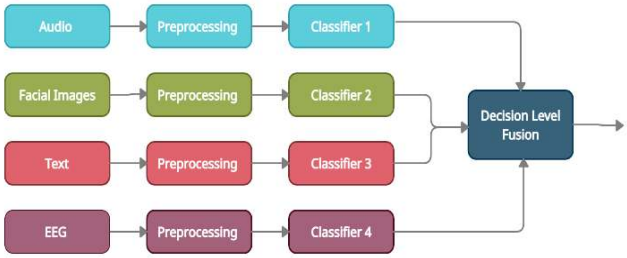

Fig. 1. General Framework of Decision-Level Fusion for Emotion Recognition.

There are various techniques of decision-level fusion that have been experimented with by researchers. This section presents each of these techniques.

A. Support Vector Regression

Support vector regression (SVR) works on the principles of the support vector machine (SVM). The ideology behind SVR is to find the best fit line suitable for the problem at hand. The best fit line in SVR is the hyperplane that contains the maximum number of points. It uses a margin of tolerance in approximation to the SVM. SVR maps to high dimensionality to estimate a function while offering nonlinear complexity [11].

In many studies, researchers have used SVR with the RBF kernel for decision-level fusion of emotion recognition [10, 12]. Haiyang Su et al. used bidirectional long short-term memory for classifying emotions individually from audio, facial images and text along with RBF kernel SVR to achieve significant improvement in classification accuracy [10]. Mihalis A. Nicolaou et al. compared the performance of bidirectional LSTM-NN and SVR on audio, facial expression and shoulder features for predicting spontaneous effect [13]. SVR can compensate for redundant information, which makes SVR suitable for decision-level fusion required for emotion recognition but it does not perform well when the number of feature extracted are more than training samples.

B. Blending Algorithms

Blending algorithm is a technique of ensemble machine learning that uses a metaclassifier to combine the outputs of different machine learning models.

A blending algorithm works in two layers. The first layer uses a traditional approach for training, where multiple basic classifiers are trained on the inputs from different modalities. The second layer is used for combining the outputs of the basic classifiers used in level 1. The outputs of each classifier are combined to form a new training set, which acts as an input to a higher level classifier called metaclassifier. The metaclassifier can use the ensemble learning techniques of bagging, boosting and stacking for combining the results based on the weight, bias and variances of base classifiers [14]. The role of level 1 basic classifiers is to classify the multimodal data, whereas the level 2 classifier is responsible for combining the outputs of base classifiers. Blending algorithms are found to have a better performance in the prediction.

Man Hao et al. used Convolution Neural network and support vector machine as base classifiers for speech and facial images along with the blending algorithm of metaclassifier to achieve multimodal emotion recognition and achieved 81.36% accuracy [14]. Lee et al. used ensemble learning on multimodal acoustic, lexical and discourse features using three classifiers to find negative emotions in spoken dialog [15]. Blending algorithm can be used as a decision-level fusion technique for emotion recognition owing to its improved performance for multimodal inputs. Blending algorithm is stable and less noisy.

C. Brain Emotion Learning Models

The brain emotional learning (BEL) system takes inspiration from the biological amygdala-orbitofrontal model to emulate the high speed of the emotional learning mechanism used by the human brain [16]. The BEL model comprises four main components. Amygdala, Orbitofrontal cortex (OFC), the Thalamus and Sensory cortex. After receiving input signals in the form of emotional stimuli from the Sensory cortex & Thalamus and a reward signals from the external environment, amygdala interacts with the OFC. The OFC evaluates the response of the amygdala based on the input received from the Sensory cortex, which leads to the prevention of improper learning connections. In a nutshell, the Sensory cortex integrates the features extracted from different unimodal inputs and the amygdala and OFC works to form a decision after interacting with the memory [17, 18].

BEL models require rewards extracted from input data and are derived from monotonic reinforcement learning. BEL-based networks are efficient in predicting peak points [19]. This method can be used as a stacking ensemble method in multimodal decision fusion techniques [17]. Zeinab Farhoudi et al. used BEL as a decision-level fusion technique for multimodal audio visual features to achieve a significant

accuracy of 73.9% for emotion recognition [17]. BEL model has good decision-making capability due to its cognitive-based structure. Due to this, Brain Emotional Learning can be used at decision-level fusion for multimodal emotion recognition.

D. Decision Tree based Approach

A decision tree is a non-parametric supervised learning method whose internal nodes represent test and leaf nodes represent classes. Filtering is applied though the nodes to achieve the required output. A decision tree tries learning decision rules from the given input data features.

Yucel Cimtay et al. used a decision tree for fusing multimodal inputs from facial expressions, galvanic skin response (GSR) and electroencephalogram (EEG) to detect emotions where the authors emphasized more on the probability vector received from facial images-based classifier. The condition of the decision tree was set to check whether this probability vector goes below a certain threshold. The research proposes using other modalities only if the specified conditions hold true [20].

Heysem Kaya et al. used a random forest decision tree based model on audio, facial and scenic inputs extracted from videos CVs to achieve decision fusion to estimate a suitable interview variable [21].

Decision trees are suitable for decision-level fusion because of their capability of learning rules according to the requirement. The predictions made by each classifier used for unimodal inputs can be combined based on the specific conditions provided at each node by using decision trees. But, decision trees are unstable data structures and they are prone to inaccuracy.

E. Rule based Approach

Rule-based classifiers are types of classifiers that make the class decision using various "if-else" rules. These classifiers are used to generate a descriptive model as the rules are easily interpretable. The "if" condition is called the antecedent and the predicted class is called the consequent in rule-based classification.

Subhasmita Sahoo and A. Routray has used a rule-based approach for emotion recognition using multimodal audio and video inputs. The rules were set to give high priority to the result of the facial image classifier and low priority to the audio classifier. The results of the audio classifier were acceptable only if there was confusion in the results of the image based classifier [22].

Rule-based classification is a way to achieve decision-level fusion as they have the advantage of fuse the multimodal signal classifier's output as per the conditions specified in the antecedent.

F. Dempster-Shafer Theory

The Dempster- Shafer (D-S) theory of evidence originated by Dempster [23], and formalized mathematically by Shafer [24]. D-S theory was developed as reasoning and modeling framework with epistemological uncertainty. This framework can be used to integrate multiple sources of evidence and to agree to a combined degree of belief for different predictions. Compared to the Bayesian model, the D-S theory is an extensive approach to address uncertainty and imprecision [25]. Due to its ability to efficiently handle the uncertainty and inconsistency of multimodal data, D-S theory is widely used in data fusion, fault detection and pattern recognition [26].

The D-S theory is drawn on the concepts of allocating suitable beliefs and probabilities to hypotheses, applying the D-S rule for fusing independent inputs from different sources and arriving at the final decision of the optimal hypothesis in a workable and reasonable manner [25]. Decision-Level fusion make use D-S evidence theory to combine the outputs of each classifier used for multimodal inputs.

Xiao-Dan Zhang used D-S theory for fusing multi-modal text and image inputs after applying KNN and SVM classifiers individually on unimodal input to improve the classification results [27]. Nazmuzzaman Khan et al. claimed that D-S theory has limitations in terms of handling conflicting data. Therefore, they used an improved version of D-S theory using a distance function and evidence-weighted penalty for combining results of different sensors to achieve improved object detection [26]. S. Nemati used D-S theory for decision-level fusion in her approach of hybrid fusion for emotion recognition. The author had applied D-S theory on audio, video and textual multimodal inputs to achieve improved emotion recognition [28]. Yu-Ting Liu et al. used Weighted Fuzzy Dempster–Shafer Framework that can adjust weights of evidence, which is inconsistent for integrating the outputs of EEG and eye movement's classifiers [25].

G. Decision Template Algorithm

The decision template (DT) algorithm is a simple and efficient fusion algorithm, which uses the average of decision profiles of each classifier used for training multimodal data inputs. These averages are treated as decision templates for each class. Assessing the resemblance between the decision profile and different decision templates helps provide accurate class prediction. Though the algorithm is used widely for decision fusion, it has limitations as it uses average decision profiles and does not emphasize differences in the performance of classifiers [29]. Reza Ebrahimpour et al. used a decision template algorithm for decision-level fusion of different classifiers for handwritten digit recognition [30]. Due to the limitations of the decision template algorithm, Aizhong Mi et al. proposed an improved weighted decision template algorithm that calculates the classifier's performance using a statistical vector and allocates appropriate weights to each classifier according to the reliability of their outputs. The authors applied this improved approach to the cataract recognition system and received improved results compared to the decision template algorithm [29].

III. Feature-level Multimodal Fusion

Feature-level fusion is a technique, which combines the features from different layers or branches. When applied to multimodal inputs, the feature-level fusion method combines the features extracted from each modality individually using different feature fusion techniques before giving it to a classifier for classification. It is considered a ubiquitous part

of several modern network architectures. Generally, feature-level fusion is implemented using the concatenation operation, but that might not be the best choice for achieving good results.

Fig. 2 depicts a general framework for feature-level multimodal fusion considering four different modalities like facial images, audio, text and EEG for emotion recognition.

Feature-level fusion works with the idea of extracting the most discriminating features from the extracted features and removing redundant information [31, 32]. It works with the conjecture of strict time synchronicity between different modalities and performs distinctively in cases where the modalities differ considerably in temporal characteristics [31, 33]. Fusion at the feature level works best for closely connected and synchronized modalities [31].

If feature-level fusion has to be used for multimodal approaches where individual modalities differ in time synchrony, the relationship between the different feature spaces needs to be explored and the features need to be made compatible [31].

Various techniques for feature-level fusion have been explored by the researchers. This section presents those techniques in detail.

A. Eigen Matrix for Fusion

The German word Eigen means characteristics. The terms Eigenvalues and Eigenvectors in computation deal with determining the characteristics of a matrix. Eigenvalues and eigenvectors are used in various applications, which include Principal Component Analysis, Spectral Clustering, and detection techniques in Computer Vision, etc. Some researchers found its use in feature-level fusion of multimodal inputs, where the individual features are extracted from each modality and serialized to form an Eigen matrix. This Eigen matrix is then normalized to get the fusion Eigen matrix [30]. The formula for Eigen matrix (E) calculation is specified in equation (1).

$$E_J^{(i)} = \frac{\left(e_j^{(i)} - \mu_j\right)}{\max e_j^{(i)} - \min e_j^{(i)}} \quad (1)$$

Where $e_j^{(i)}$ represents the j^{th} eigenvalue of i samples, μ_j represents the mean value of the j^{th} feature, and the denominator represents the range of j features.

Jian Che et al. used an Eigen matrix for multimodal feature fusion of longitudinal tear detection of a conveyor belt where the input modalities were images and sound [34]. The authors claimed to have improved accuracy due to multimodal fusion. However, the Eigen matrix technique for feature-level fusion relies on the mean and eigenvalues of each modality to achieve fusion. Therefore, this technique may not be able to use the entire range of diversity for each modality.

B. Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a method for analyzing the relationship between two multivariate sets of vectors [28, 35]. CCA is considered as an effective tool in analysis that has an ability to increase statistical power compared to univariate methods due to the use of second-order statistics [28].

CCA has an application in multimodal fusion [35]. It worked to find sets of changed variables having the maximum correlation between the two modalities [28]. CCA, when used for multimodal inputs, take linear combinations of $X1W1$ and $X1W1$that maximizes the pairwise correlations of two multimodal datasets where X1 & X2 are two multimodal datasets and W1 & W2 are canonical coefficient vectors [28].

Shahla Nemati used CCA for feature-level fusion to detect multimodal emotion recognition using audio and visual modalities [28]. The features of individual modalities are first extracted using different feature extraction methods before giving it to CCA for feature-level fusion.

CCA as a method for feature-level fusion uses the concept of maximizing the pairwise correlation between the two modalities. Therefore, applying CCA on more than two modalities will require fusion at multiple levels.

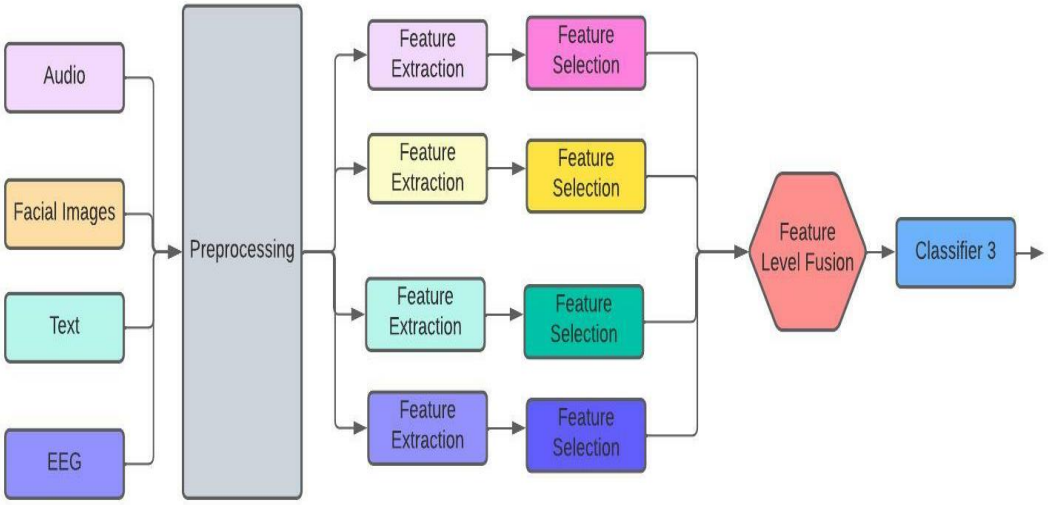

Fig. 2. General Framework for Feature-Level Fusion for Emotion Recognition.

C. Mixture of Brain Emotional Learning Model

The mixture of brain emotional learning models (MoBEL) combines the features of Mixture of Experts neural network (MoE) and Brain Emotional Learning (BEL) model. The BEL model has a good capability of decision-making due to its cognitive-based structure and the MoE model is based on working of the associative cortex of the brain, which has the capacity to integrate information from multiple sources [17, 18]. The presence of the associative cortex improves the brain's capacity of perception of the environment [36]. Therefore, MoE has the capacity to perform better in pattern recognition.

The MoBEL model structure uses the BEL model for expert and gating networks and trains all parts of the network jointly using back-propagation. MoBEL can be used as a fusion network to integrate multimodal features. The MoBEL model is more efficient in terms of processing speed, memory consumption, and neuron numbers than the MOE [17].

Zeinab Farhoudi et al. used MoBEL as a feature-level fusion technique for multimodal audio visual features to achieve a significant accuracy of 81.7% for emotion recognition [17].

D. Merging Features at Hidden Layer

This strategy works with the approach of inserting features of different time durations into different hidden layers of a training network [35]. This technique, though unexplored much, can tackle the problem arising due to asynchronous multimodal features. It also helps resolve the time compatibility issues of different multimodal features.

Given a network having multiple hidden layers and multimodal inputs with different time durations, short - duration features like audio are given as input to the first hidden layer. The output of the first hidden layer is combined with longer time features like visual features and given as input to the second hidden layer. The output of the second layer is then combined with longer duration features from the remaining modalities and given as the input to the third hidden layer. The process is repeated until the features are provided to the network [35].

Shizhe Chen et al. considered this approach for feature-level fusion in emotion recognition using three modalities: audio, images and EEG, wherein the authors used RNN-LSTM network for recognizing and fusing the features [35].

Though this approach can tackle the problems of asynchronous multimodal features, it depends upon the network used for training. Also, for modalities having similar duration, deciding the hidden layer at which these features need to be fused becomes an issue.

E. Concatenation

Concatenation generally uses consolidated dimensions to achieve fusion [37]. The concatenation formula is shown in equation 2.

$$Y = x_1 \cup x_2 \cup x_3 \ldots\ldots\ldots \cup x_k \quad (2)$$

where x_k is a set of output feature maps for the k^{th} layer and Y represents the fusion of all features.

The main goal of concatenation is to enrich the diversity of features for better recognition ability [36]. Sanghyun lee et al. used a concatenation method in a process of multimodal feature fusion for emotion recognition based on audio, visual and textual features for emotion recognition.

Schoneveld et al. used a concatenation method for multimodal feature-level fusion for emotion recognition based on audio visual modalities [38]. The authors also claimed to have improved fusion accuracy by adding hadamard product to concatenation in attention-based approach for emotion recognition [39].

It is the most popular method for feature-level fusion. But it might not be the best choice for multimodal inputs that differ in dimensions.

IV. Hybrid Multimodal Fusion

Feature-level fusion methods require synchronization between different modalities, but it is possible to produce results even when the data from one of the modality is missing. Decision-level fusion, on the other hand, does not require synchrony between different modalities, but its capabilities cannot be exploited fully even if the data from a single modality is missing.

The hybrid fusion method tries achieving the advantages of both feature level and decision-level fusion methods by combining these approaches. The hybrid fusion approach applies feature-level fusion on some synchronous modalities and combines the result of with the remaining asynchronous modalities using decision-level fusion.

The hybrid approach works well for multimodal inputs, where some inputs assume time synchrony and some others are asynchronous in time. Fig. 3 shows the general framework of hybrid multimodal fusion for emotion recognition fusion considering four different modalities like facial images, audio, text and EEG.

Shahla Nemati used a hybrid approach for multimodal emotion recognition based on three modalities viz. audio, visual, and users' comments. The author used feature-level fusion for audio and visual features since they were synchronized. The result is then fed to a classifier for classification. A separate classifier was used for user comments because of its asynchronous nature. In the next step, decision-level fusion was used to fuse the outputs of both classifiers [28].

Zhen-zhong Lan et al. used a hybrid model for multimedia event detection based on visual, audio and textual data. The authors used five classifiers for training based on five features and feature combinations, where feature fusion was performed to combine features and the result of each classifier was combined using decision-level fusion [40].

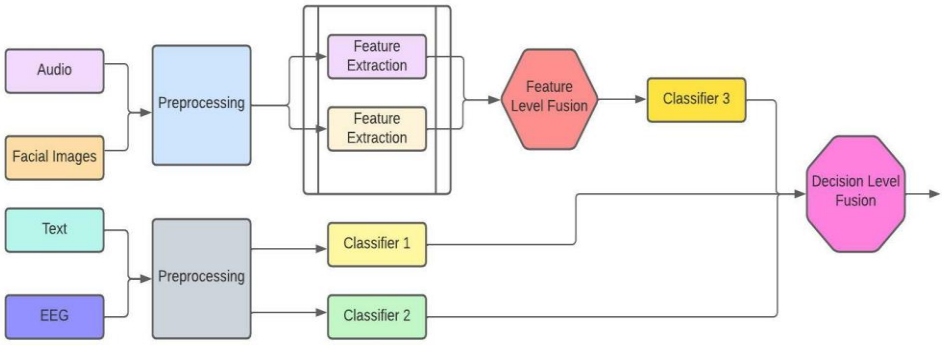

Fig. 3. General Framework for Hybrid Fusion to Emotion Recognition.

Fig. 4 depicts graphical representation of summary of different fusion methods described in review of literature.

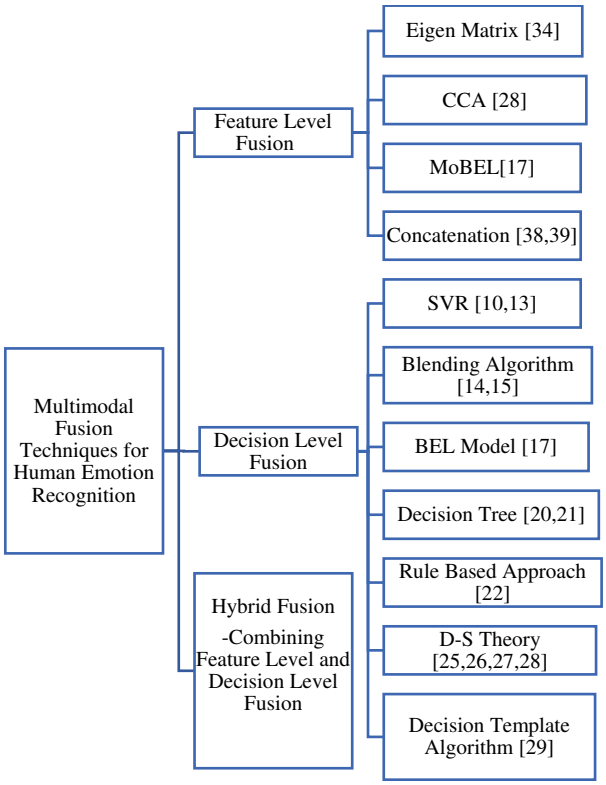

Fig. 4. Summary Diagram of Review of Literature Describing Various Fusion Methods.

V. Experimental Analysis

To analyze the effectiveness of decision level, feature level and hybrid fusion methods in emotion recognition, we focused on facial expression, audio and text as multimodal inputs. We have used RAVDESS dataset [41] for audio data, FER 2013 dataset curated by Pierre Luc Carrier and Aaron Courville [42] for facial images and EMOTION dataset [43] for textual data for the purpose of experimentation. The RAVDESS dataset contains 1440 sample recordings from 24 actors (12 female, 12 male), in a neutral North American accent showing 8 emotions namely calm, happy, sad, angry, fearful, surprise, and disgust [41].

The FER-2013 dataset curated by Pierre Luc Carrier and Aaron Courville contains 28,709 training examples of 48x48 pixel grayscale facial images depicting seven emotions namely angry, Disgust, Fear, Happy, Sad, Surprise and Neutral [42]. The EMOTION dataset contains Twitter messages in English having six basic emotions: anger, fear, joy, love, sadness, and surprise [43].

In order to analyze, we clustered these emotions into three categories i.e. positive, negative and neutral. Since the modalities used for experimentation were not following time synchrony, feature level fusion of text with other two modalities was impractical. For decision level fusion, we have used pre-trained models of CNN by FER and Wav2Vec.

2.0 for classifying facial expressions and audio each providing the accuracy of 59 % and 82.23% respectively and pre-trained BERT model for classifying textual data providing accuracy of 93.8%. The output of each classifier is fused using max voting ensemble algorithm. The decision level fusion is tested on annotated samples of CMU-MOSI dataset which provided accuracy of 59.09 %.

For hybrid fusion, feature level fusion is applied on features extracted from facial images and audio by concatenating them before applying random forest for classification. Feature level fusion provided accuracy of 86.6%. Textual data is classified using Pre-trained BERT model providing accuracy of 77%. The output of feature level fusion is combined with text classifier using max voting ensemble algorithm of decision level fusion. The feature level fusion is tested on annotated samples of CMU-MOSI dataset which provided accuracy of 66.67 %.

Table I describes performance analysis of individual classifier for facial images, audio and text each, feature level fusion of facial images and audio and decision level and hybrid fusion in terms of accuracy.

Fig. 5(a) shows graphical representation for the accuracy comparison of decision level and hybrid model. The decision level fusion is providing the accuracy of 59.09 % whereas feature level fusion gives the accuracy of 66.67% on annotated samples of CMU-MOSI dataset. Fig. 5(b) depicts the individual accuracy graph of facial images, text and audio respectively which provide the accuracy of 59%, 93.8% and 82.23% using CNN and BERT models. The last bar in the figure shows accuracy of 86.6% for feature level fusion of facial images and audio input.

TABLE I. PERFORMANCE ANALYSIS OF DECISION LEVEL AND HYBRID FUSION

Accuracy	Facial Images	Audio	Text
Individual Classification	0.59 (Using CNN)	0.8223 (Using CNN)	0.938 (Using BERT)
Feature Level Fusion	0.866 (Using concatenation & Random Forest)		
Decision Level Fusion	0.5909 (Using blending algorithm)		
Hybrid Fusion	0.6667 (Using concatenation & blending algorithm)		

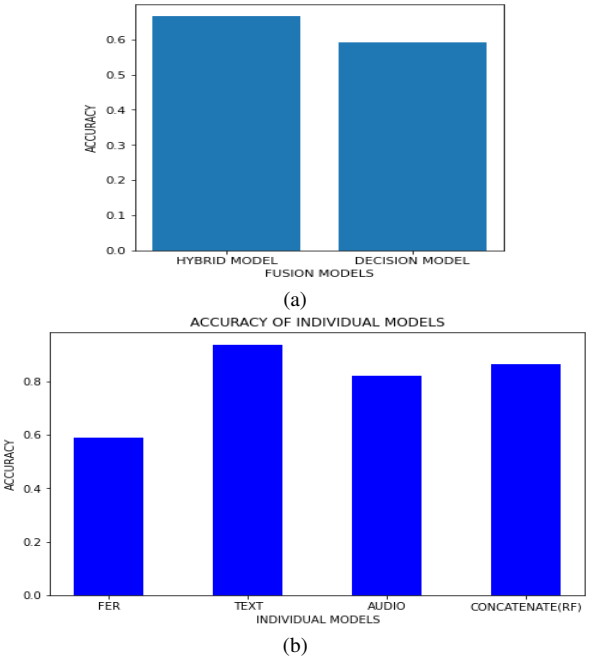

Fig. 5. (a) Accuracy Comparison Bar Graph of Hybrid and Decision Level Fusion, (b) Accuracy Graph of Individual Models and Feature Level Fusion Model of Audio and Facial Images.

VI. DISCUSSIONS

Emotions can be classified based on the discrete theory of emotions, where the emotions are categorized into discrete classes like anger, disgust, fear, joy, sadness and surprise, or based on dimensional emotional models where emotions are classified as a two valence-arousal models or three-dimensional valence-arousal-dominance models. Facial expression - based emotions are generally classified using discrete theory whereas EEG-based emotions use dimensional models for classification. Table II summarizes different multimodal fusion approaches adopted by researchers for emotion recognition.

Although a uni-modal approach can identify emotions to a certain extent, the accuracy of emotion recognition increases with the multimodal approach. While using a multimodal approach for emotion recognition, the main hurdle lies in fusing multiple modalities that differ significantly in temporal characteristics. The feature-based fusion approach is suitable for the inputs, which are synchronous in time.

Hence, this approach may not work well for emotion recognition based on multiple modalities that differ in temporal characteristics. The Eigen matrix technique for feature-level fusion relies on the mean and eigenvalues of each modality to achieve fusion. Therefore, this technique may not be able to use the entire range of diversity for each modality. Moreover, the Eigen matrix needs to be a square matrix, which further puts restrictions on the choice of features to be considered for fusion. CCA as a method for feature-level fusion uses the concept of maximizing the pairwise correlation between the two modalities. But, a traditional CCA can handle only two modalities at a time. Therefore, applying CCA on more than two modalities will require fusion at multiple levels. MoBEL model, when used for feature-level fusion, requires less processing speed, memory consumption, and neuron numbers. Since the MoBEL model is based on brain emotional learning, to exploit it fully, expertise on cognitive-based abilities of the brain is needed. Merging the features at the hidden layer works with the concept of merging the features at different hidden layers according to the duration of features to achieve feature-level fusion. Though this approach can tackle the problems of asynchronous multimodal features, it depends upon the network used for training. Also, for modalities having similar duration, deciding the hidden layer at which these features need to be fused becomes an issue. Concatenation is the most commonly used method for feature-level fusion, but it might not be the best choice for multimodal inputs that differ in dimensions.

Decision-level fusion, on the other hand, does not require synchrony between different modalities. SVR as a method of decision-level fusion can compensate for redundant information, but it underperform when the number of features exceeds the number of training samples. A blending algorithm works well as a decision-level fusion technique for emotion recognition owing to its improved performance for multimodal inputs. Blending model when used for decision-level fusion is stable and less noisy, but this technique has high time and space complexity. Brain emotional learning can be used at decision-level fusion owing to its good decision-making capabilities. In spite of this, expertise in cognitive-based abilities of the brain is needed to exploit this technique.

Decision trees are considered for decision-level fusion because of their capability of learning rules according to the requirement. But, decision trees are unstable data structures and they are prone to inaccuracy. Rule - based classification is a way to achieve decision-level fusion as they have the advantage of fuse the multimodal signal classifier's output as per the conditions specified in the antecedent. Nonetheless, this requires a deep understanding of the domain for deciding the conditions and demands a lot of manual work. The D-S theory is widely explored for decision-level fusion. It is based on the concepts of allocating appropriate beliefs and probabilities to hypotheses, applying the D-S rule for fusing independent items from different sources and arriving at the final decision of the optimal hypothesis in a workable and reasonable manner.

TABLE II. MULTI-PERSPECTIVE SUMMARIZATION OF MULTIMODAL FUSION METHODS

Paper	Modalities	Task	Dataset	Model	Fusion Method	Fusion Technique	Results
H. Su et al. [10]	Audio, Video and Text	Multi-level segmented decision-level fusion emotion recognition	AVEC2017	BLSTM	Decision Level	SVR	Improved CCC performance of 0.685 on arousal
Nicolaou et al. [13]	Audio and Video	Fusion of audio cues, facial expression and shoulder gesture for continuous emotion prediction	SAL	BLSTM-NN	Decision Level, Feature Level and Output Associative	SVR	Obtained COR as 0.796 & 0.642 and RMSE as 0.15& 0.21 for valence and arousal respectively
Man Hao et al. [14]	Audio and Video	Audio visual Emotion Recognition Framework	eNTERFACE	SVM and CNN	Decision Level	Blending Ensemble	Improved recognition SI accuracy of 81.36% and SD Accuracy of 78.42%
Zeinab Farhoudi and Saeed Setayeshi [17]	Audio and Video	Audio-visual Emotion recognition with the MoBEL fusion network	eNterface'05	CNN and RNN	Decision Level and Feature Level	BEL and MoBEL	Improved audio visual emotion recognition accuracy of 81.7%
Y. Cimtay et al. [20]	Facial expression GSR and EEG	Hybrid multimodal emotion recognition	LUMED-2 and DEAP	CNN	Decision Level	Decision Tree	Obtained maximum one-subject-out accuracy of 91.5% and mean accuracy of 53.8%
S. Sahoo and A. Routray [22]	Audio and Video	Audio visual Emotion Recognition	eNTERFACE'05	HMM and SVM	Decision Level	Rule Based	Obtained average recognition accuracy of 76% for males candidates and 86% for female candidates for subject-dependent cases
S. Nemati [28]	Audio, video and text	Emotion Recognition using audio, video and users' comments	DEAP	SVM and Naïve Bayes	Hybrid	CCA and D-S theory	Obtained 0.85% and 76% of accuracy and f1-measure respectively for Emotion Recognition
Liam Schoneveld et al. [39]	Audio and Video	Emotion recognition using recent advances in deep learning	RECOLA, AffectNet and Google FEC	MTCNN, VGG, LSTM	Feature Level	Concatenation	Obtained CCC of 0.740 for valence and 0.719 for arousal in RECOLA

Although it works well for non-conflicting data, it may lead to inaccuracy for data that are conflicting in nature. The decision template algorithm is also widely used for decision fusion, but it has limitations as it uses average decision profiles and does not emphasize differences in the performance of classifiers. The hybrid fusion method achieves the advantages of both feature level and decision-level fusion methods by combining these approaches. Feature-level fusion methods require synchronization between different modalities, but it is possible to produce results even when the data from one of the modality is missing. Decision-level fusion, on the other hand, does not require synchrony between different modalities, but its capabilities cannot be exploited fully even if the data from a single modality is missing.

The hybrid approach overcomes the limitations of feature and decision level based approaches by applying feature-level fusion on the synchronous modalities and combining the results of with the remaining asynchronous modalities using decision-level fusion.

Results of experimental analysis on audio, facial images and text for emotion recognition show that feature level fusion of text with other two modalities was impractical since the modalities differ in temporal characteristics. Decision-level fusion works well for fusing the modalities differing in time synchronicity. However, Hybrid fusion works best for emotion recognition using multiple modalities which differ in time synchronicity providing the accuracy of 66.67%. These results can further be improved by taking large data samples for testing.

VII. CONCLUSION

This paper discussed different multimodal fusion techniques for human emotion recognition. The study proposed a review of different state of art techniques for decision level, feature level and hybrid fusion to achieve multimodal emotion recognition.

The paper concentrated on using facial images, audio and text as multimodal inputs for emotion recognition. Experimental analysis of fusion methods was conducted which claimed that decision-level fusion and feature level fusion can be performed when inputs of different modalities are synchronous in time. However, when the inputs are not synchronous in time, suitable choices are decision level fusion and hybrid fusion. Hybrid fusion using concatenation and blending algorithm worked best improving accuracy by 7.6 % compared to decision level fusion. The study however had a limited test dataset. The accuracy results can be further improved if large number of annotated data samples are used for testing.

The review suggested that D-S theory is the most aggressively used method for decision-level fusion followed by rule-based and decision tree methods. Brain emotional learning is the latest approach used but it requires expertise in cognitive-based abilities of the brain. However, the blending algorithm works well for emotion recognition owing to its improved performance for multimodal inputs. For feature-level fusion, the most commonly used method is concatenation. CCA and Eigen matrix methods are also used in many applications for feature level fusion. The state of the art literature for feature fusion uses brain emotional learning but it requires expertise in cognitive-based abilities of the brain. Hybrid method is not explored much in literature. However, from experimental analysis it can be deduced that hybrid approach works best in fusing multiple modalities which differ in time synchrony. This provides a direction for the researchers to explore hybrid fusion approach on multimodal inputs.

The experimental analysis performed in this paper is based on audio, text and image inputs for emotion recognition. As a part of future work, audio-visual and physiological signals like EEG can be used to analyze different fusion techniques. Also, this paper used concatenation and meta classifier technique to evaluate effect of feature level, decision level and hybrid fusion. However, as a part of future work, there is a scope for applying different techniques of feature and decision level fusion discussed in the literature review in order to evaluate their effect on decision level, feature level and hybrid fusion.

REFERENCES

[1] J. Wang and M. Wang, “Review of the emotional feature extraction and classification using EEG signals,” Cogn. Robot., vol. 1, pp. 29-40, ISSN 2667-2413, 2021, doi:10.1016/j.cogr.2021.04.001.

[2] I. B. Mauss and M. D. Robinson, “Measures of emotion: A review,”, Cogn. Emot., vol. 23, no. 2, pp. 209-237, 2009, doi:10.1080/02699930802204677.

[3] P. Chavan and S. Desai, “A Review on BCI Emotions Classification for EEG Signals Using Deep Learning” ,IOS press journal, Advances in Parallel Computing. doi:10.3233/APC210241.

[4] M. Agrawal et al., “Models for hand gesture recognition using deep learning,” 5th International Conference on Computing Communication and Automation (ICCCA), vol. 2020. IEEE, 2020, pp. 589-594, doi:10.1109/ICCCA49541.2020.9250846.

[5] D. Agarwal and S. Desai, “Multimodal techniques for emotion recognition,” International Conference on Computational Intelligence and Computing Applications (ICCICA), vol. 2021, 2021, pp. 1-6, doi:10.1109/ICCICA52458.2021.9697294.

[6] H. Gunes and M. Pantic, “Automatic, dimensional and continuous emotion recognition,” Int. J. Synth. Emot., vol. 1, no. 1, pp. 68-99, 2010, doi:10.4018/jse.2010101605.

[7] S. Chen and Q. Jin, “Multi-modal dimensional emotion recognition using recurrent neural networks,” Proc. 5th International Workshop on Audio/Visual Emotion Challenge. New York, NY, USA: Association for Computing Machinery, 2015, pp. 49-56, doi:10.1145/2808196.2811638.

[8] B. Huang et al., “A review of multimodal medical image fusion techniques,” Comp. Math. Methods Med., vol. 2020, p. 2020, article ID 8279342, 16, 2020. doi:10.1155/2020/8279342.

[9] Chen Xiao Yu et al., “Research Progress and Trend of the Machine Learning based on Fusion,” Int. J. Adv. Comput. Sci. Appl.(IJACSA), vol. 13, no. 7, 2022. doi:10.14569/IJACSA.2022.0130701.

[10] H. Su et al., “An improved multimodal dimension emotion recognition based on different fusion methods,” 15th IEEE International Conference on Signal Processing (ICSP), vol. 2020, 2020, pp. 257-261, doi:10.1109/ICSP48669.2020.9321008.

[11] K. S. Ni and T. Q. Nguyen, “Image superresolution using support vector regression” in IEEE Trans. Image Process., vol. 16, no. 6, pp. 1596-1610, Jun. 2007, doi:10.1109/TIP.2007.896644.

[12] J. Huang et al., “Continuous multimodal emotion prediction based on long short term memory recurrent neural network” in Proc. 7th Annual Workshop on Audio/Visual Emotion Challenge (AVEC ’17). New York, NY, USA: Association for Computing Machinery, 2017, pp. 11-18. doi:10.1145/3133944.3133946.

[13] M. A. Nicolaou et al., “Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space” in IEEE Trans. Affect. Comput., vol. 2, no. 2, pp. 92-105, Apr.-Jun. 2011, doi:10.1109/T-AFFC.2011.9.

[14] M. Hao et al., “Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features.”, Neurocomputing, vol. 391, pp. 42-51, ISSN 0925-2312, 2020, doi:10.1016/j.neucom.2020.01.048. Available at: https://www.sciencedirect.com/science/article/pii/S09252 31220300990.

[15] C. M. Lee and S. S. Narayanan, “Toward detecting emotions in spoken dialogs” in IEEE Trans. Speech Aud. Process., vol. 13, no. 2, pp. 293-303, Mar. 2005, doi:10.1109/TSA.2004.838534.

[16] F. Wubing et al., “An improved fuzzy brain emotional learning model network controller for humanoid robots,” Front. Neurorobotics, vol. 13, 2, 2019 , ISSN=1662-5218, doi:10.3389/fnbot.2019.00002.

[17] Z. Farhoudi and S. Setayeshi, “Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition,” Speech Commun., vol. 127, pp. 92-103, ISSN 0167-6393, 2021, doi:10.1016/j.specom.2020.12.001.

[18] S. I. Rusu and P. C. M. A. “ Learning, memory and consolidation mechanisms for behavioral control in hierarchically organized corticobasal ganglia systems,”, Hippocampus, vol. 30, no. 1, pp. 73-98, 2020 Jan.. doi:10.1002/hipo.23167.

[19] E. Lotfi and M.. Akbarzadeh-T, “ Supervised Brain Emotional Learning” The, vol. 2012, pp. 1-6, 2012 International Joint Conference on Neural Networks (IJCNN), doi:10.1109/IJCNN.2012.6252391.

[20] Y. Cimtay et al., “Cross-subject multimodal emotion recognition based on hybrid fusion” in IEEE Access, vol. 8, pp. 168865-168878, 2020, doi:10.1109/ACCESS.2020.3023871.

[21] H. Kaya et al., “Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video CVs,” IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), vol. 2017, 2017, pp. 1651-1659, doi:10.1109/CVPRW.2017.210.

[22] S. Sahoo and A. Routray, “Emotion recognition from audio-visual data using rule based decision level fusion,” vol. 2016, pp. 7-12, 2016 IEEE Students’ Technology Symposium (TechSym), doi:10.1109/TechSym.2016.7872646.

[23] A. P. Dempster, “Upper and lower probabilities induced by a multivalued mapping,” Ann. Math. Statist., vol. 38, no. 2, pp. 325-339, 1967, doi:10.1214/aoms/1177698950.

[24] G. Shafer, A Mathematical Theory of Evidence. Princeton, NJ, USA: Princeton Univ. Press, 1976.

[25] Y.-T. Liu et al., “Weighted fuzzy Dempster–Shafer framework for multimodal information integration” in IEEE Trans. Fuzzy Syst., vol. 26, no. 1, pp. 338-352, Feb. 2018, doi:10.1109/TFUZZ.2017.2659764.

[26] N. Khan and S. Anwar, “Improved Dempster-Shafer sensor fusion using distance function and evidence weighted penalty: Application in object detection” in Proc. 16th International Conference on Informatics in Control, Automation and Robotics, vol. 1: ICINCO, ISBN 978-989-758-380-3, 2019, pp. 664-671. doi:10.5220/0007917106640671.

[27] X.-D. Zhang, “Study on Feature Layer fusion Classification Model on Text/Image Information,” Phys. Procedia, vol. 33, pp. 1050-1053, ISSN 1875-3892, 2012, doi:10.1016/j.phpro.2012.05.172..

[28] S. Nemati, “Canonical correlation analysis for data fusion in multimodal emotion recognition,” 9th International Symposium on Telecommunications (IST), vol. 2018, 2018, pp. 676-681, doi:10.1109/ISTEL.2018.8661140.

[29] A. Mi et al., “A Multiple Classifier Fusion Algorithm Using Weighted Decision Templates,” Sci. Program., vol. 2016, pp. 1-10, 2016, doi:10.1155/2016/3943859.

[30] R. Ebrahimpour and S. Hamedi, Hand Written Digit Recognition by Multiple Classifier Fusion Based on Decision Templates Approach, 2009. doi:10.5281/zenodo.1069935.

[31] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," IEEE Trans. Syst. Man Cybern. B Cybern., vol. 39, no. 1, pp. 64-84, 2009, doi:10.1109/TSMCB.2008.927269.

[32] C. Kamlaskar and A. Abhyankar, "Feature level fusion framework for multimodal biometric system based on CCA with SVM classifier and cosine similarity measure," Aust. J. Electr. Electron. Eng., 1-14, 2022, doi:10.1080/1448837X.2022.2129147.

[33] A. Corradini et al., "Multimodal input fusion in human computer interaction on the example of the on-going nice project" in Proc. NATO-ASI Conference Data Fus, Situation Monit. Incident Detect. Alert Resp. Manag., pp. 223-234, 2003.

[34] J. Che et al., "Longitudinal tear detection method of conveyor belt based on audio-visual fusion," Measurement, vol. 176, p. 109152, ISSN 0263-2241, 2021, doi:10.1016/j.measurement.2021.109152.

[35] N. M. Correa et al., "Canonical correlation analysis for data fusion and group inferences: Examining applications of medical imaging data," IEEE Signal Process. Mag., vol. 27, no. 4, pp. 39-50, 2010, doi:10.1109/MSP.2010.936725.

[36] B. E. Stein et al., "The neural basis of multisensory integration in the midbrain: Its organization and maturation," Hear. Res., vol. 258, no. 1-2, 4-15, 2009, doi:10.1016/j.heares.2009.03.012.

[37] C. Ma et al., "Multi-layers feature fusion of convolutional neural network for scene classification of remote sensing" in IEEE Access, vol. 7, pp. 121685-121694, 2019, doi:10.1109/ACCESS.2019.2936215.

[38] L. Schoneveld et al., "Leveraging recent advances in deep learning for audio-Visual emotion recognition," Pattern Recognit. Lett., vol. 146, pp. 1-7, ISSN 0167-8655, 2021, doi:10.1016/j.patrec.2021.03.007.

[39] S. Lee et al., "Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification" in IEEE Access, vol. 9, pp. 94557-94572, 2021, doi:10.1109/ACCESS.2021.3092735.

[40] Z. Lan et al., "Multimedia classification and event detection using double fusion,", Multimed. Tools Appl., vol. 71, no. 1, 333-347, 2014 doi:10.1007/s11042-013-1391-2.

[41] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PLoS ONE, vol. 13, no. 5, p. e0196391, 2018. doi:10.1371/journal.pone. 0196391.

[42] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests" in Lect. Notes Comput. Sci.. ICONIP 2013, vol. 8228, M. Lee, A. Hirose, Z. G. Hou and R. M. Kil, Eds. Neural Information Processing. Berlin, Heidelberg: Springer, 2013. doi:10.1007/978-3-642-42051-1_16.

[43] E. Saravia et al., "CARER: Contextualized affect representations for emotion recognition" in Proc. 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 3687-3697, doi:10.18653/v1/D18-1404.