IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 December 2022, accepted 18 January 2023, date of publication 30 January 2023, date of current version 2 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3240420

RESEARCH ARTICLE

A Framework to Evaluate Fusion Methods for Multimodal Emotion Recognition

DIEGO PEÑA[1], ANA AGUILERA[2], IRVIN DONGO[3,4,5], JUANPABLO HEREDIA[6], AND YUDITH CARDINALE[1,5]

[1]Departamento Computación y Tecnología de la Información, Universidad Simón Bolívar, Caracas 1080, Venezuela
[2]Escuela de Ingeniería Informática, Facultad de Ingeniería, Universidad de Valparaíso, Valparaíso 2340000, Chile
[3]Electrical and Electronics Engineering Department, Universidad Católica San Pablo, Arequipa 04001, Peru
[4]ESTIA Institute of Technology, University of Bordeaux, 64210 Bidart, France
[5]Grupo de Investigación en Ciencia de Datos (GRID), Universidad Internacional de Valencia, 46002 Valencia, Spain
[6]Computer Science Department, Universidad Católica San Pablo, Arequipa 04001, Peru

Corresponding authors: Irvin Dongo (ifdongo@ucsp.edu.pe) and Ana Aguilera (ana.aguilera@uv.cl)

ABSTRACT Multimodal methods for emotion recognition consider several sources of data to predict emotions; thus, a fusion method is needed to aggregate the individual results. In the literature, there is a high variety of fusion methods to perform this task, but they are not suitable for all scenarios. In particular, there are two relevant aspects that can vary from one application to another: (i) in many scenarios, individual modalities can have different levels of data quality or even be absent, which demands fusion methods able to discriminate non-useful from relevant data; and (ii) in many applications, there are hardware restrictions that limit the use of complex fusion methods (e.g., a deep learning model), which could be quite computationally intensive. In this context, developers and researchers need metrics, guidelines, and a systematic process to evaluate and compare different fusion methods that can fit to their particular application scenarios. As a response to this need, this paper presents a framework that establishes a base to perform a comparative evaluation of fusion methods to demonstrate how they adapt to the quality differences of individual modalities and to evaluate their performance. The framework provides equivalent conditions to perform a fair assessment of fusion methods. Based on this framework, we evaluate several fusion methods for multimodal emotion recognition. Results demonstrate that for the architecture and dataset selected, the methods that best fit are: Self-Attention and Weighted methods for all available modalities, and Self-Attention and Embracenet+ when a modality is missing. Concerning the time, the best times correspond to Multilayer Perceptron (MLP) and Self-Attention models, due to their small number of operations. Thus, the proposed framework provides insights for researchers in this area to identify which fusion methods better fit their requirements, and thus to justify the selection.

INDEX TERMS Emotion recognition, fusion methods, multimodality.

I. INTRODUCTION

People manifest emotions with verbal, spontaneous and automatic non-verbal expressions, which make them easy to interpret by other people, but complex to digitally represent and identify [1], [2], [3]. Thus, non-verbal communication (e.g., body movements, face gestures) along with actual words are useful to decipher the true emotions of a speaker, both for other people and for automatic emotion recognition systems [4]. Currently, there is an increasing interest in the scientific community in producing computational methods that can consider different information conveyed by different modalities (both verbal and non-verbal) in order to produce a more accurate reading of people's emotions [5].

In the state of the art, exists a huge diversity of methods to analyze a single modality and identify the underlying emotion (or emotions) behind the sampled data from a single source – e.g., text [6], [7], [8], voice [9], [10], [11], facial gestures [12], [13], [14]. However, there is still a lack of additional strategies to fully take advantage of the complementary information that

The associate editor coordinating the review of this manuscript and approving it for publication was Hossein Rahmani.

one modality could offer to another in certain situations. That is where multimodal data fusion can step in.

The complication behind multimodal techniques for emotion recognition is the fact that not all sources or modalities carry the same quality of information [15], [16] – i.e., in some situations, one modality may fail to produce the expected output (or any output at all) for a given sample [17]. In such scenarios, the fusion process should aggregate these different data sources in a specific way that captures the relationships between them [18], and, if possible, find a way to rely more on certain modalities when the quality of the sample from another one proves to be particularly poor [16]. Most multimodal fusion methods are supported by deep learning models that are used to obtain a new representation that fuses all different data, classifies such representation, or both [18], [19], [20], [21]. Given the fact that human interaction by nature uses multiple channels simultaneously when transmitting a message to another human, and the human brain has the capacity to process all that information [22], it is easy to understand why multimodal data fusion could be a natural fit for the problem of emotion recognition.

Despite the advantages, deep learning techniques for multimodal fusion also come with some drawbacks, mostly related to the complexity of the process and to the variability of data quality coming from different sources. For instance, many state-of-the-art models for certain modalities (like videos) can require a particularly big number of operations during their analysis [23]; moreover, a fusion model would require computations from all its individual unimodal models and also some additional for the fusion itself in order to make a prediction (or make a backward pass, if trained end-to-end); in this case, a fusion model could end up being quite computationally intensive to be executed in limited hardware or with battery saving restrictions (e.g., smartphones, social robots). Another major issue fusion methods face is the existence of scenarios where not all the sources are equally reliable (e.g., a camera in a dimly lit space identifying faces, a sound sensor in a noisy space catching the voice) [17]. Thus, in those cases, any prediction must be made by diminishing the importance of the data from such as source or even ignoring it.

In the literature, there is a high variety of fusion methods to perform multimodal emotion recognition, but they do not behave the same in all scenarios, concerning to the consideration of different levels of data quality from different sources and to the time complexity. In this context, developers and researchers need metrics, guidelines, and a systematic process to evaluate and compare different fusion methods that can fit to their particular application scenarios.

As a response to this need, we propose a framework that establishes a base to perform a comparative evaluation of fusion methods to demonstrate how they adapt to the quality differences of individual modalities and to evaluate their performance. The framework provides equivalent conditions to perform a fair assessment of fusion methods. Based on this framework, we evaluate nine fusion models in the context of multimodal emotion recognition. The framework offers three deep learning models to make prediction of emotions from three individual modalities (i.e., facial gestures, audio, and text); and then passes the results to the different fusion models. This framework proposes guidelines to perform the comparative evaluation.

Based on the comparative evaluation performed with the framework proposed, we show how some fusion methods manage to achieve a certain degree of resiliency in the absence of data for a certain modality, and that some of them even achieve an improvement from the most accurate individual modality. For the comparative experiments the Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database was used.[1] Results show that for this combination of architecture/dataset, Self-Attention and Weighted methods are the best when all modalities are available and Self-Attention and Embracenet+ bahave the best when a modality is missing. Concerning the execution time, the best times correspond to Multilayer Perceptron (MLP) and Self-Attention methods, due to both having small number of operations. Hence, the framework proposed provides insights for researchers in this area to identify which fusion methods better fit their requirements and to justify the selection. In summary, the main contribution of this paper is three-fold:

- A framework to compare different fusion methods under the same conditions to produce a fairer evaluation and assessment of their capabilities, which could be used by researchers on the topic to inform their decisions on which fusion method better fit their project.
- A demonstration on how different multimodal methods adapt to different data quality or the presence or absence of certain modalities.
- A comparative evaluation in terms of execution time of the considered fusion methods.

With the ability of evaluating these aspects, the fusion methods can be selected according to the application scenarios, resulting on more suitable and appropriate decisions.

This article is structured as follows. In Section II, a review of some recent works on the topic of fusion methods for multimodal emotion recognition is presented. In Section III the fusion methods evaluated are described. In Section IV, the evaluation framework is described. In Section V the evaluation and comparison of results are presented and discussed with a detailed explanation of the training process for the models and experimental procedure. Finally, Section VI presents the conclusions and future work.

II. RELATED WORK

The idea of multimodal fusion has been applied in numerous contexts. The survey presented in [24] describes the state of the art for multimodal fusion methods for urban data (input collected from meteorological stations, taxi GPS, pollution levels, traffic volume, among others) to make predictions on a variety of things from crowd flows to air quality levels. In a different field in [25], several fusion techniques

[1] https://sail.usc.edu/iemocap/

for early detection of Alzheimer's disease are described. Authors compare early and late fusion shallow models, like K-neighbors and random forests, with a deep learning model that uses a concatenation of an intermediate representation of the data from each modality (i.e., clinical data, imaging, and genetic). In this case, the final classification is done using a simple multilayer perceptron (MLP). This proposal is fairly common, in multiple contexts. In [26], authors use an MLP as a way to fuse data from the Persuasive Opinion Multimedia (POM) dataset in order to improve the quality of the predictions. In this case, predictions are confidence scores as to whether the sample is persuasive or not. In the work presented in [27] for the problem of music genre classification, before the fusion step, the authors apply some techniques such as matrix multiplication and trigonometric functions to embed the results from each modality (i.e., audio tracks, text reviews, and cover art images) in a new multimodal space that optimizes the similarities among modalities. The vectors obtained in this space are then concatenated with the original ones, as if all are from different sources and the resulting feature vector is used as input for the MLP. In [28], [29] it is also used an MLP with the concatenation of each of its modalities (i.e., text and images) for document classification in case of [28], and emotion classification in [29]. The latter set the analysis of emotions in images by context levels and processes each defined context as the modalities.

More complex fusion models have also been explored, that can be applied in any context. Embracenet, a deep learning model that relies on multinomial distribution to create a feature vector from each individual modality's resulting vector is presented in [17]. This mechanism not only helps to prevent overfitting but also allows discarding information from sources where data is unavailable, as it is designed to select features from sources with probability zero. In theory, the mechanism is designed to offer enough flexibility to be used in multiple contexts, although it was initially tested in the context of gas recognition (using the Gas Sensor Array dataset) [30] and in human activity recognition (using the OPPORTUNITY dataset) [31]. DeepFusion is another model aimed at mitigating the issue of missing/poor quality data from one of the modalities [16]. DeepFusion learns to calculate a weight that represents the approximate quality of the sample received from each of the sources or how informative each of them is. This way, the final classification should lean more on the data that it believes more valuable to make a prediction. DeepFusion also possesses layers that find the correlation between the samples from different modalities, which combined with the previously mentioned quality-determining layers, allows the model to obtain state-of-the-art level results in the problem of activity recognition.

In the context of emotion recognition, multimodal fusion methods have been continuously studied. A variety of models have been tested, from the concatenation of individual modalities' results passed to an MLP [32] to more sophisticated recurrent networks that allow evaluating the temporal dependencies present in the samples, such as the LSTM-based models proposed in [33], [34], [35] or the biGRU mechanism found in [36]. Attention mechanisms have also been very popular. The model proposed in [37] achieves state-of-the-art results for the IEMOCAP dataset by using an attention mechanism to fuse audio and text modalities [5]. In [38], the Multi-modal Attention (MMA) module is used to fuse features from audio, body language and text adaptively, also over the IEMOCAP dataset. Similarly, the RAVEN model is used in [15] to fuse audio, text, and video. Attention as part of the fusion mechanism has also been tested on other datasets, like CMU-MOSEI, as detailed in [39], or HEU part one and two [40]. In the latter, a multi-modal attention mechanism (cLSTM-MMA) is proposed, which facilitates the attention across three modalities in their Multi-modal Attention Network (MMAN). Others complex methods [41], [42], [43] achieve high performances with, such as DEAP, SEED or EmoFBVP. In [41] the audio and visual modalities are fused using a latent space linear map and then, their projected features into the cross-modal space are fused with the textual modality using a Dempster-Shafer (DS) theory-based evidential fusion method. Authors in [42], present the convolutional deep belief network (CDBN) models that learn salient multimodal features of expressions of emotions, which achieve better results in low-intensity emotion expressions than state-of-the-art methods. In [43] is used the restricted Boltzmann machine (RBM) model to construct a Bimodal Deep autoencoder (BDAE) whose extracted high-level representation features are shown to be effective for emotion recognition using.

Other types of models have also been proposed, such as linear regression usage and Tensorfusion. In [44], the authors use linear regression on all single systems for arousal and valence. Linear regression is used also for the fusion. Tensorfusion, a fusion method based on tensor multiplication, is described in [45], in which each tensor represents a different modality. A more efficient version of this method that relies on low-rank weighted decomposition, designed with high-dimensioned tensors in mind and tested in multiple datasets, is proposed in [46]. The results obtained for the IEMOCAP dataset, rank among the best ever obtained to date [5]. A modified Embracenet version, called Embracenet+, is proposed in [47] that proves to be slightly more accurate than the Embracenet when tested using the EMOTIC dataset. A fusion method that uses a combination of transformers and attention is described in [48]. This work considers three input modalities of text, audio (speech), and vision with features extracted from independently pre-trained Self Supervised Learning models. A method that benefits from intra-modal attention mechanisms and Factorized Bilineal Pooling (FBP) for cross-modality fusion is proposed in [49]; it obtains competitive results for the AFEW dataset, which contains ''in the wild'' samples (taken from movies and TV, rather than from laboratory sessions designed specifically to generate data).

TABLE 1. Fusion methods.

Reference	Approach and method name (if available)	Context	Dataset	Modalidaties	Limitations
Wang et al. [15]	Recurrent attention (RAVEN)	Emotion recognition	IEMOCAP	Audio, video and text	Poor performance with sequences extracted at a higher frequency for visual and acoustic modalities.
Xue et al. [16]	Attention or quality factor (DeepFusion)	Human activity recognition	Self-built dataset	Sensor signals	Extra burden and discomfort from users wearing the sensors to data capture.
Choi and Lee [17]	Probability based (Embracenet)	Gases and human, activity recognition	Gas sensor array dataset, OPPORTUNITY	Sensor signals	When only one modality is available the performance decrease due to no correlated information between modalities to exploit.
Venugopalan et al. [25]	MLP	Alzheimer detection	ADNI	Image, text, sequences	Limited size of the training dataset (only a few thousand samples in total and even fewer samples with all three modalities).
Nojavanasghari et al. [26]	MLP	Persuasion detection	POM	Face, audio and text	The model is not able to deal with the noisy training data. Features with different representations are treated equally producing a lower performance for the classifier.
Audebert et al. [28]	MLP	Document classification	RVL-CDIP, Tobacco3482	Text and images	Neither takes into account missing modalities, nor regulates the influence of each modality for a given sample. Data collected under very controlled conditions.
Oramas et al. [27]	MLP ((after additional information calculation))	Musical genre recognition	Million Song Dataset	Songs and album covers	Does not explicitly regulate the influence of each modality for a given sample.
Tripathi and Beigi [32]	MLP	Emotion recognition	IEMOCAP	Audio, text and *motion capture*	Does not take into account the missing modalities, nor does it regulate the influence of each modality for a given sample. Data collected under highly controlled conditions..
Tzirakis et al. [33]	Self-Attention, Hierarchical attention, LSTM	Emotion recognition	SEWA	Audio, text and Face	LSTM may not be the best-suited classifier for non-sequential features. The implementation may not be as straightforward as other methods.
Xu et al. [34]	LSTM with alignment mechanism	Emotion recognition	IEMOCAP	Audio and text	The use of LSTM for non-sequential features may not be the best-suited case. The method is not flexible enough to be directly applied to a different set of modalities.
Akhtar et al. [36]	BiGRU, Neural layers	Emotion and sentiment recognition	CMU-MOSEI	Audio, text and face	The method is not fault-tolerant enough to be applied in real-world situations.
Priyasad et al. [37]	Self-Attention (Self-Attention)	Emotion recognition	IEMOCAP	Audio and text	Data collected under very controlled conditions.
Choi et al. [39]	Attention matrix	Emotion recognition	CMU-MOSEI	Audio and text	Not flexible enough to be directly applied to more than two modalities without additional logic and adaptations.
Zadeh et al. [45]	Tensor multiplication, MLP (Tensor Fusion)	Sentiment Recognition	CMU-MOSI	Audio, text and face	The number of necessary operations will grow quickly with the number of modalities or features per modality, as it involves the tensor product of all feature vectors.
Liu et al. [46]	*Low-rank weighted decomposition*, MLP (Tensor Fusion)	Emotion recognition	IEMOCAP	Audio, text and video	Implementation is less straightforward than the original version of Tensor Fusion [45]
Heredia et al. [47]	Probability based (Embracenet+)	Emotion recognition	IEMOCAP	Audio, video and text	Performance is highly dependent on the individual processing of each modality.
Siriwardhana et al. [48]	Transformers, Hadamard product, MLP	Emotion and sentiment recognition	IEMOCAP, CMU-MOSEI	Audio, text and video	Uses independently trained models for each modality that could be related.
Zhou et al. [49]	Attention intra-modal, FBP	Emotion recognition	AFEW	Audio and visual	Focused on feature fusion strategies for audio and visual modalities. Training using only pre-trained networks.
Chen et al. [38]	Attention (MMA)	Emotion recognition	HEU-1, HEU-2	Face, posture and audio	Does not explicitly capture the intermodal temporal relationships. Was tested on a new dataset where most state of the art methods have not been tested.
Pan et al. [40]	Attention (MMAN)	Emotion recognition	IEMOCAP	Audio, Body language and text	Focused on speech emotion recognition.
Mittal et al. [29]	MLP	Emotion recognition	EMOTIC	Face, posture, context and depth	Model often confuses between certain class labels. Context interpretations are required to improve the model accuracies.
Poria et al. [35]	LSTM	Emotion and sentiment recognition	CMU-MOSI, MOUD, IEMOCAP	Audio, video and text	Wrong model predictions with face occlusion or noisy audio, weak and non contextual sentiments with bias towards its surrounding utterances.
Nemati et al. [41]	MFA, SVM, Dempster-Schafer	Sentiment recognition	DEAP	Audio, video and text	Intrinsic problems associated to the evidential Dempster-Shafer-based fusion.
Ranganathan et al. [42]	DBN, SVM	Emotion recognition	EmoFBVP	Face, body gestures, voice and physiological signals	Limited to DBN models with SVM as baselines in validation. Absence of modalities is not considered.
Liu et al. [43]	RBM, SVM	Emotion recognition	DEAP and SEED	physiological signals (eyes and brain)	Emotion understood as a value in four-dimensional arousal-valence-dominance-linking continuous space.
Povolny et al. [44]	Lineal regression (LR)	Emotion recognition	RECOLA	Audio, text, video and physiological signals	Limited to LR models without an explicit validation of assumptions to apply them. Focused on features derived from audio. Emotion only understood as a value in two-dimensional arousal-valence continuous space.

All these works are summarized in Table 1. They demonstrate the recent interest in developing fusion approaches for multimodal classification methods and the wide variety of proposals. To decide which is the most suitable and efficient in specific scenarios, it is important to evaluate them in terms of adaptability to different quality of modalities and time consumption. In particular, it is relevant to make a comparison of these methods in a context of multimodal emotion recognition which is a trending topic in Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI).

III. FUSION METHODS TO BE EVALUATED

Fusion methods have been categorized in different ways. The most popular categorization considers mainly three types of fusion strategies: information/data fusion (low-level fusion), feature fusion (intermediate-level fusion), and decision fusion (high-level fusion) [50]. Data fusion combines several sources of raw data to produce new raw data that is expected to be more informative and synthetic than the inputs. Feature fusion concatenates all the features into a single representation. The concatenated heterogeneous features are then directly fed into the classifiers to train the models [51]. Decision fusion methods use a set of classifiers to provide a better and unbiased result. The classifiers can be of same or different type and can also have same or different feature sets [52]. The fusion type of the methods studied in this work are based on decision fusion, sometimes called also late fusion.

In this work, we evaluate nine fusion models, which are a representative set of a variety of deep learning techniques. A brief description of each one is provided below:

- Simple dense neural network/multi-layer perceptron (DNN/MLP): It is possibly the simplest type of neural network with multiple neurons and a very common fusion method, as demonstrated by several related works.
- An attention mechanism plus an MLP: This is based on the work proposed in [39] and uses a structure similar to the one in [48], although with completely different techniques. Since this work is referred to as bimodal fusion, an adaptation is made – i.e., bimodal attention is applied to each pair of modalities. In this case, for each modality, two attention matrices are produced, which are then separately multiplied by the modalities' vector. The result is two different vectors that are reduced to one vector of the size of the original and fused with an MLP for the final classification.
- The DeepFusion model proposed in [16]: This is a unified multi-sensor deep learning framework designed to learn informative representations of heterogeneous sensory data. This framework combines different sensors' information weighted by the quality of their data and incorporates cross-sensor correlations. The framework is composed of two sub-modules, the quality weighted and the correlation sub-modules. Afterward, results from each of these sub-modules are concatenated, and a linear layer produces a vector of size four (the number of categories). This vector is then applied to a softmax fusion for final classification. In the context of our work, we additionally test each sub-module as an individual fusion method, as we explain in the following two items.
- The Weighted Combination sub-module of the DeepFusion model: It uses a single neuron with a variation of the sigmoid activation function per modality to determine a scalar weight or factor for the quality of the modality's output. A softmax function is then applied over the vector formed by the resulting factors to normalize them with samples from the test dataset. Afterward, each normalized factor is multiplied by their respective modality's output and the obtained vectors are added. This result is passed to a 2-layer Gated Recurrent Unit (GRU) that makes the final classification. A softmax layer is later applied when this module is used individually, The GRU is made to comply with the design proposed in [16]. Since the input of this network is not a sequence, a more straightforward no-recurrent type of network might have sufficed.
- The Cross-modality sub-module of the DeepFusion model: It relies on subtraction, concatenation, and a couple of layers to make its classification. First, for each modality, a couple of subtractions are performed (i.e., between the modality output vector and those of the others). The resulting vectors are then concatenated and fused with a ReLU layer. This output is a vector with the same size as the output vector (four in this case) of individual modalities. Then, the vectors obtained from this process (one for each modality) are averaged, resulting in the final classification (a softmax layer is later applied when this module is used individually, as we explained in the previous item).
- The Tensor Fusion model proposed in [45]. It is a multimodal fusion approach which explicitly aggregates unimodal, bimodal, and trimodal interactions using a 3-fold Cartesian product from modality embeddings in the context of multimodal sentiment analysis. An extra constant dimension with value 1 is added to each vector to generate the unimodal and bimodal dynamics. As a result of this process a, tridimensional tensor is obtained. Afterward, this tensor is flattened and fed to an MLP for final classification.
- The Embracenet method proposed in [17]. It is one of the easier models to adapt to the context in which it is being used and one of the few that explicitly considers the problem of missing modalities. This method uses a docking layer to transform the generated vector from each modality into a vector of a specific size using a ReLU layer. Then, a new vector is created with this size, selecting its $i-th$ element from among the $i-th$ elements of the vectors obtained in the docking layer. The $i-th$ element from a specific modality is selected with a probability learned during the training process.
- Embracenet+, presented in [53] as an alternative or improvement of the Embracenet model. This model has an architecture that involves three simple Embracenet models working to improve the modalities' correlation learning as well as the final results. Each Embracenet model used has one more linear layer and a dropout layer, which hardens the model a bit to improve learning. Being based on Embracenet, this model is flexible in terms of number of used modalities and fault or missing data tolerance.

- Self-Attention: This mechanism was originally proposed by [37]. Attention is applied individually to each modality before concatenating the resulting vectors and using a linear layer with softmax for classification. In this case, the attention value is represented by a scalar that is multiplied by the vector corresponding to that modality. This scalar is obtained by using the vector of modalities as input to a neuron with activation function *tanh*, the result of which is then passed through a sigmoid function to obtain the final value.

These methods were selected because of their performance in the works in which they appeared and for their capacity to easily adapt to the selected unimodal methods. For example, LSTM based methods were discarded because the output of the selected unimodal methods was not sequential, and as such, the strengths of these type of methods would not be put to good use. Additionally, these fusion models are representative of different approaches – i.e., MLP, Attention, and Probability based. Some of these, like Embracenet and Self-Attention had been successfully tested in the field of emotion recognition, while others like DeepFusion obtained good results in other areas and were selected based on the prospect of achieving similar results in the task of emotion recognition.

IV. EVALUATION FRAMEWORK

This section describes the framework proposed to compare different fusion methods, and its evaluation.

A. FRAMEWORK DESCRIPTION

Figure 1 shows an example of the proposed framework to perform the comparison of different fusion methods using three modalities. It is composed of the following steps:

1) Dataset selection: The dataset to be used must be selected according to the the study context.
2) Preprocessing: At this stage the data must be prepared, which may involve data cleaning, denoising, input normalization, sample selection, etc.
3) Individual modality training: The training process for each modality must be performed, normally based in different machine learning techniques. Then, with the data preprocessed in the previous stage as input, the data must be adjusted using different techniques and their performance must be evaluated.
4) Selection and testing of the fusion method: The testing stage consists on considering different fusion methods to evaluate. Since most fusion methods require an input for each modality, for each sample where the information from one modality is missing, default values are supplied for said modality.
5) Evaluation and comparison of results: Using different quantitative model evaluation metrics, i.e., precision, recall, F1 score, and accuracy, the results obtained by each fusion method are compared.

In the following section, we describe how the comparative evaluation of several fusion methods is performed based on the proposed framework.

B. THE COMPARATIVE EVALUATION PROCESS: STEP BY STEP

The first steps of our framework (dataset selection, preprocessing and individual model training) were considered in previous work developed in the context of multimodal emotion recognition [47]. This work proposed an architecture to process face, audio, and text modalities aggregated with a fusion method called Embracenet+. Each modality is processed individually and its result is an input for the fusion method. Finally, the fusion method produces a recognized emotion (i.e., happiness, neutral, sadness, and anger). Figure 2 shows the basic structure of this work, which serves as the basis for the evaluation of the different fusion methods. The main considerations on each step assumed in this work are summarized as follows.

1) DATASET SELECTION

In this work, we keep the dataset selected in the previous work [47]: the IEMOCAP dataset [54]. This dataset contains recordings from 10 actors interacting in pairs during five different sessions. Some interactions were scripted, others improvised. The dataset includes video and audio recordings from these interactions and their respective transcriptions. Approximately 12 hours of audiovisual data were obtained, including video, speech, motion capture of the face, and text transcriptions. Motion capture (MOCAP) data are also included, but they are not considered in the experiments described in this paper. There are 10039 samples, each containing data from different modalities.

Ten emotions considered during the samples' annotation process (i.e., neutral, happiness, sadness, anger, surprise, fear, disgust, frustration, excitement, and others). However, most research on this dataset considers only a group of four emotions (i.e., happiness/excitement, sadness, anger, neutral) [5]. All the models in this work are trained only with samples that have these labels, the only labels the authors of the dataset intended to use initially. Table 2 shows the distribution of the original ten emotions of the IEMOCAP dataset. Table 3 synthesizes the number of samples in each of the four categories that were classified by at least one of the models used for the individual modalities. In the original dataset, 2507 samples (24.97% of the total) were unlabeled. The first four sessions of IEMOCAP were used for training, while the fifth one was used for testing.

2) PREPROCESSING

This stage was also covered in the previous work [47]. For image, the face region was recognized and processed separately. Audio was processed in two ways: an MFCC features extractor was used to obtain a graphic representation of

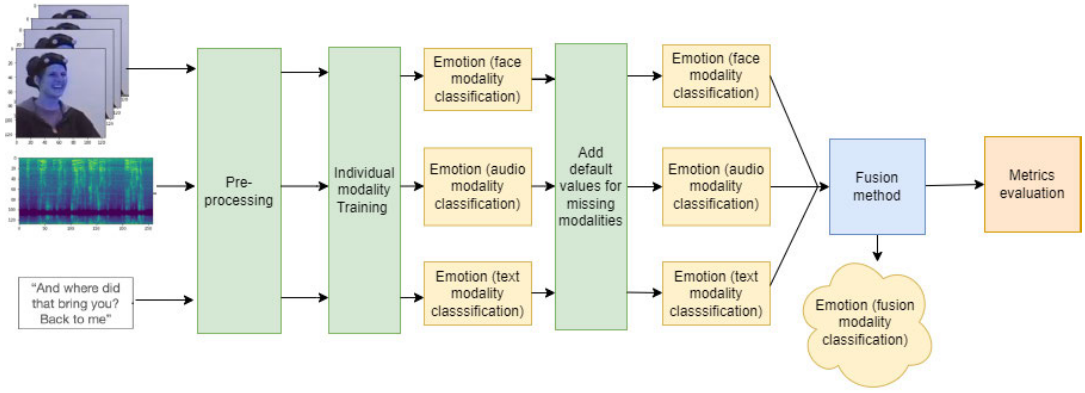


FIGURE 1. Framework to compare different fusion methods applied to three different modalities (face, audio, text).

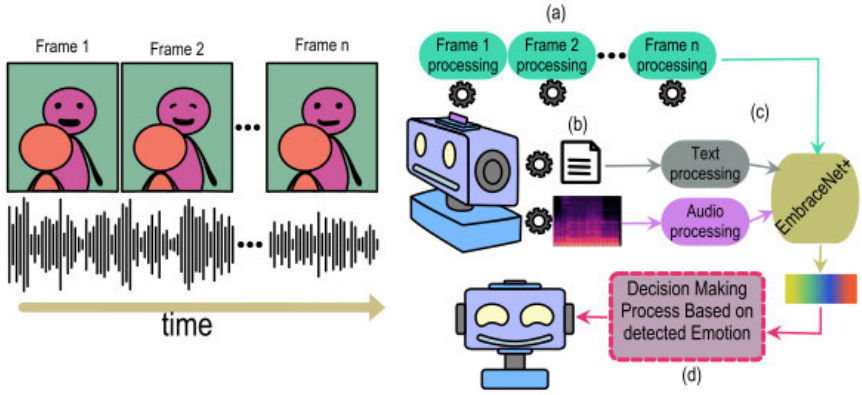


FIGURE 2. Architecture used for the multimodal emotion recognition [47].

TABLE 2. Number of samples from each category in the original IEMOCAP dataset.

	Neutral	Happiness	Sadness	Anger	Surprise	Fear	Disgust	Frustration	Excitement	Other	Total (labeled)
Number of samples	1708	595	1084	1103	107	40	2	1849	1041	3	7532
%	17,01	5,93	10,80	10,99	1,06	0,40	0,02	18,42	10,37	0,03	75,03

TABLE 3. Number of samples from each category that were classified by at least one of the individual models.

Set	Happiness/Excitement	Neutral	Sadness	Anger	Total
Training	1194	1324	839	993	4290
Test	442	384	245	170	1241
Total	1636	1708	1084	1103	5531

speech, and a conversion into text (transcription of speech). Text was analyzed with NLP techniques.

3) INDIVIDUAL MODALITY TRAINING

For evaluation of the fusion models, we have kept the same individual emotion recognition models used in the work presented in [47], which were trained and validated in IEMOCAP dataset:

- Face modality: For this modality, a variation of the model proposed by Parkhi et al. in [55] was used. A VGG19 was considered instead of the VGG16 used by Parkhi and the image sizes were reduced to 48×48. This latter helped also to reduce the processing time and the numbers of parameters in the last layers. Aside from the VGG19, which outputs a 512 feature vector, a linear layer was used to make the classification. For this network, in order to work with the video samples from IEMOCAP the samples were pre-processed as follows: 8 frames that contained a face where extracted from each video sample and were individually processed by the network. The final result was the average of

the classification vectors for each of the 8 frames. The accuracy of this modality was 0.44. But since face information was only available for about 41% of the test dataset, this modality only correctly recognized about 18% of the test dataset.

- Audio modality: To process audio, a CNN-based approach based on the work described in [56] was used, in which the voice audios were represented as Mel Frequency Cepstral Coefficients (MFCC) and processed as images by the CNN. The audio and video models were trained using Adam Optimizer, with a learning rate of 0.001 and the cross-entropy function as the loss function. The face model was trained with 22 epochs, and the audio model with 60. The difference in the number of epochs was because the face model was overfitting. In the case of the audio model, a scheduler was also used, which reduced the learning rate to 0.000001. The accuracy reached by this modality was 0.58.
- Text modality: The model selected to process the text obtained from the audio transcription was based on the DialogXL, proposed in [57], which has obtained state of the art results when compared to other text processing models [58]. The output of the DialogXL was then passed to a neural network where each layer contained dialog-aware self-attention and utterance recurrence components. A dense neural network (DNN) was used as a last layer for classification purposes. Since the pre-trained version of DialogXL was designed to recognize six emotions and the rest of the models used worked with four, some adaptations were necessary. Emotion and happiness were merged and samples labeled with fear were omitted from training and subsequent fusion. The accuracy of this modality was 0.84. But since text information was only available for about 80% of the test dataset, this modality only correctly recognized about 68% of the test dataset.

As for the output of each individual network, by taking into account a given sample, the outputs of both the audio and the face networks have the same format – i.e., a probability distribution for the category to which the sample belongs to (in this case, the categories are the four emotions considered). Meanwhile, the output for this sample with the text network is a vector of ones and zeros, with a value of one in the position representing the predicting category and zeros in the rest. When one of the individual models cannot make a prediction, for fusion purposes the output is replaced by a vector of zeros. Afterwards, the output of each individual modality is passed as input to the Embracenet+ method for the fusion, which finally returns as output the recognized emotion [47].

4) SELECTION AND TESTING OF THE FUSION METHODS

The selection included fusion methods of the state-of-the-art that showed good performance in each of their domains; these were previously described in Section III.

All the models were trained with Adam Optimizer using a learning rate of 0.001 and a batch size of 32. Cross-entropy was selected as the loss function for most of the models, with the exception of the DeepFusion model. As specified in [16], the function selected for this one is the sum of the cross-entropy loss for the whole network and the weighted sum of the two individual components' entropy loss (both Weighted Combination and Correlation modules). For each configuration of each method, the training process was repeated from scratch 15 times.

There is an imbalance in the data, even after merging the excitement and happiness and removing some of the less featured categories. The training was done by assigning a weight to each label and applying a weighted version of the respective loss function to mitigate the effects of imbalance in order to obtain more accurate results.

For the networks that involved MLPs (MLP, attention, and Tensor Fusion models), the considered hyperparameters were the number of layers and the number of neurons per layer. Different combinations of these hyperparameters were tested to determine which was a better fit for our problem. For models like DeepFusion, the number of layers was the same as the one specified in its original design [16]. Additionally, some steps from the final phase of DeepFusion were omitted since they seek to reduce the dimensionality of the vectors produced by its submodules and equate the size of the vectors; both modules already produce vectors of 1×4. The only hyperparameters that were varied and tested for DeepFusion were the weights of the submodules' loss functions in the final weighted sum (which acts as the loss function).

Finally, all trained models were evaluated on versions of the dataset in which one of the modalities is missing. This was done to study how much each modality influences the final results.

For each model, we relied on accuracy and F1-score metrics (the weighted average of the score for all categories) to evaluate the methods' quality and capabilities.

The final step of our framework is described in the next section.

V. EVALUATION AND COMPARISON OF RESULTS

This section describes the results obtained in the final step of the comparative evaluation of the nine fusion methods selected. The results correspond to tests on two scenarios: (1) considering all available modalities and (2) considering that one modality is always missing. In both, the accuracy and F1 performance metrics are reported.

A. USING ALL AVAILABLE MODALITIES

Table 4 shows the accuracy measure and F1 score for the best configuration of hyperparameters for each fusion method using weighted samples. Similarly, Table 5 shows the accuracy measure and F1 score for the best configuration of hyperparameters for each fusion method using unweighted samples.

TABLE 4. Accuracy and weighted F1 values for the best hyperparameter configuration for each model according to these metrics (when using weighted objective function).

Method	Avg. Acc.	Min Acc.	Max. Acc.	Var. Acc.	Avg. F1	Min. F1	Max. F1	Var. F1
DeepFusion	0.78909	0.78163	0.79694	0.00002	0.78965	0.78228	0,79759	0,00002
Weighted Combination	0.78888	0.76068	0.79371	0.00006	0.78981	0.75821	0.79467	0.00007
Self-Atenttion	0.78711	0.78646	0.78807	4E-07	0.78810	0.78743	0.78906	3E-07
Embracenet+	0.78426	0.78082	0.78888	4E-06	0.78545	0.77978	0.79001	7E-06
Embracenet	0.76906	0.74617	0.78566	0.00009	0.76998	0.73168	0.77908	0.00012
Cross Modality	0.72259	0.71878	0.74214	0.00006	0.73071	0.72835	0.74342	0.00002
Tensor Fusion	0.69530	0.50846	0.78243	0.00855	0.65609	0.37156	0.78366	0.01921
DNN/MLP	0.64835	0.45447	0.78727	0.01461	0.58778	0.30821	0.78804	0.03221
Attention MLP	0.59635	0.35616	0.77357	0.01477	0.51852	0.18708	0.77551	0.02881

TABLE 5. Accuracy and weighted F1 values for the best hyperparameter configuration for each model according to these metrics (when using unweighted objective function).

Method	Avg. Acc.	Min Acc.	Max. Acc.	Var. Acc.	Avg. F1	Min. F1	Max. F1	Var. F1
Self-Atenttion	0.78936	0.78807	0.79049	3E-07	0.79093	0.78963	0.79214	3E-07
DeepFusion	0.78926	0.78324	0.79452	0.00001	0.790147	0.78441	0.79522	0.00001
Weighted Combination	0.78920	0.78646	0.79130	2E-06	0.79020	0.78719	0.79243	2E-06
Embracenet+	0.78571	0.78163	0.78888	5E-06	0.78707	0.78005	0.79447	0.00001
Embracenet	0.76820	0.75584	0.78082	0.00006	0.76864	0.74364	0.78244	0.00010
Cross Modality	0.73092	0.67123	0.74456	0.00037	0.73164	0.62385	0.74752	0.00090
Tensor Fusion	0.67827	0.53344	0.78163	0.00820	0.62643	0.43276	0.78337	0.01859
DNN/MLP	0.63196	0.35616	0.78163	0.01485	0.55846	0.18708	0.78225	0.03136
Attention MLP	0.59500	0.35616	0.75020	0.00983	0.51361	0.18708	0.75272	0.020971

Regarding the effect of weighted loss for training, roughly half of the methods do not report any gains using this technique (Weighted Combination, Self-Attention, Embracenet+, Cross-Modality, and DeepFusion) either in their accuracy or their F1-score; they are slightly better with unweighted objective functions. In fact, three of the four methods that do present improvements with weighted objective functions, rely heavily on MLPs (MLP Simple, Attention MLP, and Tensor Fusion). As described in Section III, Tensor Fusion and Attention MLP use a lot of tensors or matrix multiplications which do not involve any kind of learnable parameters before actually passing the processed data to an MLP for classification, which means that for these methods, most of the learning is done by simple MLP used to fuse data. Among the MLP-dependent models, Tensor Fusion and MLP simple benefit the most, both in their accuracy and F1. These two methods improve 0.017 and 0.016 respectively, while Embracenet and Attention MLP only improve 0.009 and 0.013, respectively.

In the different attempts over training models with these configurations, these methods showed high variability and contrast between the best and worst case scenarios. In the worst case, some models performed similarly to the least accurate individual modality (in other words, the fusion failed to fulfil its objective). In the best case, the models offer results similar to those of the Embracenet. As mentioned before, there is a visible imbalance in the training data, and since these MLP-based models do not appear to possess mechanisms, like Weighted Combination or Embracenet, that purposely help them overcome this hurdle, this may be the reason behind these models' highly variable behaviour. The use of class weights (a technique used to account for imbalance in the dataset) mitigates the variability present in the results of these models. Most notably in the MLP, where the worst case scenario for accuracy improves around 10% and 12% for F1. In a different context, authors in [26] mention that their MLP classifier, trained on the probabilities for each class given by each modality (like the classifiers tested for this work), had high variability issues.

Furthermore, on average, Tensor Fusion and the MLP do not reach the quality of the results reported in [32] and [46], respectively. Most likely, these methods are not a good fit for the architecture used for the individual models. Both groups work with different individual models in their papers than those used here. Additionally, they do not "fuse" the results from each modality; they use intermediate results. That changes the input space to one with fewer local minima, in which the dataset imbalance may not represent the problem. Therefore, it may make it easier to train the network to achieve state-of-the-art results.

The confusion matrices for the validation set (once the training has finished) for representative instances of each method can be seen in Figure 3. From these, it is easy to confirm the problems that cause dataset imbalance in MLP-type models. The MLP (Figure 3a) and Tensor Fusion (Figure 3d) never assign the categories sadness and anger, respectively, which are the least present categories in both the training and validation datasets. Even more worrying is the case of Attention MLP (Figure 3e), which never classifies a sample as belonging to the neutral category, despite this being the most

common in the training. This, together with the poor results of accuracy and F1, makes this method not at all advisable, at least for the problem of emotion recognition with these modalities and conditions (i.e., unbalanced dataset, number of classes, etc.).

As for the other methods, a pattern can be observed in which happiness is often confused with neutral and to a lesser extent neutral with happiness and sadness. Regarding the first two cases (happiness and neutral) it is normal that most of the confusions involve these two categories, since they have the greatest presence in the dataset and therefore, a percentage of error comparable to those of the other categories, translates into a greater number of errors in absolute terms. As for the confusion of neutral with sadness, it is possible that this is due to the fact that sadness is the category with the lowest presence in the training set, and therefore, the methods may not be able to adjust their weights to correctly discriminate certain edge cases.

According to the values in Table 5 (results with unweighted objective function), Self-Attention is the best method according to both accuracy and F1, while according to Table 4 (results with weighted objective function), DeepFusion and Weighted Combination achieve the best results in accuracy and F1 respectively, even though the difference between all these methods in all these metrics is really small. This particularly tiny difference between the values of Weighted Combination and DeepFusion calls into question whether the contribution Cross-Modality offers to DeepFusion is really worth the addtional computations, specially in terms of time, as it offers only a minuscule improvement over using Weighted Combination, and only in certain metrics. This goes to show that the key factor behind DeepFusion's results is Weighted Combination.

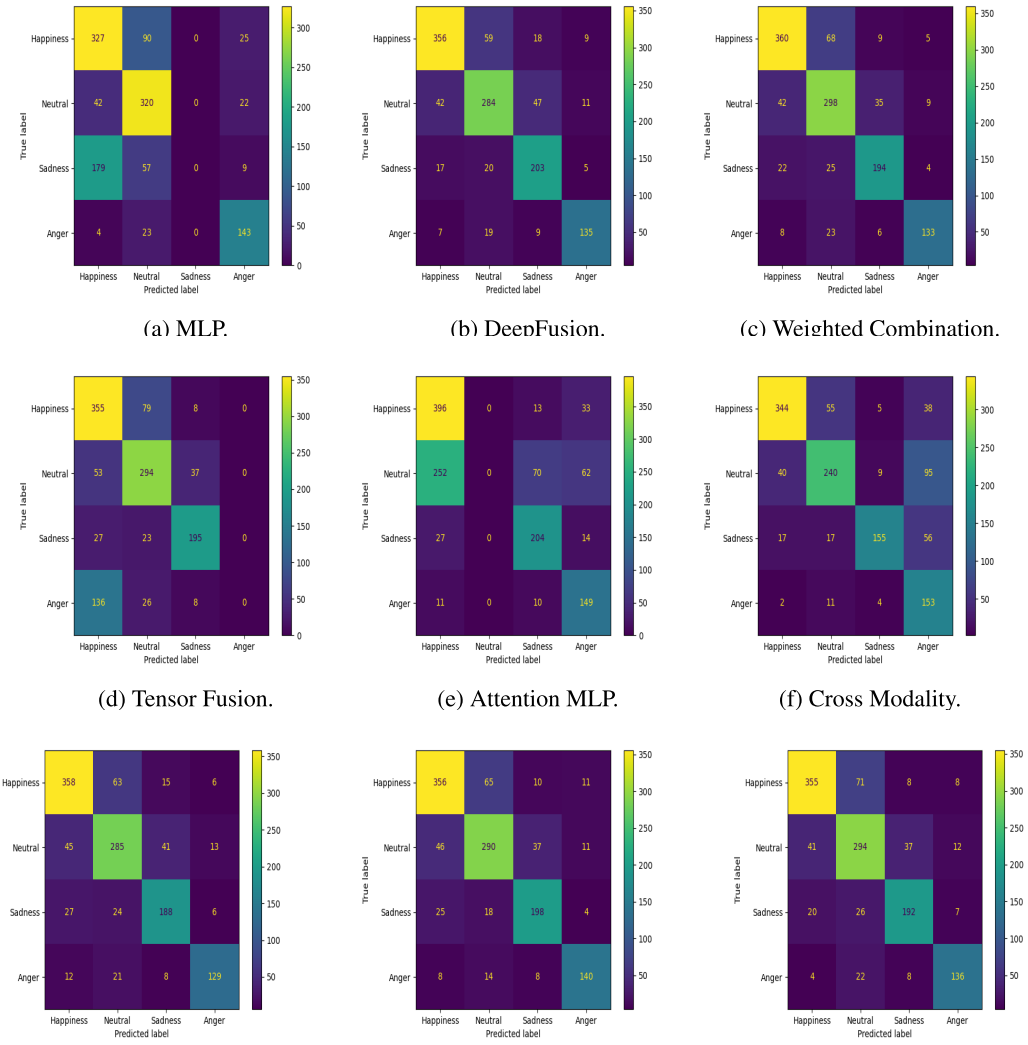
The success of these methods might be because of the quality factors or weights used by these type of methods. The ideas behind Self-Attention and Weighted Combination are very similar after all, as explained in Section III (Weighted Combination is inspired in Attention methods [16]). Given the simplicity of the process used for determining these values, it is easy to offer a possible interpretation of how these method work: the neuron used in a modality learns the emotions predicted by such modality with more precision and outputs a greater weight depending both on whether the probability assigned to the sample belongs to a specific category (emotion) is high and how precise the model is at recognizing the emotion. Even though these type of methods could be grouped under the ''Attention'' label, in this work they will be referred as ''quality factor'' methods to avoid confusions with Attention MLP, which obtained very different results. A more detailed view of a quality factor method (specifically weighted combination) is shown in Figure 4, in which the weight assigned to the text modality is vital for the GRU's input vector to favor the correct emotion heavily. While this appears to be the main idea behind how this mechanism works, a potential problem for this use case might be that the text modality expresses its results as a vector of zeros and an one, and the others as a more elaborated probability distribution, which may cause some bias in the model. Although the Weighted Combination method improves the accuracy of the text model by a decent margin (around 10% when accounting for the complete test dataset), if we observe the test dataset and one of the obtained models (one with an over-the-average accuracy), we can deduce that this difference comes from the samples where the text modality is missing. There is not a single case where the other modalities manage to ''correct'' a wrong prediction from the text modality. This is not to say that this method cannot use the results from a less accurate modality to correct a more accurate one, which will be expanded in the following Section.

B. USING ONLY SOME MODALITIES

To study the behaviour of these models when one modality is missing and the importance of each modality for the final classifier, the same models previously trained were also tested on versions of the test dataset that purposely omitted one of the modalities. Table 6 shows the average accuracy and the F1 score for each of these tests with weighted samples, while Table 7 shows the average accuracy and the F1 score for their corresponding unweighted samples. The performance for all models is similar when omitting the audio or face modalities (w/o audio and w/o face, respectively in Table 6 and Table 7) as when using all three available modalities; in many cases, omitting one of the modalities results in more accurate or greater F1 values. That suggests that these modalities contribute roughly the same to the samples where text is missing or fails, even though the accuracy of the individual audio model was notably better than the face model.

Even when comparing the accuracy of these methods only in the samples where the text modality is absent, audio accuracy is even greater than face. The explanation of this comes of fact that when considering only the face and text modalities, between both of them they only cover approximately 88% of the test dataset, whereas just the audio modality has data for every sample in the dataset. This suggests that the overall *accuracy* over the whole dataset, when omitting audio may be a little lower, for every method. Hence, when just considering the samples where either text or face data exist, the accuracy of the text model is 0.75; thus, the models that surpass this threshold can be considered successful, as they improved the performance of the best single modality available.

But when comparing these results to those obtained by the different methods when the text modality is absent, a significant difference is noticed: all metrics for all the tested techniques go down significantly. Most significantly, there is a consistent sense of proportionality: the methods with superior results in the other scenarios are still the ones with superior results in this case, with one marked exception, the Cross-Modality method. As seen in Table 6, its accuracy dropped 56%, which is not only the sharpest drop seen in these experiments but also, when comparing Table 4 and Table 5 with Table 6 and Table 7, respectively, one can notice



(a) MLP. (b) DeepFusion. (c) Weighted Combination.

(d) Tensor Fusion. (e) Attention MLP. (f) Cross Modality.

(g) Embracenet. (h) Embracenet+. (i) Self-Attention.

FIGURE 3. Confusion matrices for the validation set for a representative instance of the best configuration for each method.

TABLE 6. Accuracy and F1 results for the best configurations of hyperparameters for each method (omitting a modality in all samples) (applying weighted samples during training).

Method	w/o face Acc.	w/o audio Acc.	w/o text Acc.	w/o face Acc.	w/o audio Acc.	w/o text Acc.
Self-Attention	0.78899	0.79019	0.53736	0.78966	0.79319	0.52209
Embracenet+	0.78598	0.79177	0.56084	0.78637	0.79368	0.55355
Weighted Combination	0.78550	0.79703	0.57303	0.78647	0.79865	0.55963
DeepFusion	0.78689	0.79533	0.56326	0.78716	0.79759	0.55503
Embracenet	0.77561	0.79158	0.56422	0.77625	0.79184	0.56232
Cross Modality	0.72291	0.77899	0.16497	0.73119	0.78216	0.05586
Tensor Fusion	0.70191	0.70445	0.46613	0.66230	0.66813	0.41835
DNN/MLP	0.65222	0.66194	0.43535	0.59083	0.60468	0.35440
Attention MLP	0.59731	0.61387	0.37207	0.51955	0.53791	0.25541

that Cross-Modality goes from being the sixth best method (according to accuracy) when all modalities available are used, to being the worst one when text is missing. It appears that the Cross-Modality method relied excessively on the text modality to make the predictions and it cannot compensate for its absence. It does not appear to be learning anything valuable from the other modalities. Its results, in terms of accuracy and F1 score, are significantly worse than the

FIGURE 4. Representation of the process followed by the Weighted Combination model to obtain the quality weights (factors) for each modality's sample.

TABLE 7. Accuracy and F1 results for the best configurations of hyperparameters for each method (omitting a modality in all samples) (without applying weighted samples during training).

Method	w/o face Acc.	w/o audio Acc.	w/o text Acc.	w/o face Acc.	w/o audio Acc.	w/o text Acc.
Self-Attention	0.78732	0.78936	0.52737	0.78875	0.79233	0.50443
Embracenet+	0.78619	0.78856	0.56035	0.78700	0.79075	0.55083
Weighted Combination	0.78555	0.79776	0.59737	0.78650	0.79949	0.59674
DeepFusion	0.79012	0.79740	0.57857	0.79068	0.79938	0.57463
Embracenet	0.77497	0.79055	0.57250	0.77574	0.79090	0.57208
Cross Modality	0.73022	0.78002	0.26140	0.73154	0.77979	0.13309
Tensor Fusion	0.68509	0.68483	0.47311	0.63251	0.63590	0.42074
DNN/MLP	0.63546	0.64493	0.43975	0.56170	0.57585	0.35951
Attention MLP	0.59425	0.60793	0.41445	0.51315	0.52540	0.32886

worst-performing individual and fusion models (with its F1 almost reaching 0). Looking at Table 6, it is evident that this issue is limited to Cross-Modality when the text modality is unavailable. Whether this happens because of the different nature of the data input from the text when compared against the others, because of the imbalance in the dataset or simply because the text is more accurate than the others is an interesting question that can be answered by testing this method with other datasets. If this method develops such a notable bias towards the more accurate modality, it might not be suitable for in-the-wild contexts, which is the ultimate goal of this type of method and where different modalities' data are frequently lost.

Despite the behavior of Cross-Modality, DeepFusion's results do not decrease as much, and approximately in the same proportion as the other methods which achieve similar levels of accuracy and F1 when using all modalities. Two explanations are possible: Cross-Modality does not significantly contribute to DeepFusion's results (as suggested in Section V-A), or Cross-Modality, without text, is so biased that DeepFusion may notice said biases and try to account for them. If DeepFusion does that, it may be able to get more useful information from Cross-Modality than its low accuracy reflects. A fact that supports this theory is that when weights are not used during training, the best Cross-Modality configuration when text is absent, is one where Cross-Modality has more influence on the loss function, but when weights are used or other modality (or none) are omitted, the best configurations are those where Weighted Combination has more influence, or both have equal contribution to the loss function.

When removing the text modality from the test dataset, independently of the method used, the precision for the categories sadness and anger dropped considerably, yet its recall value did not lower as much. The opposite happened with happiness. Overall, this could be explained by an increase in the number of false positives for the sadness and anger categories while maintaining a similar number of true positives and a reduction of true positives while keeping a comparable quantity of false positives for happiness, with the correct guesses still outnumbering the wrong ones for said category. It appears that the information provided by the text modality is fundamental to correctly classifing certain samples belonging to the happiness category, as the number of predictions made for that category diminished significantly and went (mostly incorrectly) to other categories (frequently

TABLE 8. Precision, recall, and F1 for each category with and without text for a representative instance of weighted combination.

Emotion	Precision	Recall	F1
Happiness	0.83	0.81	0.82
Happiness (no text)	0.78	0.45	0.57
Neutral	0.72	0.78	0.75
Neutral (no text)	0.54	0.62	0.58
Sadness	0.80	0.79	0.79
Sadness (no text)	0.55	0.75	0.63
Anger	0.88	0.78	0.83
Anger (no text)	0.55	0.69	0.61

TABLE 9. Precision, recall, and F1 for each category with and without text for a representative instance of embracenet.

Emotion	Precision	Recall	F1
Happiness	0.82	0.79	0.81
Happiness (no text)	0.71	0.50	0.59
Neutral	0.72	0.76	0.74
Neutral (no text)	0.54	0.54	0.54
Sadness	0.77	0.78	0.78
Sadness (no text)	0.53	0.72	0.61
Anger	0.82	0.76	0.79
Anger (no text)	0.54	0.67	0.60

TABLE 10. Precision, recall, and F1 for each category with and without text for a representative instance of DeepFusion.

Emotion	Precision	Recall	F1
Happiness	0.84	0.81	0.82
Happiness (no text)	0.77	0.41	0.54
Neutral	0.74	0.74	0.74
Neutral (no text)	0.54	0.58	0.56
Sadness	0.73	0.83	0.78
Sadness (no text)	0.52	0.76	0.62
Anger	0.84	0.79	0.82
Anger (no text)	0.51	0.72	0.60

TABLE 11. Precision, recall, and F1 for each category with and without text for a representative instance of self-attention.

Emotion	Precision	Recall	F1
Happiness	0.85	0.80	0.82
Happiness (no text)	0.51	0.58	0.54
Neutral	0.71	0.77	0.74
Neutral (no text)	0.51	0.58	0.54
Sadness	0.75	0.84	0.79
Sadness (no text)	0.53	0.74	0.62
Anger	0.83	0.80	0.82
Anger (no text)	0.44	0.76	0.56

to sadness and anger, which would explain their precision results). Table 8, Table 9, Table 10, and Table 11 show the precision, recall, and F1 for instances of various methods, with and without using the text modality.

As for other aspects of the adaptability of the methods, Weighted Combination and Embracenet are the most successful. For every combination of modalities tested, these models succeed at improving the results of the best modality available or at least equaling its performance: when the text modality is present, they manage to get superior results both in accuracy and F1. Consequently, audio is the most reliable modality, Embracenet gets similar results, and Weighted Combination significantly improves it. Even if this last part does not seem like much, it could be inferred from these results that during their training, these networks, learn a sort of "quality ranking," which allows them to determine which of the available modalities present in a sample would be more informative to the final classification.

To better understand how this 'quality ranking' mechanism works, the process of classifying a sample with a Weighted Combination is analyzed at in more detail. Weighted Combination process is represented in Figure 5 (where the process of calculating the quality factors is shown), and Figure 6 (where the process of classifying the samples using the quality factors is shown) using an authentic sample from the test dataset, with the text modality removed. From the weights in the neurons used to determine each modality's quality factor, it is pretty difficult for the face modality to produce values that contribute significantly to the final vector used by the GRU to make a classification, unless the face modality predicts a category with a large percentage of certainty. These neurons' weights serve an additional purpose: their value appears to maintain proportionality concerning the quality of the modality when compared to the others. This seems to be what allows the model to determine which modality should have more influence on the vector that ends up being the GRU's input. For example, the least reliable modality, face, is the only one with negative weights in its neuron, and in consequence, the quality factor it produces will be very low. And even though, in this instance, the weights from the audio neuron are heavier than those of the text neuron, due to the different nature of the format of the text and audio vectors, the text modality remains most influential. As can be seen in Figure 7, despite the quality factor for audio (0.3962) being greater than the one for text (0.3953), the value for the correct category (sadness, third element of the vector) for the text modality, after applying the quality factors is 0.3952, while for audio is 0.1843 and for text is 0.070. Since these values are simply added to determine the value of the category in the GRU's input vector, it is obvious that text is the modality that contributes the most to the correct category.

Figure 5 and Figure 6 also show that these quality weights are not the only piece responsible for the accurate classification. There exist certain instances, like the one shown, in which the GRU's input vector seems to favor a particular emotion, but the final output goes for another. After reviewing how much these events affect the result, at least for the studied instance of Weighted Combination and always omitting the text modality, it appears that there are 168 samples where all the available modalities signal one category, but the final classification goes for another. Around 59.5% of these changes were indeed appropriate corrections, which positively impacts the results. From the 168 instances in which the modalities and the final classifier disagree, in 94% of the cases the GRU classified the sample as belonging to the neutral category, and the other six distributed between anger and sadness. Despite happiness being the second most prevalent emotion, this change that the GRU makes never outputs happiness. That shows that this bias only occurs

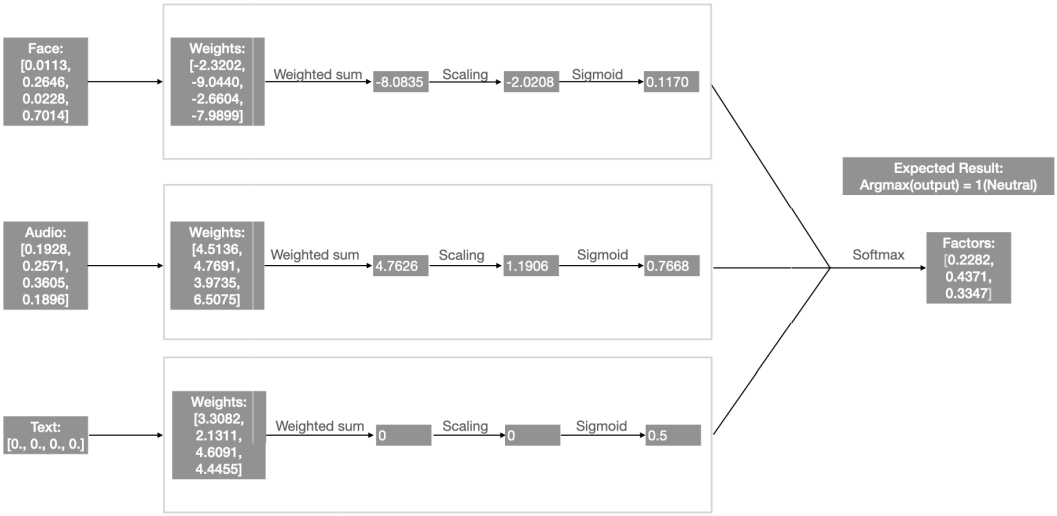


FIGURE 5. Process of the DeepFusion model to obtain the quality weights for each of the sample's modalities without text modality.

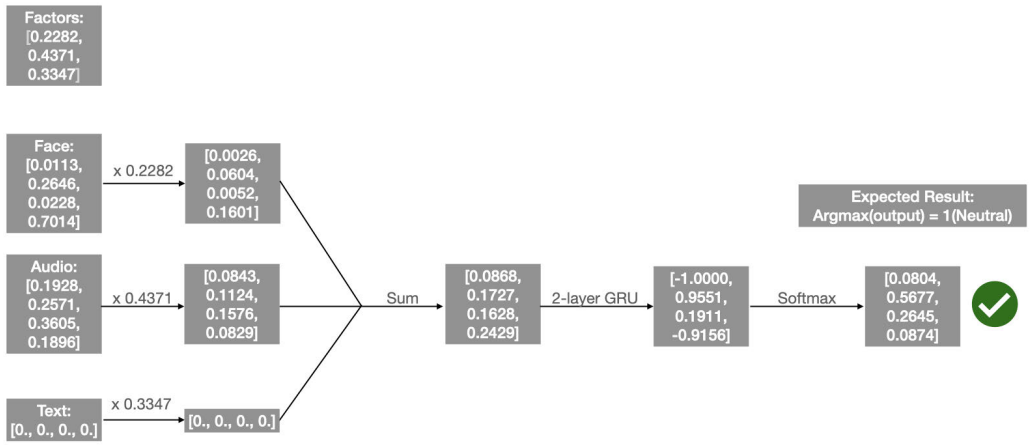


FIGURE 6. Classification process of the DeepFusion model once these weights are obtained without text modality.

towards the most common categories and is not proportional to their number. Most likely, when the GRU's input vector does not appear to favor a category in particular, the GRU has learned to go for the most common category from the group of categories with higher probabilities. It is possible that in a more balanced dataset, this does not happen, or maybe the number of changes to the wrong category decreases, which would be ideal.

C. VISUALIZATION OF THE TRAINING PROCESS HISTORY

This section visualizes the history of the behavior of each model during the learning process by means of their accuracy and loss curves.

Figure 8 shows the loss curve for each fusion method. Methods such as MLP (Figure 8a) and Attention MLP (Figure 8e) can get stuck in local minima for prolonged periods. In contrast, others methods like Weighted Combination (Figure 8c), Self-Attention (Figure 8i), and Cross-Modality (Figure 8f) have smoother descents (although Cross-Modality appears to have a small overfitting problem). Tensor Fusion (Figure 8d) and DeepFusion (Figure 8b) did overfit very clearly and probably could have used fewer epochs. Finally, the Embracenet (Figure 8g) and Embracenet+ (Figure 8h) show many peaks, even though it was more or less stable after a few iterations.

Similarly, Figure 9 shows the accuracy curve for each fusion method, which reaffirms the discussed behavior for loss curves in Figure 8. In terms of validation accuracy, the Tensor Fusion method (Figure 9d) has a very poor and chaotic behavior, with more peaks than the loss function (Figure 8d), which implies that the method is very sensible, where small changes in the loss value cause bigger changes in the accuracy. This contrasts with DeepFusion (Figure 9b), where what appeared to be a notable case of overfitting according to the loss value (Figure 8b), is not as significant taking in account the accuracy; which indicates a more stable model in terms of accuracy, in opposite of Tensor Fusion. Even though, Self-Attention (Figure 9i) and Weighted Combination (Figure 9c) have the better performance, the model Attention MLP

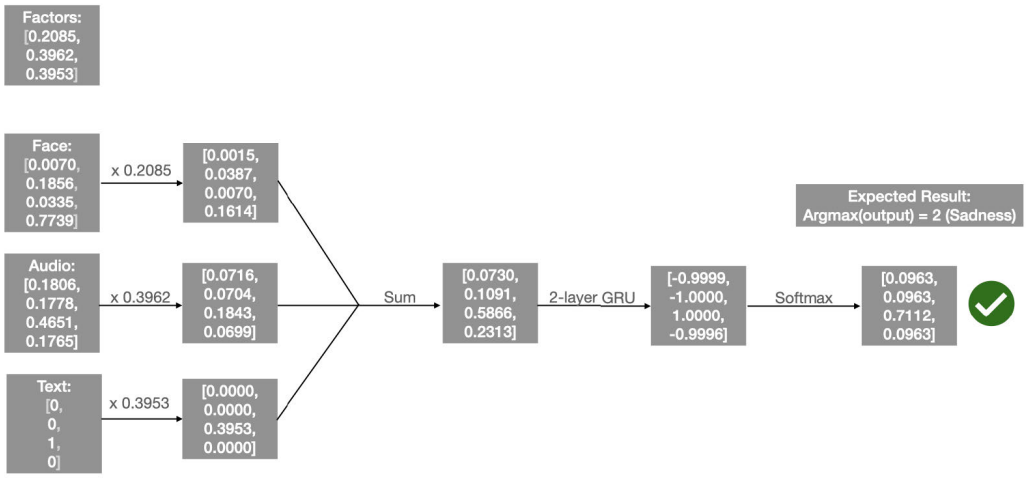


FIGURE 7. Representation of the process followed by Weighted Combination model to classify a sample.

(Figure 9e) has the smallest gap between accuracy curves. This is not as good as it seems, as it can be seen that when compared to the rest of the plots in Figure 9, it is the only one that has a training accuracy below 0.75. It can also be seen, that the peaks shown both in the loss and accuracy curves for Embracenet (Figure 8g and Figure 8g) are more pronounced that those of the Embracenet+ (Figure 8h and Figure 9h), where the peaks in accuracy are smaller than those of its loss, which signals a more stable model. This fact, along with the variance and accuracy results seen in Table 4 and Table 5 proof that the Embracenet+ not only achieved the goal of improving the accuracy (specially when all modalities are present) and reduce variance issues, but also created a more stable method.

D. AVERAGE EVALUATION TIMES

The evaluation time is measured from the moment where the fusion method being evaluated receives all inputs until it produces the classification. By measuring time this way, all the processing related to the individual modalities and the wait for the slowest of the unimodal models to finish occur before the timer starts running, so the time spent in this processes does not factor into the time measuring. Additionally, since in this study all the methods share the same unimodal architectures, the time spent in this stage should be, in average, the same for all of the fusion methods. The real difference is then in the time used in the fusion stage of the process.

The speed of MLP compared to the other methods is not surprising, it is a fairly straightforward mechanism and the configurations that record these times use very few neurons (two layers of seven neurons and one of four in the case of the first table and one of ten with another of four in the second). It does not perform any kind of matrix operations, like Tensor Fusion or Attention MLP, nor does it involve additional processing of each modality before classification, like Embracenet or methods involving the calculation of quality factors, since it concatenates the vectors it receives and starts directly to process them as if they were one.

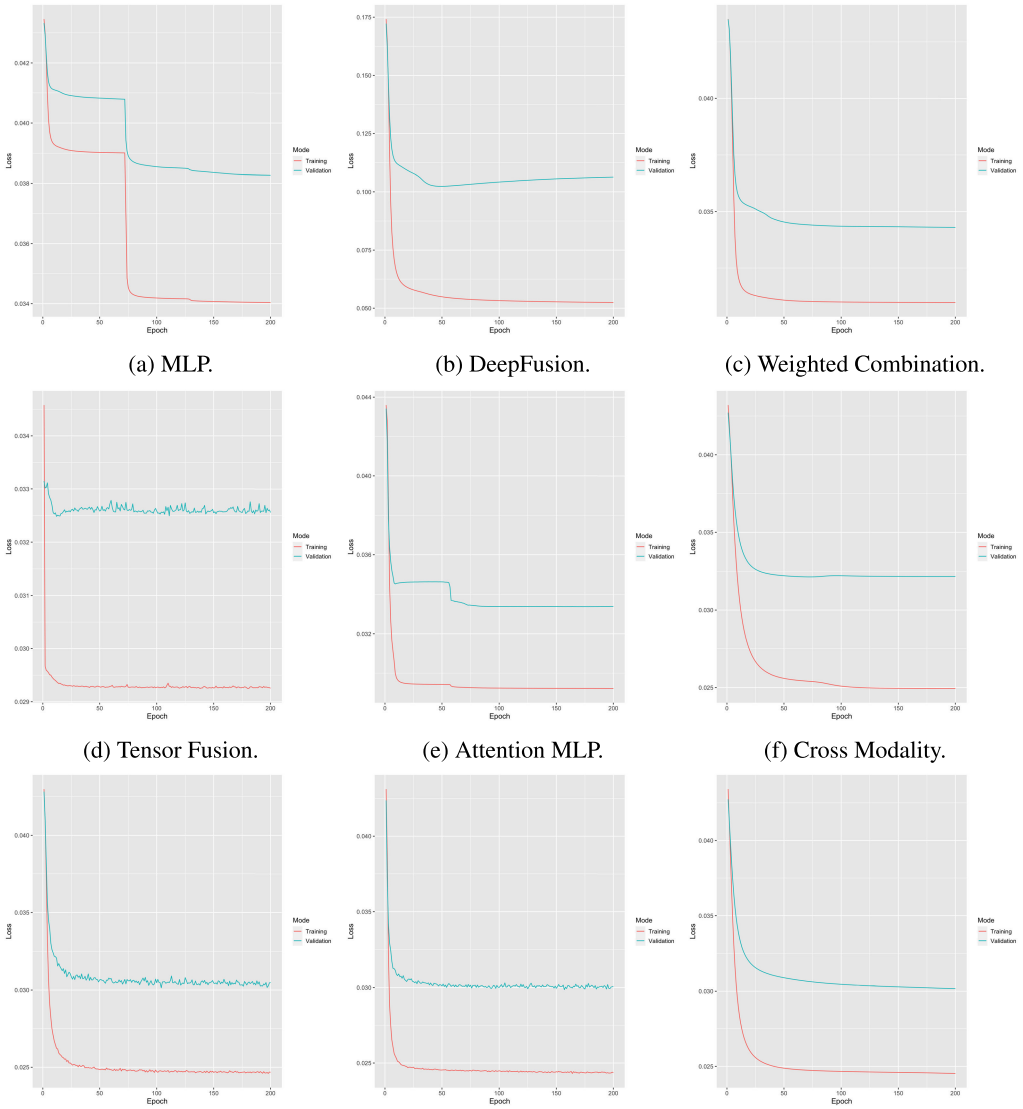
TABLE 12. Average prediction time (in seconds) for the best configuration of each model (weighted).

Method	Avg.	Min.	Max
DNN/ MLP	1.92334E-05	1.78814E-05	1.36137E-04
Self-Attention	3.61209E-05	3.48091E-05	2.16246E-04
Cross Modality	4.33195E-05	4.07696E-05	3.10183E-04
Tensor Fusion	5.00094E-05	4.88758E-05	8.70228E-05
Weighted Combination	8.2514E-05	7.98702E-05	4.32968E-04
Embracenet	1.2032E-04	6,69956E-05	3.39913E-03
DeepFusion	1.24300E-04	1.21832E-04	1.95980E-04
Attention MLP	1.45965E-04	1.40905E-04	4.48942E-04
Embracenet+	2.75565E-04	1.59025E-04	5.31817E-03

TABLE 13. Average prediction time (in seconds) for the best configuration of each model (unweighted).

Method	Avg.	Min.	Max
DNN/MLP	1.49454E-05	1.35899E-05	4.41074E-05
Self-Attention	3,60392E-05	3,48091E-05	6,8903E-05
Cross Modality	4,31005E-05	4,07696E-05	1,59979E-04
Tensor Fusion	5.00377E-05	4.88758E-05	9.01222E-05
Weighted Combination	8,21137E-05	7,98702E-05	1,50919E-04
Embracenet	1.19302E-04	6.67572E-05	3.94177E-03
DeepFusion	1.24233E-04	1.21832E-04	2.06232E-04
Attention MLP	1.55214E-04	1.50204E-04	2.99931E-04
Embracenet+	2.7557E-04	1.56879E-04	3.294945E-04

It is striking, among the quality factor methods, that Self-Attention is much faster than Weighted Combination despite being methods that share similar approaches and even similar processing. The most logical explanation is that the difference in time comes from the final classifier used by each: Weighted Combination uses a two-layer GRU network of four neurons, while Self-Attention opts for a single layer of four neurons. Additionally, there are certain additional steps performed by Weighted Combination, such as the Softmax layer to normalize the factors obtained. Other elements may play a role, such as the fact that one method concatenates the vectors of each modality and the other sums them. But, the Weighted Combination classifier is more conceptually



(a) MLP. (b) DeepFusion. (c) Weighted Combination.

(d) Tensor Fusion. (e) Attention MLP. (f) Cross Modality.

(g) Embracenet. (h) Embracenet+. (i) Self-Attention.

FIGURE 8. **Training and validation loss.**

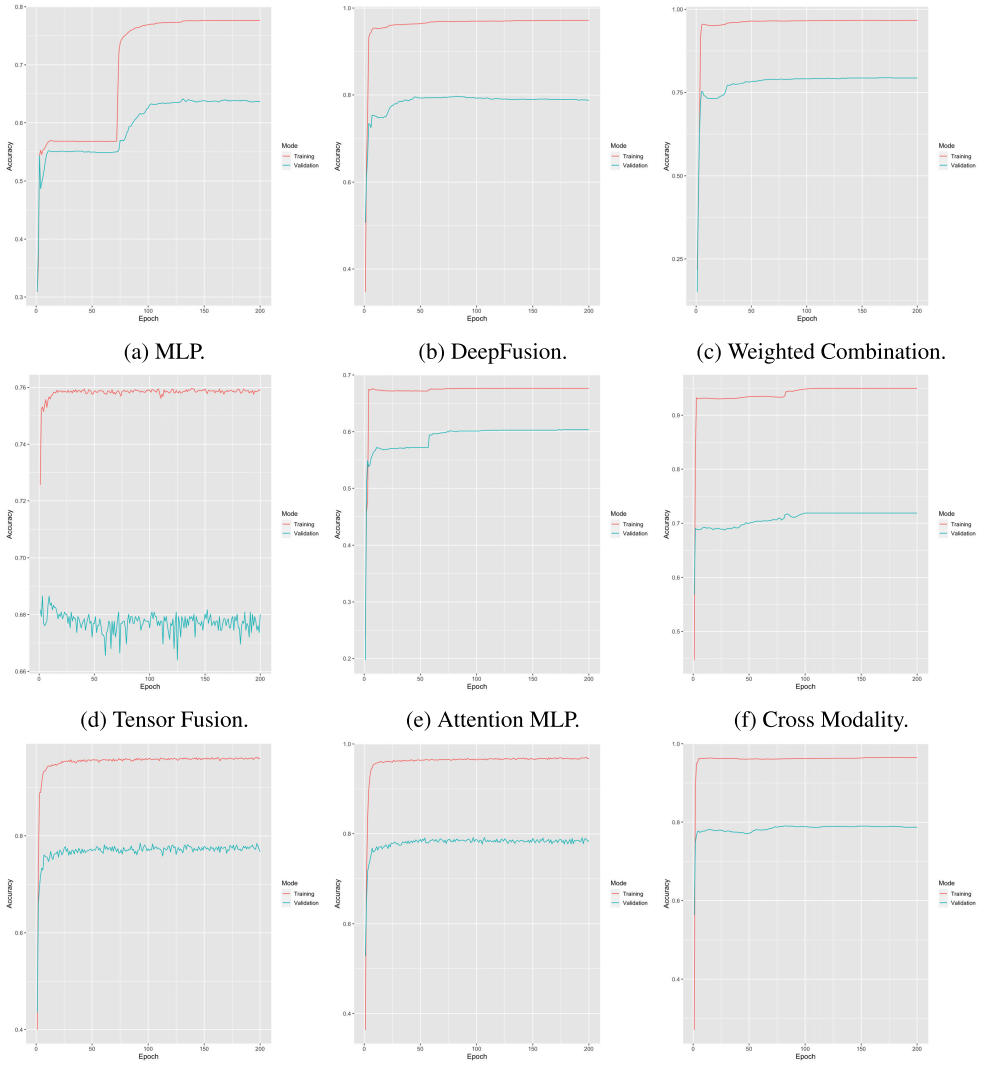
complex and uses twice as many neurons as Self-Attention (however few); which is the differentiating factor in this case.

One positive result is how little time Tensor Fusion takes, unlike Attention MLP, which is also a method that relies heavily on matrix multiplications and was one of the worst performers in this metric, as in all others. A more careful analysis of the performance of these two methods easily reveals why this difference: Tensor Fusion only computes one matrix, which it then multiplies by a single vector, while Attention MLP must compute three matrices, then transpose them and multiply each by a vector. These operations performed by Attention MLP are so slow, that even taking into account that the Tensor Fusion configuration has one MLP with 170 neurons as classifier, versus the four MLP used by Attention MLP that in total have about 47 neurons, Tensor Fusion is a much faster method. Of course, the speed of Tensor Fusion can only be highlighted given the small number of modalities involved and the reduced number of dimensions in the vectors of each of these, since the number of operations needed to perform these multiplications grows as a function of these two variables.

Obviously DeepFusion and Embracenet+, which are combinations of other methods (recall that Embracenet+ uses two Embracenets) take longer than their respective individual methods. Thus, these models with low performance are not suitable in contexts where response time is the key.

E. FINAL CONSIDERATIONS ABOUT THE RESULTS

After careful consideration of the different results presented throughout this section, it is quite evident that the quality factor methods (DeepFusion, Weighted Combination, Self-Attention) and the Embracenet can be considered successful in general terms, as they improve the results of the best individual modality (text). They can correctly classify most



(a) MLP. (b) DeepFusion. (c) Weighted Combination.

(d) Tensor Fusion. (e) Attention MLP. (f) Cross Modality.

(g) Embracenet. (h) Embracenet+. (i) Self-Attention.

FIGURE 9. Training and validation accuracy.

of the samples that are correctly classified when only that modality is used, but also many of the samples for which text information was unavailable. They also achieve this type of results consistently, which contrasts the performance of the MLP based models (MLP simple, Attention MLP, Tensor Fusion), which not only perform much worse in average, but also have high variability and stability issues, and also proved to be very sensible to imbalance problems in the dataset. Cross-Modality achieves respectable values in both accuracy and F1 under most circumstances, although slightly inferior to those of the quality factor and Embracenet and Embracenet+ methods, and most importantly, it shows a significant degree of dependency on the information provided by the text modality, which calls into question this method's ability to adapt to contexts where information may be lost frequently.

Embracenet and Embracenet+ methods were proposed to explicitly consider the problem of dealing with missing information from different modalities. Results shown that they actually achieve their objective (see Table 6 and Table 7). However, the quality factor methods, beside of tolerating missing modalities, are able to magnify or reduce the influence of each modality depending on how informative each of them is. Given that there are simpler mechanisms that achieve better results that the Embracenets, both when using every modality and only some of them, and the fact that these methods, specifically Embracenet+, proved to be particularly slow in terms of evaluation time (which also makes it inappropriate for in situation where hardware resources for processing are very limited), it appears that the quality factor methods are preferable.

Among the quality factor methods, although DeepFusion achieves the best accuracy in many of the scenarios tested, it uses Cross-Modality as one of its modules, and given this method's extreme dependency on a single modality, until further experiments are carried out to determine why this

happened, it will remain a possibility that the low performance of Cross-Modality when certain modalities are missing, at least in this context, could affect DeepFusion when using the method in similar circumstances. Additionally, when DeepFusion's results are better that those of Weighted Combination, they are only marginally better, which may not justify the additional computations (DeepFusion's evaluation time is approximately the double of Weighted Combination). Between Weighted Combination and Self-Attention, the choice is more complicated: both achieve some of the best results, and each of them has its merits. Self-Attention has a really low variance, which is a sign of stability, and its average evaluation time proves it is particularly fast when processing samples, specially when compared to Weighted Combination. But this last method shows a better tolerance to missing modalities, as it has better accuracy and F1 values both when audio or text are missing (the two most accurate modalities). Its results when text is missing are 1% better than those of Self-Attention, which is more than the usual difference in the metrics between these methods, and indicates a lesser degree of dependency on the most accurate modality.

The decision on which one is the better method is then tied to the context and set of circumstances in which one plans to implement the methods, since both proved to be more than adequate at the task of emotion recognition using face, audio, and text. Even in a specific field, like social robotics, one should consider additional details about the situation to make this decision: Weighted Combination may be more suited to robots deployed at big events with lots of noise and movement, like parties or stadium activities, like concerts or sports while Self-Attention could be more useful for museums, libraries or even daily house life, since in these contexts information may not go missing as frequently and in consequence, better response times may be preferable to better tolerance against missing modalities.

Since all models were trained and tested only on the IEMOCAP dataset, they were only tested on a single variant of the problem of emotion recognition (considering different datasets as different variants). This limits the scope of the conclusions and makes it more difficult to make generalizations on the behavior of these models for the problem of emotion recognition. Moreover, IEMOCAP data were collected under controlled conditions, and although this dataset has historically been very useful in this field of study [5], additional tests are needed on ''in the wild'' datasets to determine whether this methods manage to achieve a comparable level of success when the data distribution and other factors (like the percentage of missing data) more closely resemble those one would expect in application contexts in which multimodal fusion methods would be deployed (e.g., interacting with people in museums or hospitals).

Furthermore, as mentioned in Section III, the unimodal architectures selected to feed the fusion methods are not ideal with implementations based on RNN that fuses the data from the different modalities, as the models used for audio and text do not produce a sequential output. This situation prevented methods that heavily relied on recurrent networks from being featured in the comparisons made in this study, despite being one of the most researched approaches to this problem [34], [35].

VI. CONCLUSION

In this paper a strategy to make comparisons between different fusion methods is described. It specifically allows comparing different fusion methods in terms of performance metrics, such as accuracy, F1-score, average evaluation time, and behavior in the training process. Nine methods (MLP simple, Attention MLP, DeepFusion, Weighted Combination, Cross-Modality, Tensor Fusion, Embracenet, Embracenet+, and Self-Attention) are compared to determine which among them is the best in the context of the emotion recognition and their potential use cases. The same experimental conditions to test method are considered – i.e., the selection of the same dataset (IEMOCAP) and unimodal models (face, audio, and text).

In order to evaluate the tolerance of models to missing modalities, experiments in two scenarios are considered: using all available modalities and purposefully suppressing the information of one of the modalities. From the results, it was determined that the methods that use a quality factor (a form of attention), behave better, specifically Weighted Combination and Self-Attention, methods that frequently topped the metrics calculated in the different scenarios. The most suitable method depends on the circumstances in which they are used. Self-Attention remarkable stability and speed may be more adequate in certain contexts (like museums or small gatherings), but Weighted Combination results when modalities are missing (even the most accurate ones) may make it more useful in other contexts where missing modalities are more frequent (like concerts and sport events). Also, the quality factor methods internally use a ''quality ranking'' which allows them to determine which available modality is the most informative, and that MLP based methods are too sensible to common problems, such as dataset imbalance or missing modalities. Additionally, results show that the text modality is the most important for the final classification, as it is the only one whose absence causes a sharp drop in results.

As future work, we plan to test the studied fusion models on a more balanced dataset, to determine the extent to which the imbalance in the dataset affects their performance. Also, we will test them on emotion recognition ''in the wild'' datasets, as IEMOCAP's data was collected under controlled conditions in a laboratory. Thus, we will show the usefulness of deploying these methods in a non-controlled environment.

ACKNOWLEDGMENT

The authors would like to thank Prof. Ralph Grove, from Norfolk, VA, USA, who helped them with the spelling of this article.

REFERENCES

[1] M. Wiener, S. Devoe, S. Rubinow, and J. Geller, "Nonverbal behavior and nonverbal communication," *Psychol. Rev.*, vol. 79, no. 3, pp. 185–214, 1972.

[2] F. B. Mandal, "Nonverbal communication in humans," *J. Hum. Behav. Social Environ.*, vol. 24, no. 4, pp. 417–421, 2014.

[3] A. Pease and B. Pease, *The Definitive Book of Body Language*. New York, NY, USA: HarperCollins Publishers Australia, 2017.

[4] P. Furley and G. Schweizer, "Body language in sport," in *Handbook of Sport Psychology*. Hoboken, NJ, USA: Wiley, 2020, pp. 1201–1219.

[5] P. Koromilas and T. Giannakopoulos, "Deep multimodal emotion recognition on human speech: A review," *Appl. Sci.*, vol. 11, no. 17, p. 7962, Aug. 2021.

[6] E. Batbaatar, M. Li, and K. H. Ryu, "Semantic-emotion neural network for emotion recognition from text," *IEEE Access*, vol. 7, pp. 111866–111878, 2019.

[7] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decis. Support Syst.*, vol. 115, pp. 24–35, Nov. 2018.

[8] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: A review of BERT-based approaches," *Artif. Intell. Rev.*, vol. 54, no. 8, pp. 5789–5829, Dec. 2021.

[9] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[10] M. Sajjad and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020.

[11] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed. Signal Process. Control*, vol. 59, May 2020, Art. no. 101894.

[12] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognit. Lett.*, vol. 115, pp. 101–106, Nov. 2018.

[13] H. Zhang, A. Jolfaei, and M. Alazab, "A face emotion recognition method using convolutional neural network and image edge computing," *IEEE Access*, vol. 7, pp. 159081–159089, 2019.

[14] E. Pranav, S. Kamal, C. Satheesh Chandran, and M. H. Supriya, "Facial emotion recognition using deep convolutional neural network," in *Proc. 6th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2020, pp. 317–320.

[15] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L. P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 7216–7223.

[16] H. Xue, W. Jiang, C. Miao, Y. Yuan, F. Ma, X. Ma, Y. Wang, S. Yao, W. Xu, A. Zhang, and L. Su, "DeepFusion: A deep learning framework for the fusion of heterogeneous sensory data," in *Proc. 20th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, Jul. 2019, pp. 151–160.

[17] J.-H. Choi and J.-S. Lee, "EmbraceNet: A robust deep learning architecture for multimodal classification," *Inf. Fusion*, vol. 51, pp. 259–270, Nov. 2019.

[18] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Comput.*, vol. 32, no. 5, pp. 829–864, May 2020.

[19] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 439–448.

[20] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "CentralNet: A multilayer approach for multimodal fusion," in *Proc. Eur. Conf. Comput. Vis. Workshops*, Sep. 2018, pp. 1–15.

[21] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, no. 2, 2020, pp. 1359–1367.

[22] J. Jiang, B. Dai, D. Peng, C. Zhu, L. Liu, and C. Lu, "Neural synchronization during face-to-face communication," *J. Neurosci.*, vol. 32, no. 45, pp. 16064–16069, Nov. 2012.

[23] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, "The computational limits of deep learning," 2020, *arXiv:2007.05558*.

[24] J. Liu, T. Li, P. Xie, S. Du, F. Teng, and X. Yang, "Urban big data fusion based on deep learning: An overview," *Inf. Fusion*, vol. 53, pp. 123–133, Jan. 2020.

[25] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, "Multimodal deep learning models for early detection of Alzheimer's disease stage," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, Feb. 2021.

[26] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 284–288.

[27] S. Oramas, F. Barbieri, O. Nieto, and X. Serra, "Multimodal deep learning for music genre classification," *Trans. Int. Soc. Music Inf. Retr.*, vol. 1, no. 1, pp. 4–21, Sep. 2018.

[28] N. Audebert, C. Herold, K. Slimani, and A. Vidal, "Multimodal deep networks for text and image-based document classification," in *Machine Learning and Knowledge Discovery in Databases*. Cham, Switzerland: Springer, 2020, pp. 427–443.

[29] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "EmotiCon: Context-aware multimodal emotion recognition using Frege's principle," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14234–14243.

[30] A. Vergara, J. Fonollosa, J. Mahiques, M. Trincavelli, N. Rulkov, and R. Huerta, "On the performance of gas sensor arrays in open sampling systems using inhibitory support vector machines," *Sens. Actuators B, Chem.*, vol. 185, pp. 462–477, Aug. 2013.

[31] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Forster, G. Troster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J. D. R. Millan, "Collecting complex activity datasets in highly rich networked sensor environments," in *Proc. 7th Int. Conf. Netw. Sens. Syst. (INSS)*, Jun. 2010, pp. 233–240.

[32] S. Tripathi, S. Tripathi, and H. Beigi, "Multi-modal emotion recognition on IEMOCAP dataset using deep learning," 2018, *arXiv:1804.05788*.

[33] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Inf. Fusion*, vol. 68, pp. 46–53, Apr. 2021.

[34] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," 2019, *arXiv:1909.05645*.

[35] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2017, pp. 873–883.

[36] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," 2019, *arXiv:1905.05812*.

[37] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Attention driven fusion for multi-modal emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3227–3231.

[38] J. Chen, C. Wang, K. Wang, C. Yin, C. Zhao, T. Xu, X. Zhang, Z. Huang, M. Liu, and T. Yang, "HEU emotion: A large-scale database for multimodal emotion recognition in the wild," *Neural Comput. Appl.*, vol. 33, no. 14, pp. 8669–8685, Jul. 2021.

[39] W. Y. Choi, K. Y. Song, and C. W. Lee, "Convolutional attention networks for multimodal emotion recognition from speech and text data," in *Proc. Grand Challenge Workshop Hum. Multimodal Lang. (Challenge-HML)*, 2018, pp. 28–34.

[40] Z. Pan, Z. Luo, J. Yang, and H. Li, "Multi-modal attention for speech emotion recognition," 2020, *arXiv:2009.04107*.

[41] S. Nemati, R. Rohani, M. E. Basiri, M. Abdar, N. Y. Yen, and V. Makarenkov, "A hybrid latent space data fusion method for multimodal emotion recognition," *IEEE Access*, vol. 7, pp. 172948–172964, 2019.

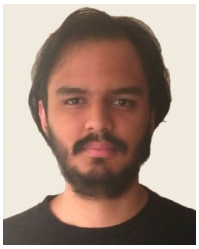
[42] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[43] W. Liu, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal deep learning," in *Neural Information Processing*. Cham, Switzerland: Springer, 2016, pp. 521–529.

[44] F. Povolny, P. Matejka, M. Hradis, A. Popková, L. Otrusina, P. Smrz, I. Wood, C. Robin, and L. Lamel, "Multimodal emotion recognition for AVEC 2016 challenge," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, Oct. 2016, pp. 75–82.

[45] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," 2017, *arXiv:1707.07250*.

[46] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. Pu Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," 2018, *arXiv:1806.00064*.
[47] J. Heredia, E. Lopes-Silva, Y. Cardinale, J. Diaz-Amado, I. Dongo, W. Graterol, and A. Aguilera, "Adaptive multimodal emotion detection architecture for social robots," *IEEE Access*, vol. 10, pp. 20727–20744, 2022.
[48] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176274–176285, 2020.
[49] H. Zhou, D. Meng, Y. Zhang, X. Peng, J. Du, K. Wang, and Y. Qiao, "Exploring emotion features and fusion strategies for audio-video emotion recognition," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 562–566.
[50] B. V. Dasarathy, *Decision Fusion*. Washington, DC, USA: Computer Society Press, 1994.
[51] J. Cai, M. Merler, S. Pankanti, and Q. Tian, "Heterogeneous semantic level features fusion for action recognition," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 307–314.
[52] U. G. Mangai, S. Samanta, S. Das, and P. R. Chowdhury, "A survey of decision fusion and feature fusion strategies for pattern classification," *IETE Tech. Rev.*, vol. 27, no. 4, pp. 293–307, 2010.
[53] J. Heredia, Y. Cardinale, I. Dongo, and J. Díaz-Amado, "A multi-modal visual emotion recognition method to instantiate an ontology," in *Proc. 16th Int. Conf. Softw. Technol.*, 2021, pp. 453–464.
[54] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.
[55] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, M. W. Jones X. Xie, and G. K. L. Tam, 2015, pp. 41.1–41.12.
[56] K. Venkataramanan and H. R. Rajamohan, "Emotion recognition from speech," 2019, *arXiv:1912.10458*.
[57] W. Shen, J. Chen, X. Quan, and Z. Xie, "DialogXL: All-in-one XLNet for multi-party conversation emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 13789–13797.
[58] W. Graterol, J. Diaz-Amado, Y. Cardinale, I. Dongo, E. Lopes-Silva, and C. Santos-Libarino, "Emotion detection for social robots based on NLP transformers and an emotion ontology," *Sensors*, vol. 21, no. 4, p. 1322, Feb. 2021.



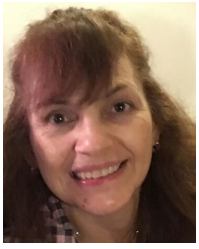
DIEGO PEÑA is currently pursuing the B.Sc. degree in computer science with Simó n Bolí var University. He has been collaborated as a Student in the Project Robots for Urban Tourism, Autonomous and Semantic (RUTAS), since 2021. His research interests include machine learning, neural networks, and deep learning for data classification.

ANA AGUILERA received the B.S. degree (Hons.) in computer science engineering from Lisandro Alvarado West-Central University (UCLA), Barquisimeto, Venezuela, in 1994, the M.S. degree in computer science from Simó n Bolí var University, Caracas, Venezuela, in 1998, and the Ph.D. degree in medical informatics from the University of Rennes I, Rennes, France, in 2008. She is currently a Full Professor with the Faculty of Engineering, Escuela de Ingeniería Informática, University of Valparaíso, Valparaíso, Chile. Her research interests include fuzzy databases, data mining, social networks, and medical informatics. She was accredited in Program for Researcher Promotion of Venezuela, Candidate Level, in 1998. Since 2011, she has been a member of the Program Encouragement for Research and Innovation Researcher (PEII) Level C, Venezuela. She is a member of the Venezuelan Association for the Advancement of Science (AsoVAC) and a member of Venezuelan Computer Society (SVC). She received the Magna Cum Laude Award for B.Sc. degree from UCLA and the "Très Honorable" Award in the Ph.D. thesis from the University of Rennes I.

IRVIN DONGO received the B.Sc. degree in computer science from Catholic San Pablo University, Peru, in 2012, and the M.Sc. and Ph.D. degrees from the University of Pau, France, in 2014 and 2017, respectively. From 2018 to 2020, he was a Postdoctoral Fellow at the École Supérieure des Technologies Industrielles Avancées, ESTIA, France, where he is currently an Associate Researcher of computer science. He is also a full-time Professor with Catholic San Pablo University. His research interests include normalization and anonymization of web resources, knowledge-bases modeling (semantic web), policies and management of credentials, security model and anonymization technique, and machine/deep learning techniques for an analysis and classification of data to discover patterns, gesture recognition, and affective computing.

JUANPABLO HEREDIA received the B.Sc. degree in computer science from Catholic San Pablo University, Arequipa, Peru, in 2021.

He has been participating as a Research Student in the Project Robots for Urban Tourism, Autonomous and Semantic (RUTAS) Web Based, since April 2020, where he developed his undergraduate thesis and participated in other research. His research interests include machine/deep learning models, computer vision algorithms, graph-based machine/deep learning models, affective computing, and the integration between artificial intelligence and neuroscience.



YUDITH CARDINALE received the degree (Hons.) in computer engineering from the Universidad Centro-Occidental Lisandro Alvarado, Venezuela, in 1990, and the M.Sc. and Ph.D. degrees in computer science from Universidad Simón Bolívar (USB), Venezuela, in 1993 and 2004, respectively. She has been a Full Professor and a Researcher at Universidad Internacional de Valencia, Spain, since 2018. She has been also a Full Professor with the computer science Department, USB, since 1996. She is currently a Principal Researcher at the Research Group in Data Science (GRID). She is also an Associate Researcher with Universidad Católica San Pablo, Arequipa, Perú. She has been the Coordinator of a variety of international research projects. She has written a range of scientific papers published in international journal, books, and conferences. Her research interests include parallel processing, distributed object processing, operating systems, digital ecosystems, high performance on grid and cloud platforms, collaborative frameworks, and web services composition, including semantic web. She has participated as a member of program committees of several international conferences and journals. She is the President of the Venezuelan Computer Society.