IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 July 2024, accepted 20 August 2024, date of publication 27 August 2024, date of current version 6 September 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3450674*

RESEARCH ARTICLE

# Uncertainty-Based Learning of a Lightweight Model for Multimodal Emotion Recognition

**ANAMARIA RADOI, (Member, IEEE), AND GEORGE CIOROIU**
Department of Applied Electronics and Information Engineering, NUST Politehnica Bucharest, 060042 Bucharest, Romania

Corresponding author: Anamaria Radoi (anamaria.radoi@upb.ro)

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CCCDI-UEFISCDI, project number PN-III-P2-2.1-PED-2021-3486 (MES-ER), within PNCDI III.

**ABSTRACT** Emotion recognition is a key research topic in the Affective Computing domain, with implications in marketing, human-robot interaction, and health domains. The continuous technological advances in terms of sensors and the rapid development of artificial intelligence technologies led to breakthroughs and improved the interpretation of human emotions. In this paper, we propose a lightweight neural network architecture that extracts and performs the analysis of multimodal information using the same audio and visual networks across multiple temporal segments. Undoubtedly, data collection and annotation for emotion recognition tasks remain challenging aspects in terms of required expertise and effort spent. In this sense, the learning process of the proposed multimodal architecture is based on an iterative procedure that starts with a small volume of annotated samples and allows a step-by-step improvement of the system by assessing the model uncertainty in recognizing discrete emotions. Specifically, at each epoch, the learning process is guided by the most uncertainly annotated samples and integrates different modes of expressing emotions through a simple augmentation technique. The framework is tested on two publicly available multimodal datasets for emotion recognition, i.e. CREMA-D and RAVDESS, using 5-folds cross-validation. Compared to state-of-the-art methods, the achieved performance demonstrates the effectiveness of the proposed approach, with an overall accuracy of 74.2 % on CREMA-D and 76.3 % on RAVDESS. Moreover, with a small number of model parameters and a low inference time, the proposed neural network architecture represents a valid candidate for the integration on platforms with limited memory and computational resources.

**INDEX TERMS** Convolutional neural networks, entropy, multimodal emotion recognition, uncertainty-based learning, MTCNN, CREMA-D, RAVDESS.

## I. INTRODUCTION

The fast technological development that characterizes the human-computer interaction domain is mainly driven by the evolution that occurred in the Artificial Intelligence (AI) domain. Numerous human-computer interaction applications are based on emotion recognition, e.g. social robotic applications [1], healthcare systems [2], [3], customer service and marketing [4], entertainment industry [5], politics [6], and even surveillance, policing and criminology systems [7]. Being at the border between Affective Computing and Social Signal Processing, emotion recognition systems target the recognition, processing and simulation of human affects, while focusing on the analysis of verbal and non-verbal information in various social-oriented scenarios [8].

Humans express emotions in a multimodal manner, e.g., facial expressions, speech inflection, or vocal intensity contain relevant information for the identification of the emotional state of a subject. In order to improve the performance of emotion recognition models, information collected from multiple modalities can be efficiently fused, leading to higher accuracy levels if compared to monomodal approaches [9]. However, most of the human-computer interaction applications may be requested to operate on

The associate editor coordinating the review of this manuscript and approving it for publication was Thomas Canhao Xu.

embedded platforms with limited resources (i.e., low level of memory and low computational power). Although many of the current approaches fuse information extracted from the entire video sequence, the development of lightweight models should start by considering a limited number of video frames and temporal windows extracted from the audio signal. Moreover, for a further integration of the multimodal emotion recognition system on platforms with limited memory and computational resources, both the number of model parameters and the complexity of the algorithms used for feature extraction should be limited.

Considering the increased number of application domains and the diversity of modes to express emotions, delivering a personalized experience in terms of human-computer interaction relies on the ability of AI-based systems to interpret the emotional state of a user from multimodal data in an accurate and timely manner. However, training such systems requires the existence of large labeled datasets. The acquisition of multimodal datasets targeting the emotion recognition domain is a difficult task due to an expensive annotation procedure, both in terms of time and expertise required for a correct identification of emotions. Therefore, other solutions to counterbalance this limitation are required in order to reduce the amount of annotated data needed for training an efficient emotion recognition system.

In this paper, a lightweight multimodal emotion recognition framework is described. In order to obtain a model with a reduced number of parameters, the feature extraction modules for audio and visual information retrieval use the same convolutional neural networks across multiple temporal segments. The proposed approach ensures obtaining time-invariant embeddings for the representation of the audio and visual information, and, thus, allows a sliding window approach for real-time processing of streaming multimodal data. Moreover, considering that the availability of labeled datasets represents a constant challenge when training emotion recognition frameworks and that not all the training samples lead to extracting meaningful information regarding the expressed emotions, we propose an uncertainty-based technique for the selection of the most relevant samples when training the model. Based on an iterative approach, the uncertainty of the recognition system is reduced gradually. More precisely, the training procedure starts with a limited amount of annotated data and gradually increases the training volume with samples for which the emotion recognition system is still the most uncertain. The main contributions of the paper are five-fold, namely: (i) we propose an efficient end-to-end lightweight framework for emotion recognition, which extracts audio and visual information from consequent temporal segments using the same neural network modules, (ii) the learning procedure relies on an uncertainty-based sampling technique, which allows the training process to start with a limited amount of annotated data and yields a robust emotion recognition system, (iii) due to its decreased computational complexity and a processing time shorter than 14 ms, the proposed model represents a valid candidate for online emotion recognition using a sliding window approach, (iv) due to its small model size (i.e., on average, the model occupies 10.3 MB), the proposed model is a valid candidate for deployment in a real-time autonomous development platform with limited memory and computational power, and (v) the effectiveness of the proposed approach is demonstrated on two publicly available datasets, namely CREMA-D [10] and RAVDESS [11].

The rest of the paper is organized as follows. After a thorough discussion of existing emotion recognition methods in Section II, Section III presents the proposed lightweight convolutional neural network architecture for audio-visual emotion recognition, along with the uncertainty-based learning approach. Section IV is dedicated to presenting the CREMA-D and RAVDESS datasets, whereas the experimental results, accompanied by comparisons with existing methods, are discussed in Section V. Finally, Section VI concludes the paper.

## II. RELATED WORK

Numerous monomodal approaches towards emotion recognition have been proposed in the last years. For example, approaches designed for speech emotion recognition integrate basic audio features (e.g., signal energy, loudness, mel frequency cepstral coefficients (MFCC), pitch, formants, spectral shape descriptors) [12], that can be extracted using the largely deployed openSMILE toolkit [13]. Besides the basic audio features mentioned before, time-frequency representations also lead to achieving competitive performance in terms of speech-based emotion recognition [9], [14]. With the advancement of deep learning techniques, convolutional neural networks (CNN) have been frequently used to extract relevant time-frequency descriptors from spectrograms [15], [16]. Other methods integrate CNN-based modules in x-vector models using multi-head attention for utterance level speaker embedding extraction [17]. Tzirakis et al. propose a different approach towards feature extraction from speech, namely the application of CNNs directly on the raw speech signals [18]. Considering that fully convolutional networks (FCN) extract spatial information, Zhao et. al consider inserting an additional attention-based bidirectional Long Short-Term Memory (LSTM) module in order to emphasize the temporal characteristic of the spectrogram [14].

Facial expressions represent relevant cues for understanding human emotions, which can be described though a mix of several Facial Action Units (FAUs) [19]. In this regard, Ekman identifies several facial attributes that allow emotion recognition from face expressions (e.g., morphology, symmetry, duration, coordination of facial muscles) [20], [21]. CNNs, either pre- or end-to-end trained state-of-the-art architectures (e.g., VGG [22], GoogleNet [23], ResNet [24]), have been frequently used as feature extractors for facial emotion recognition tasks (e.g., DeXpression [25], MERML [12], Tzirakis et al. [18], Dixit and Satapathy [26]). Moreover, combining the results yield by multiple

CNNs, both randomly initialized or pre-trained, leads to improved performance [27]. Furthermore, approaches based on ensemble methods show promising results in many domains [28], [29], including multimodal emotion recognition [30]. Although improved performance can be obtained using multiple architectures in parallel, the main drawback is the increased complexity, which limits their usage on platforms with limited resources.

An enhanced emotion recognition system can be obtained by combining the analysis of facial expressions in video frames with the information retrieved from speech [9], [12], [31], [32], [33], [34], [35]. Various intra-modal and cross-modal fusion strategies, including attention mechanisms to highlight important emotion features, feature concatenation and factorized bilinear pooling (FBP) for cross-modal feature fusion, have been explored in [36]. An efficient combination of multiple modalities can be obtained by assigning dynamic weights in generalized mixture functions applied at decision level [33]. Another multimodal fusion strategy is achieved by concatenating the features that were extracted from each modality via independent fully connected layers, whilst the result is provided as input to an additional fully concatenated layer that performs the final classification [37]. A slow modality fusion can be achieved by inserting multimodal transfer modules at different levels of the feature hierarchy in an intermediate fusion approach [38].

Embedding the audio and visual content onto a metric space is the solution proposed in [32] for reducing the gap between modalities. The temporal joint embeddings are obtained by connecting multiple LSTM cells that lead to an uncertainty-based learning of the audio-visual information across time, taking, thus, into account the broader context and the temporal evolution of the emotion throughout the entire video sequence. In [32], the visual features were extracted with 3D-CNN [39], whereas the audio features were extracted from raw signals using soundNet [40], by transferring and synchronizing discriminative knowledge across visual and sound networks. In a similar vein, a metric learning paradigm, called Multimodal Emotion Recognition Metric Learning (MERML), is designed in [12] in order to obtain a better discrimination and an enhanced representation in a latent space for both modalities. The learned metric is used as a distance for the Radial Basis Function (RBF) kernel incorporated in a Support Vector Machine (SVM) classifier [12]. The main limitations of these models are related to the complexity of the proposed solutions, which may hinder both real-time processing and inclusion of the models on platforms with limited capabilities.

The emotions that each person share may differ in intensity and modes of expression. This aspect may induce a certain degree of uncertainty in assessing the emotion expressed by each person. Therefore, the approaches that target person-specific emotion consider the insertion of a group of neural networks which act as personalized emotional memories retaining individual aspects of emotional expressions [41]. Apart from the person-specific neural network, a model that contains an adversarial autoencoder, is configured for representing the general aspects of emotions [41]. The role of the adversarial autoencoder is, on one hand, to learn general representations of facial expressions and, on the other hand, to generate new images of expressions for a particular person using conditional emotional information. This collection of expressions is then used to initialize a Grow-When-Required (GWR) neural network that functions as a personalized affective memory which captures the particular expressions of emotions for a subject. However, the technique proposed in [41] is difficult to apply for an online emotion recognition task, where the processing of instantaneous emotion expressions is a requirement.

Attention mechanisms that increase or decrease the importance of particular time-windows and modalities are often used in multimodal learning and fusion frameworks dealing with emotion recognition [42], [43], [44]. In order to capture the dynamic characteristics between video frames, Beard et. al proposed a recurrent multi-attention (RMA) mechanism with shared external memory that allows the temporal information to persist over multiple hops before being updated during the analysis framework [31]. By means of an attention-based fusion between facial and audio temporal features, the recognition performance achieved by the method proposed in [31] is comparable to crowd-sourced human rating [10]. The emotion recognition system with attention mechanism proposed in [45] consists of two encoder sub-networks, one for each modality, integrated in a Multi-Head Self-Attention framework. The original video sequence is divided into separate time windows, each window containing a sequence of frames, whereas the sub-networks, based on pretrained VGG-based architectures, are used to extract features from the video frames and audio signals, similar to the approach shown in [46]. Considering an inter-modality-based scheme for the attention mechanism, the multimodal emotion recognition system proposed in [42] uses attention to guide the extraction of visual features by means of the information extracted from the audio signal.

Transformers currently play an instrumental role in image recognition [47] and numerous linked sub-domains, including emotion recognition [16], [48], [49]. In [16], a multimodal transformer with three branches for audio self-attention, video self-attention, and audio-video cross-attention, respectively, is proposed, along with a block embedding that captures the temporal information within the video frames. In an attempt to find links between facial and audio cues, audio-visual transformers have been also proposed for building improved feature extraction models [48]. However, training transformer models requires large volumes of data, which are, in general, difficult to obtain for emotion recognition tasks. A possible solution for mitigating the problem of limited labeled data is to use active learning strategies that can boost the performance of

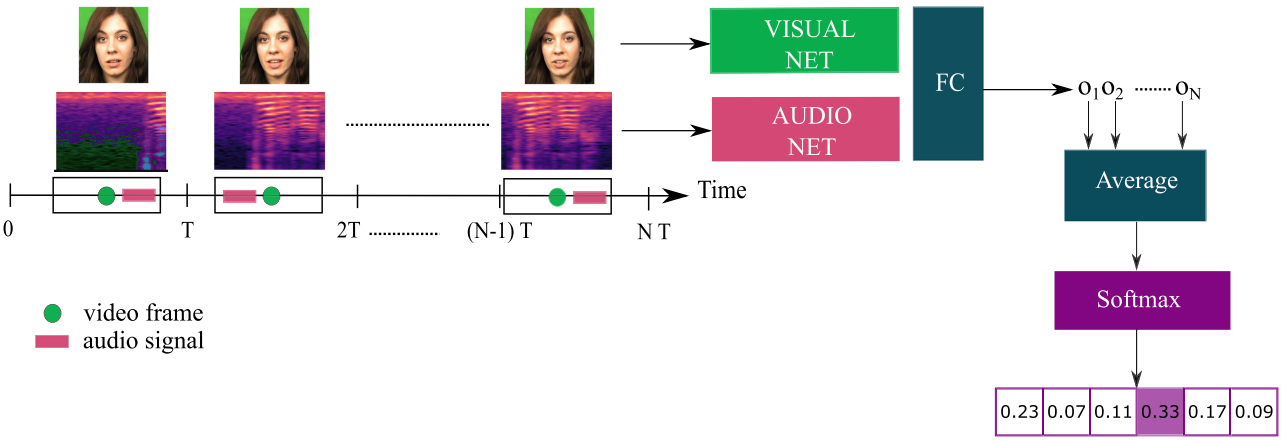



**FIGURE 1.** Proposed multimodal neural network architecture using the same Visual and Audio Neural Networks across all temporal segments.

deep learning methods, or to use models that have been pre-trained on large datasets designed for different tasks, e.g. the VoxCeleb2 dataset, designed for large scale speaker recognition, was used for pre-training the model proposed in [48].

## III. PROPOSED METHOD

### A. DEFINITION AND NOTATIONS

We consider a labeled multimodal dataset $\mathcal{D}$, consisting of $M$ audio-visual pairs and corresponding labels representing discrete human emotions:

$$\mathcal{D} = \{(\mathbf{X}_1^a, \mathbf{X}_1^v, L_1), (\mathbf{X}_2^a, \mathbf{X}_2^v, L_2), \ldots, (\mathbf{X}_M^a, \mathbf{X}_M^v, L_M)\}$$

where $(\mathbf{X}_i^a, \mathbf{X}_i^v)$ is the $i^{th}$ audio-visual pair and $L_i$ is the corresponding discrete emotion label. The goal is to predict the emotion expressed by the subject in each audio-visual test pair. The total number of possible discrete human emotion labels is denoted by $C$.

The dataset $\mathcal{D}$ is split into three subsets, namely, $\mathcal{D}_{train}^{(0)}$ used for initial training, $\mathcal{U}$ used for uncertainty-based learning, and $\mathcal{D}_{test}$ used for evaluation. The subsets are independent at actors' level, i.e., an actor pertains to only one of the three subsets.

### B. LIGHTWEIGHT MULTIMODAL NEURAL NETWORK FOR EMOTION RECOGNITION

The multimodal neural network proposed in this paper fuses information retrieved from the visual and audio domains by means of convolutional neural network (CNN) architectures trained end-to-end. The multimodal signal is divided into $N$ temporal segments, and, for each temporal segment, we extract both visual and audio-related features using two distinct neural networks, one for each modality. The scheme of the proposed multimodal neural network is shown in Figure 1, whereas the structure of the architecture is depicted in Table 1. The usage of the same audio and visual neural networks across different temporal segments ensures an overall architecture with fewer parameters to learn and, also, invariance with respect to time shifts at segment level. The invariance characteristic leverages the possibility to apply a sliding window approach for processing streaming multimodal data (Figure 2) without the need to extract the features for each temporal segment several times.

#### 1) TEMPORAL FUSION OF MULTIMODAL INFORMATION

The temporal fusion of multimodal information is obtained by dividing the multimodal signal into $N$ temporal segments, extracting relevant information from each modality and then merging the information retrieved from both modalities into a compressed representation. The aggregation of audio and visual information is performed in an asynchronous manner, allowing a temporal offset between the two modalities and a natural augmentation of the dataset. More precisely, at each training epoch, a different pair of audio and visual content is selected for learning, within a random temporal offset. The multimodal temporal-binding approach, i.e., combining different modalities within a range of offsets, has been previously introduced in [50] for egocentric action recognition and, then, extended for emotion recognition in [51]. The main difference with respect to the method our team proposed in [51] consists in using the same audio and visual neural networks for the feature extraction process, across all temporal segments. Thus, the proposed approach yields time-invariant audio and visual embeddings with respect to shifts of the analysis window at temporal segment level. Similarly, the weights of the fully connected layer performing the classification are the same for all the temporal segments. By contrast, the architecture presented in [51] is composed of $N$ distinct audio and visual neural networks and fully connected layers. Thus, the approach in [51] yields a model architecture that linearly grows with the number of temporal segments, which hinders the architecture's integration on embedded platforms with limited resources, e.g., low memory and low computational power.

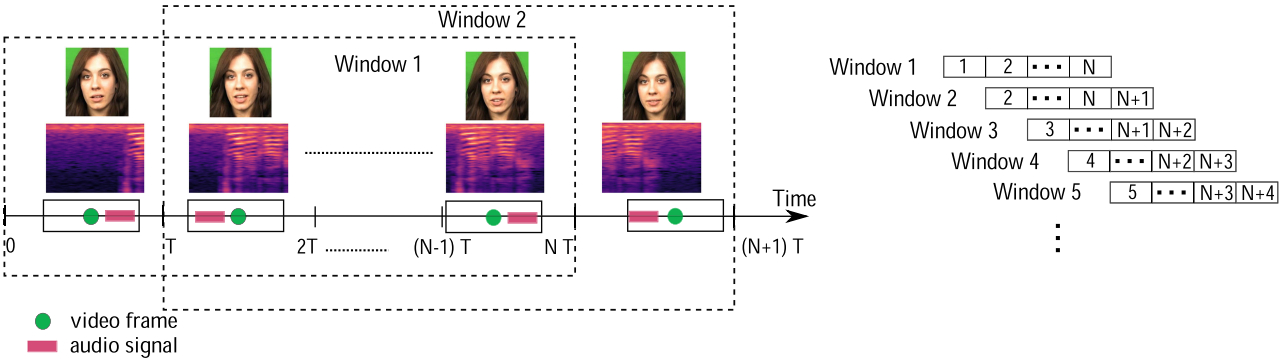



**FIGURE 2.** Sliding window approach for processing streaming multimodal data.

**TABLE 1.** Multimodal neural network architecture with CNN for both Audio Net (AN) and Visual Net (VN).

| # | Layer | Type | Size | Kernel | (stride, pad) | No. parameters |
|---|---|---|---|---|---|---|
| Visual Net (VN) | input | input | $1 \times 98 \times 80$ | – | – | 0 |
| | conv1_v | Conv2D + BN + ReLU | $16 \times 96 \times 78$ | $3 \times 3$ | (1, 0) | 192 |
| | pool1_v | MaxPool | $16 \times 48 \times 39$ | $2 \times 2$ | (2, 0) | 0 |
| | conv2_v | Conv2D + BN + ReLU | $32 \times 46 \times 37$ | $3 \times 3$ | (1, 0) | 4704 |
| | pool2_v | MaxPool | $32 \times 23 \times 18$ | $2 \times 2$ | (2, 0) | 0 |
| | conv3_v | Conv2D + BN + ReLU | $64 \times 21 \times 16$ | $3 \times 3$ | (1, 0) | 18624 |
| | pool3_v | MaxPool | $64 \times 10 \times 8$ | $2 \times 2$ | (2, 0) | 0 |
| | linear_v | linear | 512 | – | – | 2621952 |
| Audio Net (AN) | input | input | $1 \times 120 \times 192$ | – | – | – |
| | conv1_a | Conv2D + BN + ReLU | $4 \times 118 \times 190$ | $3 \times 3$ | (1, 0) | 48 |
| | pool1_a | MaxPool | $4 \times 59 \times 95$ | $2 \times 2$ | (2, 0) | 0 |
| | conv2_a | Conv2D + BN + ReLU | $8 \times 57 \times 93$ | $3 \times 3$ | (1, 0) | 312 |
| | pool2_a | MaxPool | $8 \times 28 \times 46$ | $2 \times 2$ | (2, 0) | 0 |
| | conv3_a | Conv2D + BN + ReLU | $12 \times 26 \times 44$ | $3 \times 3$ | (1, 0) | 900 |
| | pool3_a | MaxPool | $12 \times 13 \times 22$ | $2 \times 2$ | (2, 0) | 0 |
| | linear_a | linear | 128 | – | – | 64640 |
| FC | drop | dropout | 640 | – | – | 0 |
| | fc | linear | 6 or 8 | – | – | 3846 or 5128 |
| Total | – | – | – | – | – | 2715218 or 2716500 |

Considering the high redundancy level among consecutive frames, the analysis of the visual information is performed only over one randomly selected frame from each temporal segment. Therefore, instead of processing the entire video sequence, only $N$ video frames are analyzed using the same visual neural network. In order to link the audio content to the visual information, we consider sequences of audio signals of length $d$, whereas the center of each audio sequence is randomly located within a certain offset with respect to the selected frame. For each temporal segment $i \in \{1, 2, \ldots, N\}$, the concatenated visual and audio information is jointly interpreted by means of a fully connected (FC) layer with $C$ output neurons, $C$ being the number of discrete emotion category labels. For each temporal segment $i$, the output of the FC layer is denoted by $\mathbf{o}_i$. In order to combine the temporal information, the outputs of the FC layer are averaged along the $N$ temporal segments:

$$\mathbf{z} = \frac{\mathbf{o}_1 + \mathbf{o}_2 + \ldots + \mathbf{o}_N}{N} \tag{1}$$

The result, $\mathbf{z} = [z_1\ z_2\ \ldots\ z_C]$, is passed through a Softmax function, which transforms the $C$ real values into a vector of normalized elements that sum to 1 [52]:

$$y_c = \frac{e^{z_c}}{\sum_{k=1}^{C} e^{z_k}}. \tag{2}$$

Considering the structure shown in Table 1, the number of parameters of the overall neural network architecture slightly varies with the number of neurons of the FC layer, which is associated to the number of discrete emotion category labels $C$, i.e. 6 labels in the case of CREMA-D dataset and 8 labels in the case of RAVDESS. Moreover, apart from using a random selection of the video frames and audio signals inside each temporal segment, a dropout rate of 0.2 is considered in order to reduce overfitting [53].

### 2) VISUAL NEURAL NETWORK

The visual neural network consists of convolutional layers, with sequences of 2D convolution – batch normalization (BN) – Rectified Linear Unit (ReLU) activation function,

followed by a max pooling layer (Table 1). The analysis is performed on only $N$ frames that correspond to each temporal segment. It is worth mentioning that, at each training epoch, the frames are randomly selected from the temporal segment of the video sequence, capturing, thus, various facial expressions of the subjects within the same emotional state. This yields a natural augmentation of the training set.

The convolutional layers have 16, 32 and 64 filters of $3 \times 3$ and a stride of 1. The role of the batch normalization layer is to increase the stability of the convolutional neural network [54], whereas the max pooling layer reduces the dimensionality of the representation at each step. The last layer of the visual neural network is a fully connected layer which translates the output of the last convolutional layer into a vector of length 512.

In order to remove the unnecessary information regarding the background, the analysis at frame level requires a pre-processing step related to face detection. In video analysis, the quick movements of the subject's head have a direct impact on the success of the entire recognition framework. A reliable method for face detection and alignment is the deep cascaded multi-task framework based on convolutional neural networks, i.e. MTCNN proposed in [55]. MTCNN is composed of three stages of processing with convolutional neural networks. The role of the first stage, called P-Net, is to propose several candidate windows containing the face of the subject. The false candidates are removed in the second stage, called R-Net. The last stage, called O-Net, performs the analysis of five facial landmarks that contain the most information regarding the face.

#### 3) AUDIO NEURAL NETWORK

For the raw audio signal, the feature extraction module consists in deriving the spectrogram for each short-term signal in the temporal segments and applying the sequence of convolutional layers to extract the audio-based features. The spectrogram is a powerful time-frequency analysis tool for speech processing [56] and speech recognition [57], including the speech emotion recognition task [58]. One of the most commonly used 2D time-frequency representations of raw audio signals is the discrete Short-Time Fourier Transform (STFT) [59]. However, STFT does not lead to a perceptually-inspired processing with respect to the human auditory characteristics [60]. By contrary, applying the Mel scale over the frequency range leads to a better approximation of the human perception. The Mel scale is, in fact, an approximation to the cochlea's non-linear frequency scaling, which is obtained from the linear frequency scale by applying a non-linear transformation [61]. In order to obtain the Mel spectrogram, a filterbank of triangular filters is applied over the raw signal in the frequency domain. Considering that the logarithmic variant of the Mel-filtered spectrogram represents a common choice for numerous approaches based on CNN architectures for speech recognition tasks, we also adopt the Log-Mel variant of the spectrogram [56].

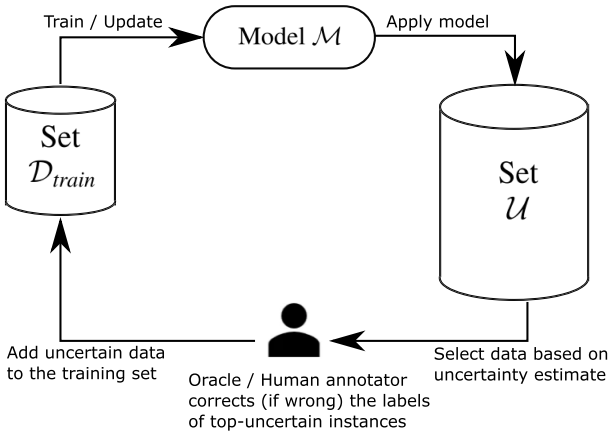



**FIGURE 3.** Workflow for uncertainty-based learning of the emotion recognition model.

Similar to the visual neural network, the audio neural network is composed of sequences of 2D convolution – batch normalization (BN) – Rectified Linear Unit (ReLU) activation function, followed by a max pooling layer (Table 1). The convolutional layers have 4, 8 and 12 filters of $3 \times 3$ and a stride of 1. The last layer of the audio neural network is a fully connected layer which translates the output of the last convolutional layer into a vector of length 128.

#### 4) END-TO-END TRAINING OF THE MULTIMODAL NEURAL NETWORK

The multimodal neural network architecture is trained end-to-end using the cross-entropy loss, computed between input logits and target [52]:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{B}\sum_{c=1}^{C} \delta_{i,c} \log(y_{c,i}) \tag{3}$$

where $y_{c,i}$ is the $c^{\text{th}}$ output of the last fully connected (FC) layer that corresponds to sample $\mathbf{x}_i$ in the minibatch of size $B$ and $\delta_{i,c}$ is the $c^{\text{th}}$ element of the target vector and equals 1 or 0 if the class for input $\mathbf{x}_i$ is $c$ or not.

### C. UNCERTAINTY-BASED LEARNING APPROACH

As already mentioned in Section II, the variability in expressing emotions and changes in intensity induce uncertainty in assessing the emotion category. Moreover, training robust multimodal neural network architectures requires a large number of annotated samples, that might be expansive to obtain for the emotion recognition task. The proposed method for training a robust emotion recognition system is centered around an uncertainty-based learning approach, that involves the identification of most uncertainly-labeled examples pertaining to a different set than the one used for training.

For the first epoch, the training procedure starts with a small annotated dataset, namely, $\mathcal{D}_{train}^{(0)}$. In addition,

**Algorithm 1** Proposed Training Algorithm With Most Uncertainly Labeled Instances

**Require:** Initial training set $\mathcal{D}_{train}^{(0)}$, labeled set $\mathcal{U}$, number of epochs $N_{epoch}$
**Ensure:** Parameters of model $\mathcal{M}$
1: Train initial model $\mathcal{M}$ on set $\mathcal{D}_{train} = \mathcal{D}_{train}^{(0)}$
2: **for** $n = 0$ to $N_{epoch} - 1$ **do**
3: **for** each instance $\mathbf{x} \in \mathcal{U}$ **do**
4: Compute the neural network output for instance $\mathbf{x}$, i.e. $\mathcal{M}(\mathbf{x})$
5: Compute the entropy over the output probability distribution of $\mathcal{M}(\mathbf{x})$
6: **end for**
7: Sort entropy values in descending order
8: Add first $N_u$ data instances to the training set $\mathcal{D}_{train}$
9: Check if the assigned labels for the $N_u$ data instances are correct and, if not, correct them
10: Remove same data instances from $\mathcal{U}$
11: Update the parameters of model $\mathcal{M}$
12: **end for**

we consider having access to another set of instances $\mathcal{U}$. The dataset $\mathcal{D}_{train}^{(0)}$ is randomly selected, with actors that are not part of either $\mathcal{D}_{test}$ or $\mathcal{U}$. The training procedure is iterative, and, after each training epoch, the training dataset is augmented with the most uncertain examples from $\mathcal{U}$. If the labels associated by the recognition system are incorrect, the labels are corrected by an oracle or human annotator, as shown in Figure 3. The overall training strategy for uncertainty-based learning is summarized in Algorithm 1.

The majority of the classification errors are produced by the examples that the recognition system is uncertain about. In information theory, the uncertainty of a random variable is measured through entropy [62]. Therefore, the most uncertainly labeled examples from $\mathcal{U}$ are determined by means of computing the entropy over the output probability distribution. Given a probability distribution $\mathbf{y} = [y_1\ y_2\ \dots\ y_C]$ over the predicted output classes, the entropy is computed as [62]:

$$H(\mathbf{y}) = -\sum_{c=1}^{C} y_c \log_2 y_c. \tag{4}$$

The entropy is maximum when the probabilities are equal and the system is characterized by a high degree of uncertainty.

The selection of the most uncertainly labeled examples from $\mathcal{U}$ is made in the descending order of the entropy values. At each training epoch, the first $N_u$ data samples for which the system was most uncertain are inserted in the training dataset $\mathcal{D}_{train}$ and removed from $\mathcal{U}$. Therefore, the emotion recognition system is trained to identify particular emotions from expressions that are less common, enhancing the generalization capacity of the classifier.

## IV. DATASETS

The algorithm developed for multimodal emotion recognition was validated on publicly available datasets CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) [10] and RAVDESS (The Ryerson Audio-Visual Database of Emotional Speech and Song) [11].

CREMA-D is composed of 7442 video sequences with 91 actors expressing an emotion from a set of predefined emotions ( i.e., anger, fearness, disgust, happiness, neutral, and sadness). The actors are selected from a wide variety of races and ethnicity, i.e., African-American, Asian, European American, Hispanic, and are divided into 48 males and 43 females. The videos have an average length of 2.54 seconds, with lengths that vary between 1.27 seconds to 5 seconds. Apart from the actors' interpretations of particular emotions, the video sequences were passed through a crowd-sourced labeling process, i.e. each video sequence was rated by more than 7 participants to this process. The study led to the following results with respect to human-level accuracy in terms of emotion recognition: 40.9 % accuracy when only audio information is labeled, 58.2 % when only video is available, and 63.6 % when the audio and video information is assessed [10].

RAVDESS [11] contains 1440 videos recorded from 24 professional actors (12 male and 12 female), where only the speech signals have been considered. The actors were asked to read two statements using a neutral North American accent while emphasizing 8 categories of emotions (i.e., neutral, calm, happy, sad, angry, fearful, surprise, and disgust). The average length of the audio-video sequences is 6.88 seconds, ranging from 5.49 seconds to 9.76 seconds.

## V. EXPERIMENTAL RESULTS

### A. EXPERIMENTAL SETUP

The multimodal emotion recognition system relies on both audio and video inputs, which are divided into $N$ temporal segments. Considering that the length of the videos is longer on average for RAVDESS than for CREMA-D, the number of temporal segments for CREMA-D is set to 5, whilst for RAVDESS, 12 temporal segments are considered. The frames of the video sequences are processed using the MTCNN method [55] for face detection and alignment. However, in order to obtain the same dimensions of the images containing the cropped faces, all images are resized to $80 \times 98$ pixels. Regarding the audio signal analysis, the length $d$ of the audio window is set to 1.5 seconds, whereas the center of the audio window is randomly selected within a maximum offset of 10 ms with respect to the chosen video frame for each temporal segment. The time-frequency representation of each audio signal is the Log-Mel spectrogram, computed using 128 evenly-spaced frequencies on the Mel scale, hop length of 512, 2048 FFT window samples, and a rate of the audio signal of 16 KHz for CREMA-D or 48 kHz for RAVDESS.

The selection of the number of temporal segments and the maximum offset value follows the recent work we have

**TABLE 2.** Hyper-parameters of the model.

| Parameter | Value |
|---|---|
| Dropout | 0.2 |
| Learning rate | 1e-4 (1e-5) |
| Optimizer | Adam |
| Batch size | 16 |
| Weight initialization | Xavier |

developed in [51] and [63], where a similar procedure for the temporal aggregation of the multimodal information was proposed. Compared to the approach proposed in [51] for extracting multimodal information, the main difference consists in using the same neural network architectures to extract audio and visual information across multiple temporal segments, i.e., one for the audio information, one for the visual information. On the opposite, the TA-AVN architecture proposed in [51] uses different networks for each temporal segment, leading to a linear increase of the number of parameters with respect to the number of temporal segments considered.

In order to train the multimodal emotion recognition system, the optimization of the objective functions is performed using the Adam algorithm [64]. The learning rate was set to 1e-4, and the training was performed over 100 epochs, with a multiplicative factor of the learning rate decay of 0.1 after 50 epochs. The weights of the architecture were initialized using the Xavier method with a uniform probability distribution [65]. The main hyper-parameters of the model are summarized in Table 2.

The performance of the proposed approach for multimodal emotion recognition is assessed through 5-fold cross validation at subject level, i.e., there is no overlap between subjects in the sets $\mathcal{D}_{train}^{(0)} \cup \mathcal{U}$ and $\mathcal{D}_{test}$. The training of the model begins with an initial subset $\mathcal{D}_{train}^{(0)}$, composed of videos acquired from 1/3 subjects from $\mathcal{D}_{train}^{(0)} \cup \mathcal{U}$. The rest of the subjects from the train dataset form the subset $\mathcal{U}$. This ensures that $\mathcal{D}_{train}^{(0)}$, $\mathcal{D}_{test}$ and $\mathcal{U}$ sets do not overlap at subject level. During the iterative training procedure, the volume of the training samples, $\mathcal{D}_{train}$, is augmented with uncertain examples sampled from $\mathcal{U}$. At each epoch, the number of uncertain samples is $N_u = 20$ for CREMA-D and $N_u = 5$ for RAVDESS. By contrary, when the uncertainty-based strategy is not considered, the update of the parameters is performed over the entire training set for each epoch in part. A natural augmentation process is achieved in both cases, i.e., with and without the uncertainty-based learning strategy, since the frames are randomly selected inside the temporal segments, whereas the audio windows are randomly selected within a small offset with respect to the chosen video frames.

### B. DISCUSSION

#### 1) PERFORMANCE ASSESSMENT

The performance achieved by the proposed method for multimodal emotion recognition has been measured through the cross-validation technique using 5 folds. The overall accuracy reached 74.2 % on the CREMA-D dataset and 76.3 % on the RAVDESS dataset.

In order to check the impact of the uncertainty-based approach on the training procedure, we show the variation of the loss and the overall accuracy values with respect to training epochs in Figure 4 and Figure 5. If compared to training over the whole training dataset from the beginning, the uncertainty-based learning strategy keeps the performance with respect to train loss in a similar range, whilst the learning process is guided through the examples about which the recognition system is most uncertain. As shown in Figure 4, the decay of the loss function is similar when the iterative learning procedure is considered. This behavior is achieved even if the training procedure is performed over a smaller amount of data, as in the case of RAVDESS. In addition, as shown in Figure 5, the maximum level of accuracy is achieved very fast, in less than 60 epochs of training.

The confusion matrices for both datasets are provided in Figure 6. In both cases, the proposed system recognizes happiness, neutral and anger states with a high level of precision. However, fear is easily confused with anger, sadness, neutrality or surprise. This behavior is similar to a large extent to the performance of human annotation when a crowd-sourcing labeling experiment was conducted over the CREMA-D dataset [10].

Moreover, the training time is improved when the uncertainty-based learning strategy is adopted. For example, in the case of the RAVDESS dataset, the training procedure takes almost 8 hours and 30 minutes when it is performed without the uncertainty-based learning strategy, and 7 hours and 40 minutes (including the inference over set $\mathcal{U}$) when the uncertainty-based strategy is considered. We mention that the experiments were carried on an 12th Gen Intel(R) Core(TM) i9-12900KF, 3.19 GHz, with 32 GB of RAM, and equipped with NVIDIA GeForce RTX 3080 Ti GPU with 12 GB of dedicated GPU memory.

#### 2) COMPARISONS WITH OTHER METHODS

Comparisons with recent approaches are provided in Table 3, along with human labeling accuracy [10], [11]. The proposed recognition system outperforms previous approaches, e.g. intermediate fusion of multiple modalities [38], methods based on recursive attention [31], deep metric learning [12], [32], attention mechanisms across modalities [43], [44], [45], or audio-visual transformers [48], [49]. Following a similar approach for the temporal aggregation of multimodal features extracted by different convolutional neural networks, TA-AVN [51] reaches a smaller overall accuracy in a 5-folds cross-validation setup. However, as detailed in the next section, it is important to note that the number of parameters that characterize the architectures achieving top-performance results is several times higher than in the case of the proposed architecture. Furthermore, if compared to the performance achieved by multimodal transformers [16], the class precision reported in [16] is high when retrieving anger (76.1 %) and

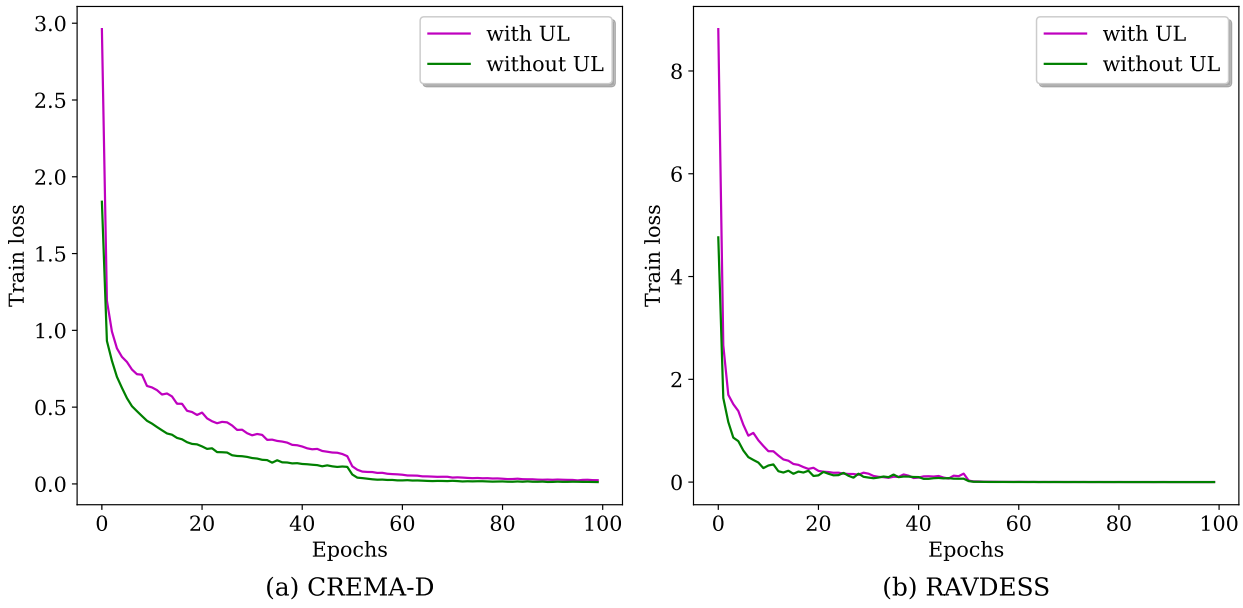



**FIGURE 4.** Loss decay when training with and without the uncertainty-based learning (UL) strategy for (a) CREMA-D and (b) RAVDESS.

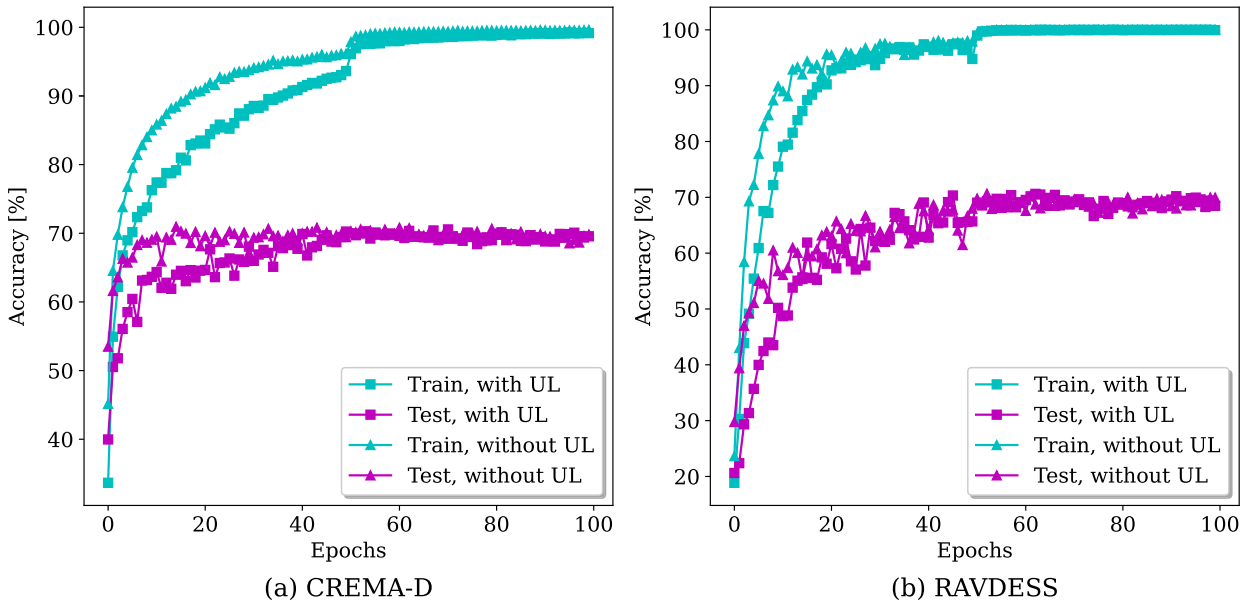
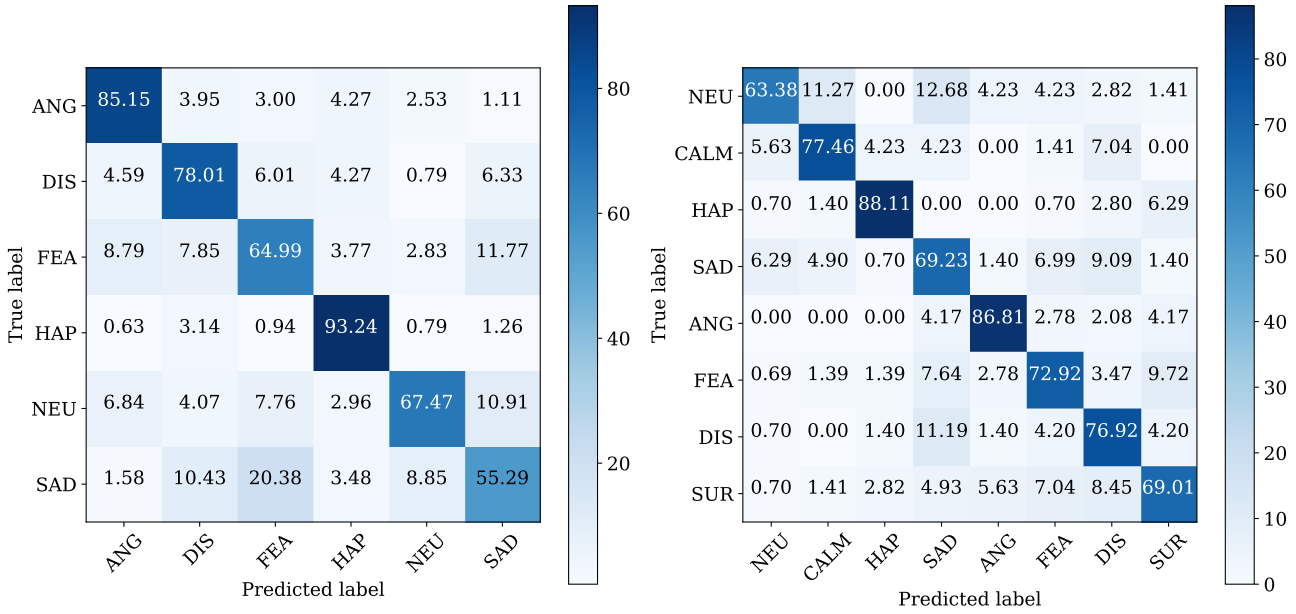



**FIGURE 5.** Overall accuracy on the train and test sets, with and without the uncertainty-based learning (UL) strategy, for (a) CREMA-D and (b) RAVDESS.

happiness (74.2 %), but still below the performance achieved by the proposed method.

Recently, Wav2Vec2.0 models achieved top performance results in numerous speech-related tasks [66]. Fine-tuning pretrained Wav2Vec2.0 models led to improved accuracy in speech emotion recognition [67], i.e., on the RAVDESS dataset, the reported overall accuracy is 84.30 % when a Wav2Vec2.0-FT variant is deployed [68], 86.70 % when a pre-trained xlsr-Wav2Vec2.0 transformer is used to extract speech-related features [69], and 82.75 % when a dual-stream representation and cross-attention fusion based on Wav2Vec2.0 are considered [70]. Inspired by the potential of speech representations based on Wav2Vec2.0, an additional experiment has been conducted by replacing the CNN-based architecture of the AudioNet with a Wav2Vec2.0 architecture for retrieving relevant audio embeddings mapped to each audio signal. The variant used for the AudioNet is the Wav2Vec2.0 base model that has




(a) CREMA-D (b) RAVDESS

**FIGURE 6.** Confusion matrices for (a) CREMA-D and (b) RAVDESS when the proposed multimodal fusion architecture uses CNN (AN) & CNN (VN). The category label is abbreviated through the first three letters (i.e., ANG stands for *anger*, FEA stands for *fear*).

been initially pretrained on the LibriSpeech dataset, which is a corpus of approximately 1000 hours of 16kHz read English speech [71]. The pretrained Wav2Vec2.0 model is publicly available online [72]. The model is fine-tuned for the multimodal emotion recognition task using the same training strategy as for the CNN-based architecture proposed in Section III. The overall accuracy values are reported in Table 3. In the case of CREMA-D, the performance results are close to the ones achieved by the proposed multimodal fusion approach that used CNN-based models for retrieving both audio and visual features. An improvement can be observed in the case of the RAVDESS dataset, for which the performance achieved when using Wav2Vec2.0 as AudioNet is similar to human performance [11]. However, the number of parameters is 36 times higher than the model proposed in Section III. This fact limits the possibility of developing emotion recognition applications based on the Wav2Vec2.0 model on embedded platforms.

### 3) NUMBER OF PARAMETERS

The main advantage of the CNN-based multimodal approach proposed in this paper resides in the small number of parameters of the overall architecture. Specifically, the convolutional blocks for feature extraction and fully connected layers have the same weights across all temporal segments, which leads to a reduction in the number of parameters for the proposed approach.

Table 4 lists the number of parameters associated to various architectures included in Table 3. Attention-based networks, and especially transformers, involve a considerably larger number of parameters than the proposed approach, e.g. 36 times higher when using the Wav2Vec2.0 architecture instead of the CNN-based model for the AudioNet. Consequently, although achieving a high level of accuracy when fine-tuning a pre-trained Wav2Vec2.0 model for the AudioNet, the large number of parameters, i.e., almost 97.2 million parameters, limits the integration of this architecture on embedded platforms. Similarly, the TA-AVN architecture that our team proposed in [51] consists of 18.7 million parameters when 12 temporal segments are considered and the number of parameters linearly increases with the number of temporal segments since each temporal segment is analyzed by different audio and visual neural networks.

Using the same audio and visual neural network modules (i.e., same weight values) across all the temporal segments leads to obtaining a constant number of parameters irrespective the number of temporal segments (i.e., approximately 2.7 million parameters as shown in Table 1). As a direct consequence, the small number of parameters and a model size of approximately 10.3 MB (with no additional quantization schemes applied) allow the integration of the proposed architecture on embedded platforms with limited resources (e.g., low memory, low computational power). The number of parameters and, thus the model size, vary only with the number of classes due to changes in the last classification layer, being constant with respect to the number of temporal segments.

**TABLE 3. Comparisons with other methods.**

| Dataset | Model | Validation method | Accuracy [%] |
|---|---|---|---|
| CREMA-D | Human performance [10] | - | 63.6 |
| | Recursive attention [31] | not mentioned | 65.0 |
| | MATER [45] | 10 folds | 67.2 |
| | MERML [12] | 10 folds | 66.5 |
| | MulT Base [48] | random split train 60% validation 20% test 20% | 68.9 |
| | MulT Large [48] | random split train 60% validation 20% test 20% | 70.2 |
| | **Proposed multimodal fusion pretrained Wav2Vec2.0 (AN) & CNN (VN)** | 5 folds | **71.4** |
| | TA-AVN [51] | 5 folds | 71.4 |
| | AuxFormer [49] | random split train 70% validation 15% test 15% | 71.7 |
| | Multimodal Transformers [16] | not mentioned | 72.5 |
| | Deep metric learning [32] | 10 folds | 74.0 |
| | **Proposed multimodal fusion CNN (AN) & CNN (VN)** | 5 folds | **74.2** |
| RAVDESS | Human performance [11] | - | 80.0 |
| | Recursive attention [31] | not mentioned | 58.3 |
| | TA-AVN [51] | 5 folds | 63.1 |
| | Deep metric learning [32] | 10 folds | 67.7 |
| | MMTM [38] | not mentioned | 73.1 |
| | MSAF [43] | 6 folds | 74.9 |
| | CFN-SR [44] | 5 folds | 75.8 |
| | MATER [45] | 12 folds | 76.3 |
| | **Proposed multimodal fusion CNN (AN) & CNN (VN)** | 5 folds | **76.3** |
| | **Proposed multimodal fusion pretrained Wav2Vec2.0 (AN) & CNN (VN)** | 5 folds | **81.9** |

**TABLE 4. Number of parameters.**

| Model | # Params |
|---|---|
| **Proposed multimodal fusion CNN (AN) & CNN (VN)** | **2.70 M** |
| TA-AVN (12 segments) [51] | 18.70 M |
| MSAF [43] | 25.94 M |
| CFN-SR [44] | 26.30 M |
| MMTM [38] | 31.97 M |
| MulT Base [48] | 38.30 M |
| MulT Large [48] | 89.20 M |
| **Proposed multimodal fusion pretrained Wav2Vec2.0 (AN) & CNN (VN)** | 97.20 M |

#### 4) INFERENCE TIME

Apart from the number of parameters and, thus, model size, inference time is a key aspect when designing real-time systems. The inference time was measured for a single instance (i.e., audio and video), which was passed throughout the proposed architecture. The results are presented in Table 5.

**TABLE 5. Inference time on various NVIDIA platforms(s).**

| Dataset | RTX 3080 Ti | RTX 4090 | Jetson Nano |
|---|---|---|---|
| CREMA-D | $5.1 \cdot 10^{-3}$ s | $1.8 \cdot 10^{-3}$ s | 0.44 s |
| RAVDESS | $13.1 \cdot 10^{-3}$ s | $4.1 \cdot 10^{-3}$ s | 0.84 s |

Furthermore, the audio and visual embeddings extracted using the two corresponding convolutional neural networks are invariant with respect to time shifts at temporal segment level. The processing of the streaming multimodal data can be optimized through the sliding window approach shown in Figure 2. This approach yields a decrease of the time needed to analyze $N$ temporal segments containing audio and visual information since the extraction of features can be performed only once for each temporal segment, whilst the temporal aggregation is performed by simply averaging the outputs and applying a Softmax function.

## VI. CONCLUSION

In this paper, we present a new and robust lightweight multimodal emotion recognition framework, which allows the processing of streaming multimodal data in a real-time setting. With a lightweight structure containing a small number of parameters and currently not using any quantization technique [73], the proposed approach represents an optimal candidate for the integration of the multimodal emotion recognition framework on embedded platforms and can leverage the opportunity of parallelization of the entire emotion recognition framework. Specifically, in a real-time setting, where the multimodal data needs a continuous processing, the proposed framework allows a single-time analysis of the multimodal information per temporal segment since the same neural networks are used to extract audio and visual features. Similarly, the fully connected layer, which is the same for all temporal segments, provides an indication of the short-term emotional state of the subject based on the multimodal information extracted over each temporal segment. The temporal information is aggregated by simply averaging the outputs of the fully connected layers over multiple temporal segments. As a direct consequence of using the same weights across multiple temporal segments, the model size is greatly reduced if compared to other frameworks that analyze the information extracted from different temporal segments with different neural networks. The proposed framework is trained end-to-end using an uncertainty-based learning approach that includes the most uncertain samples in the training loop. The experimental results show that such an approach requires a smaller amount of annotated training data at each epoch, while maintaining a high performance level. Moreover, the training procedure encompasses a natural augmentation of the training dataset through a random selection of the video frames and associated speech signals within a given temporal segment.

## REFERENCES


[1] F. Cavallo, F. Semeraro, L. Fiorini, G. Magyar, P. Sinčák, and P. Dario, "Emotion modelling for social robotics applications: A review," *J. Bionic Eng.*, vol. 15, no. 2, pp. 185–203, Mar. 2018.

[2] M. Dhuheir, A. Albaseer, E. Baccour, A. Erbad, M. Abdallah, and M. Hamdi, "Emotion recognition for healthcare surveillance systems using neural networks: A survey," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Jun. 2021, pp. 681–687.

[3] R. Guo, H. Guo, L. Wang, M. Chen, D. Yang, and B. Li, "Development and application of emotion recognition technology—A systematic literature review," *BMC Psychol.*, vol. 12, no. 1, p. 95, Feb. 2024.

[4] M. Płaza, R. Kazała, Z. Koruba, M. Kozłowski, M. Lucińska, K. Sitek, and J. Spyrka, "Emotion recognition method for call/contact centre systems," *Appl. Sci.*, vol. 12, no. 21, p. 10951, Oct. 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/21/10951

[5] S. Cosentino, E. I. S. Randria, J.-Y. Lin, T. Pellegrini, S. Sessa, and A. Takanishi, "Group emotion recognition strategies for entertainment robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 813–818.

[6] V. Bakir and A. McStay, "Empathic media, emotional AI, and the optimization of disinformation," in *Affective Politics of Digital Media*. Evanston, IL, USA: Routledge, 2020, pp. 263–279.

[7] L. Podoletz, "We have to talk about emotional AI and crime," *AI Soc.*, vol. 38, no. 3, pp. 1067–1082, May 2022, doi: 10.1007/s00146-022-01435-w.

[8] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, Oct. 2013, pp. 3–10.

[9] N.-C. Ristea, L. C. Dutu, and A. Radoi, "Emotion recognition system from speech and visual information based on convolutional neural networks," in *Proc. Int. Conf. Speech Technol. Human-Computer Dialogue (SpeD)*, Oct. 2019, pp. 1–6.

[10] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct. 2014.

[11] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391.

[12] E. Ghaleb, M. Popa, and S. Asteriadis, "Metric learning-based multimodal audio-visual emotion recognition," *IEEE Multimedia*, vol. 27, no. 1, pp. 37–48, Jan./Mar. 2020.

[13] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2010, pp. 1459–1462, doi: 10.1145/1873951.1874246.

[14] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97515–97525, 2019.

[15] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[16] V. John and Y. Kawanishi, "Audio and video-based emotion recognition using multimodal transformers," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 2582–2588.

[17] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7169–7173.

[18] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[19] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.

[20] P. Ekman, "An argument for basic emotions," *Cognition Emotion*, vol. 6, nos. 3–4, pp. 169–200, May 1992.

[21] P. Ekman, "Facial expression and emotion," *Amer. Psychologist*, vol. 48, no. 4, p. 384, 1993.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[25] P. Burkert, F. Trier, M. Zeshan Afzal, A. Dengel, and M. Liwicki, "DeXpression: Deep convolutional neural network for expression recognition," 2015, *arXiv:1509.05371*.

[26] C. Dixit and S. M. Satapathy, "A customizable framework for multimodal emotion recognition using ensemble of deep neural network models," *Multimedia Syst.*, vol. 29, no. 6, pp. 3151–3168, Oct. 2023, doi: 10.1007/s00530-023-01188-6.

[27] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ACM Int. Conf. Multimodal Interact.* New York, NY, USA: Association for Computing Machinery, Nov. 2015, pp. 435–442, doi: 10.1145/2818346.2830595.

[28] T. N. Rincy and R. Gupta, "Ensemble learning techniques and its efficiency in machine learning: A survey," in *Proc. 2nd Int. Conf. Data, Eng. Appl. (IDEA)*, Feb. 2020, pp. 1–6.

[29] S. Akbar, A. Raza, T. A. Shloul, A. Ahmad, A. Saeed, Y. Y. Ghadi, O. Mamyrbayev, and E. Tag-Eldin, "PAtbP-EnC: Identifying anti-tubercular peptides using multi-feature representation and genetic algorithm-based deep ensemble model," *IEEE Access*, vol. 11, pp. 137099–137114, 2023.

[30] D. Valles and R. Matin, "An audio processing approach using ensemble learning for speech-emotion recognition for children with ASD," in *Proc. IEEE World AI IoT Congr. (AIIoT)*, May 2021, pp. 55–61.

[31] R. Beard, R. Das, R. W. M. Ng, P. G. K. Gopalakrishnan, L. Eerens, P. Swietojanski, and O. Miksik, "Multi-modal sequence fusion via recursive attention for emotion recognition," in *Proc. 22nd Conf. Comput. Natural Lang. Learn.*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 251–259.

[32] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2019, pp. 552–558.

[33] M. A. Razzaq, J. Hussain, J. Bang, C.-H. Hua, F. A. Satti, U. U. Rehman, H. S. M. Bilal, S. T. Kim, and S. Lee, "A hybrid multimodal emotion recognition framework for UX evaluation using generalized mixture functions," *Sensors*, vol. 23, no. 9, p. 4373, Apr. 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/9/4373

[34] S. E. Eskimez, Y. Zhang, and Z. Duan, "Speech driven talking face generation from a single image and an emotion condition," *IEEE Trans. Multimedia*, vol. 24, pp. 3480–3490, 2022.

[35] P. Bhattacharya, R. K. Gupta, and Y. Yang, "Exploring the contextual factors affecting multimodal emotion recognition in videos," *IEEE Trans. Affective Comput.*, vol. 14, no. 2, pp. 1547–1557, Apr./Jun. 2023, doi: 10.1109/TAFFC.2021.3071503.

[36] 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, doi: 10.1145/3340555.3355713.

[37] J. D. S. Ortega, M. Senoussaoui, E. Granger, M. Pedersoli, P. Cardinal, and A. L. Koerich, "Multimodal fusion with deep neural networks for audio-video emotion recognition," 2019, *arXiv:1907.03196*.

[38] H. R. Vaezi Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: Multimodal transfer module for CNN fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13286–13296. [Online]. Available: https://api.semanticscholar.org/CorpusID:208176099

[39] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jul. 2017, pp. 4724–4733, doi: 10.1109/CVPR.2017.502.

[40] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2016, pp. 892–900.

[41] P. Barros, G. I. Parisi, and S. Wermter, "A personalized affective memory neural model for improving emotion recognition," 2019, *arXiv:1904.12632*.

[42] G. Cioroiu and A. Radoi, "Multimodal emotion recognition with attention," in *Proc. Int. Symp. Signals, Circuits Syst. (ISSCS)*, Jul. 2023, pp. 1–4.
[43] L. Su, C. Hu, G. Li, and D. Cao, "MSAF: Multimodal split attention fusion," 2020, *arXiv:2012.07175*.
[44] Z. Fu, F. Liu, H. Wang, J. Qi, X. Fu, A. Zhou, and Z. Li, "A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition," 2021, *arXiv:2111.02172*.
[45] E. Ghaleb, J. Niehues, and S. Asteriadis, "Multimodal attention-mechanism for temporal emotion recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 251–255.
[46] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 131–135.
[47] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[48] M. Tran and M. Soleymani, "A pre-trained audio-visual transformer for emotion recognition," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 4698–4702.
[49] L. Goncalves and C. Busso, "AuxFormer: Robust approach to audiovisual emotion recognition," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7357–7361.
[50] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "EPIC-fusion: Audio-visual temporal binding for egocentric action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5491–5500.
[51] A. Radoi, A. Birhala, N.-C. Ristea, and L.-C. Dutu, "An end-to-end emotion recognition framework based on temporal aggregation of multimodal information," *IEEE Access*, vol. 9, pp. 135559–135570, 2021.
[52] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). Berlin, Germany: Springer-Verlag, 2006.
[53] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html
[54] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[55] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[56] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Deep representation learning in speech processing: Challenges, recent advances, and future trends," 2020, *arXiv:2001.00378*.
[57] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.
[58] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.
[59] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, Nov. 1977.
[60] J. Gibson, M. V. Segbroeck, and S. S. Narayanan, "Comparing time-frequency representations for directional derivative features," in *Proc. Interspeech*, Singapore, 2014, pp. 612–615. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2014/i14_0612.html
[61] D. O'Shaughnessy, *Speech Communications—Human and Machine*, 2nd ed., New York, NY, USA: Wiley-IEEE, 2000.
[62] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2005, ch. 2, pp. 13–55, doi: 10.1002/047174882X.ch2.
[63] A. Birhala, C. N. Ristea, A. Radoi, and L. C. Dutu, "Temporal aggregation of audio-visual modalities for emotion recognition," in *Proc. 43rd Int. Conf. Telecommun. Signal Process. (TSP)*, Jul. 2020, pp. 305–308.
[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.
[65] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[66] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12449–12460.
[67] L.-W. Chen and A. Rudnicky, "Exploring Wav2vec 2.0 fine tuning for improved speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[68] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. Interspeech*, Aug. 2021, pp. 3400–3404.
[69] C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J. M. Montero, and F. Fernández-Martínez, "A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset," *Appl. Sci.*, vol. 12, no. 1, p. 327, Dec. 2021. [Online]. Available: https://www.mdpi.com/2076-3417/12/1/327
[70] S. Yu, J. Meng, W. Fan, Y. Chen, B. Zhu, H. Yu, Y. Xie, and Q. Sun, "Speech emotion recognition using dual-stream representation and cross-attention fusion," *Electronics*, vol. 13, no. 11, p. 2191, Jun. 2024. [Online]. Available: https://www.mdpi.com/2079-9292/13/11/2191
[71] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.
[72] *Hugging Face*. Accessed: Feb. 7, 2024. [Online]. Available: https://huggingface.co/facebook/wav2vec2-base
[73] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2704–2713.

**ANAMARIA RADOI** (Member, IEEE) received the B.Sc. degree in electronics and telecommunications from the University Politehnica of Bucharest, Romania, in 2010, the M.Sc. degree in communication systems from the Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland, in 2012, and the Ph.D. degree in electronics and telecommunications from the University Politehnica of Bucharest, in 2015. She is currently working as an Associate Professor with the Department of Applied Electronics and Information Engineering, NUST Politehnica Bucharest. She is also a Principal Investigator in one project dealing with the development of multimodal emotion recognition frameworks integrated on embedded systems. Previously, she has been a Principal Investigator in projects focusing on multimodal data analysis and time series analysis in the remote sensing domain. Her research interests include signal and image processing, machine learning, pattern recognition, time series analysis, and multimodal data interpretation. She regularly acts as an External Technical Expert of the Horizon Europe and Horizon 2020 Programs and as a reviewer of various IEEE publications.

**GEORGE CIOROIU** received the B.Sc. degree from the Faculty of Automatic Control and Computers, University Politehnica of Bucharest, in 2017, and the M.Sc. degree from the Master Program Complex Signal Processing in Multimedia Applications, University Politehnica of Bucharest, in 2019. He is currently pursuing the Ph.D. degree with a focus on developing multimodal emotion recognition systems integrated on platforms with limited resources. His research interests include deep learning, learning representation, diffusion, distillation, and state space models.