




## Article

# Enhancing Emotion Recognition through Federated Learning: A Multimodal Approach with Convolutional Neural Networks

Nikola Simić <sup>1,\*</sup> , Siniša Suzić <sup>1</sup>, Nemanja Milošević <sup>2</sup>, Vuk Stanojev <sup>1</sup> , Tijana Nosek <sup>1</sup>, Branislav Popović <sup>1</sup>  and Dragana Bajović <sup>1</sup>

<sup>1</sup> Faculty of Technical Sciences, University of Novi Sad, 21000 Novi Sad, Serbia; sinisa.suzic@uns.ac.rs (S.S.); vukst@uns.ac.rs (V.S.); tijana.nosek@uns.ac.rs (T.N.); bpopovic@uns.ac.rs (B.P.); dbajovic@uns.ac.rs (D.B.)

<sup>2</sup> Faculty of Sciences, University of Novi Sad, 21000 Novi Sad, Serbia; nmilosev@dmf.uns.ac.rs

\* Correspondence: nikolasimic@uns.ac.rs

**Abstract:** Human–machine interaction covers a range of applications in which machines should understand humans’ commands and predict their behavior. Humans commonly change their mood over time, which affects the way we interact, particularly by changing speech style and facial expressions. As interaction requires quick decisions, low latency is critical for real-time processing. Edge devices, strategically placed near the data source, minimize processing time, enabling real-time decision-making. Edge computing allows us to process data locally, thus reducing the need to send sensitive information further through the network. Despite the wide adoption of audio-only, video-only, and multimodal emotion recognition systems, there is a research gap in terms of analyzing lightweight models and solving privacy challenges to improve model performance. This motivated us to develop a privacy-preserving, lightweight, CNN-based (CNNs are frequently used for processing audio and video modalities) audiovisual emotion recognition model, deployable on constrained edge devices. The model is further paired with a federated learning protocol to preserve the privacy of local clients on edge devices and improve detection accuracy. The results show that the adoption of federated learning improved classification accuracy by ~2%, as well as that the proposed federated learning-based model provides competitive performance compared to other baseline audiovisual emotion recognition models.

**Keywords:** artificial intelligence; emotion recognition; federated learning; machine learning; multimodal



**Citation:** Simić, N.; Suzić, S.; Milošević, N.; Stanojev, V.; Nosek, T.; Popović, B.; Bajović, D. Enhancing Emotion Recognition through Federated Learning: A Multimodal Approach with Convolutional Neural Networks. *Appl. Sci.* **2024**, *14*, 1325. <https://doi.org/10.3390/app14041325>

Academic Editor: Douglas O’Shaughnessy

Received: 31 December 2023

Revised: 21 January 2024

Accepted: 23 January 2024

Published: 6 February 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Speech is commonly referred to as the most natural way of human-to-human communication [1]. As humans, we can distinguish how others feel based on their voices. Emotion detection affects our interpretation of the spoken content, our behavior, and the consequent actions. Emotions do not only affect our speech. They also affect our mood, facial expressions, medical features, and body gestures. Based on someone’s facial expressions, people can often recognize how others are feeling, even without speech content in the case of non-verbal communication. Although people can recognize emotions only from speech or only from facial expressions, commonly, both of these modalities are naturally used for overall recognition [2]. Similar concepts are employed in order to design state-of-the-art systems for human–computer interaction.

Automatic emotion recognition has been a research area for more than two decades, and in that time frame, numerous methods based on machine learning (ML) and deep learning (DL) have been proposed [3–5]. In the beginning, unimodal speech emotion recognition and facial emotion recognition systems were developed due to hardware limitations and the lack of available data for creating complex models. Some of the most preferred speech emotion recognition algorithms throughout the years have been based on the hidden Markov model (HMM), Gaussian mixture models (GMMs), support vector

machines (SVMs), and artificial neural networks (ANNs). Other classification techniques usually involve methods based on decision trees (DTs), k-nearest neighbors (k-NN), k-means, as well as Naive Bayes classifiers. An overview of the existing methodologies can be found in [6]. Recent efforts have been directed toward the implementation of convolutional neural networks (CNNs), deep learning, and transformer-based networks, which commonly require speech samples to be preprocessed and converted into spectrograms [7,8]. Besides spectrograms that are obtained using Fourier transform, scalograms can be also found in the literature as a visual representation of speech in the case of applied wavelet transform for different tasks involving audio signal processing and classification using convolutional neural networks [9]. F. Andayani et al. proposed a combination of a long short-term memory (LSTM) and a transformer encoder network to learn the long-term dependencies in speech signals and classify emotions [8].

Similar classifiers have been considered for facial emotion recognition, whereby CNNs are the prevailing choice [10]. Besides audio and video modalities, an analysis of brain electroencephalography (EEG) signals, collected using wearable devices, also attracted researchers' interest [11].

Transformer-based architectures are considered cutting-edge in the field of sequence modeling, and they have achieved remarkable success in various domains. However, their effectiveness is reached at the cost of quadratic computational and memory complexity [12] and in turn they require higher energy consumption compared to traditional machine learning techniques, making them challenging for implementation on edge devices [13]. On the other hand, sentiment analysis is considered a sensitive topic and there are privacy concerns related to data processing, such as potential user profiling, making edge-based applications desirable for detection. The technology proposed in this paper is developed within the MARVEL project, whose ubiquitous edge-to-fog-to-cloud architecture and solutions are primarily designed for smart city environments [14].

In recent research, there has been a shift from unimodal to multimodal approaches, with a strong focus on designing audiovisual models to ensure higher accuracy. In [15], the authors proposed an end-to-end network that incorporates LSTM besides a CNN. In [16], the authors exploit an attention mechanism to improve the efficacy of the DL network, whereas in [17], a multimodal emotion recognition metric learning is introduced. A correlation-based graph convolutional network (C-GCN) for audiovisual emotion recognition task is introduced in [18]. Reference [19] introduces an audio-visual fusion model of deep learning features with a Mixture of Brain Emotional Learning. In this method, a CNN as well as recurrent neural network (RNN) were employed.

So far, not much attention has been given to the generalization and significance of incorporating federated learning procedures. Motivated by the COVID-19 pandemic, Chikara et al. [20] made a first step in this direction by introducing a framework for monitoring the emotional state of an individual without sending data to a centralized server. The latter was achieved by incorporating federated learning techniques based on weighted averaging and utilizing both audio and video modalities. The principal goal of their research was to predict the state of depression during the post-processing phase, after classifying emotions based on the outputs from the audio and video modalities. To process video signals, Chikara et al. employed a convolutional neural network. On the other hand, an ensemble of seven machine learning classifiers has been utilized to process audio signals. In the end, they briefly described the procedure for combining these modalities and employing federated learning, without providing multimodal results to validate the approach.

Another multimodal emotion recognition model, accompanied by federated learning, was proposed by Nandi et al. [21]. They introduced a federated learning method for real-time classification of emotional states from multimodal streaming. Their focus was primarily on utilizing physiological data captured from wearable sensors. Apart from multimodal approaches from [20,21], which addressed specific objectives, there is only a limited number of federated learning-based methods for unimodal emotion detection

systems that consider either the video modality [22] or the audio modality [23,24]. Our approach represents a step further in the field of privacy-preserving emotion recognition, as we conduct an experiment focused on the application of federated learning to the multimodal audiovisual emotion recognition task, considering relatively small classification models that can be deployable at the edge.

### 1.1. Motivation and Contributions

The aim of this research is to design a privacy-preserving audiovisual emotion recognition model (AVER), capable of classifying emotions at the edge and supporting continuous model updates within a decentralized system. Unlike the model proposed in [20], we do not exploit an ensemble of classifiers for processing audio signals but rather implement a convolutional neural network on the audio modality as well. This way, post-processing could be simplified, as we do not need to train several ML classifiers, analyze their performance, and choose a subset of classifiers for decision-making. Furthermore, we exploit transfer learning and fine-tuning for detecting facial expressions, since training convolutional neural networks on a small set of data cannot guarantee satisfactory generalization. To validate the proposed approach, we perform experiments on the eNTERFACE'05 dataset and provide classification accuracy, F1-score, recall, and precision results obtained using multimodal data processing.

One of the key aspects to consider in this context is the availability of training data and the composition of a set of speakers. There are two commonly discussed groups of emotion recognition models: speaker-independent and speaker-dependent [25]. These groups are related to different problems in the field. The aim of speaker-independent emotion recognition models is to generalize well and accurately predict the emotions of speakers that have not been seen before during the model training, i.e., to operate on the open-set of speakers. However, due to the lack of available datasets with accurate annotations, current models face challenges in achieving very high detection accuracy as a general solution. On the other hand, speaker-dependent emotion recognition systems require being trained for each user, maximizing the performance of the system for a closed set of speakers. This way, systems could be more robust and provide higher performance in an industrial environment compared to the general speaker-independent models. This approach is beneficial for applications designed for specific individuals, such as voice-based personal assistants or speech-based authentication systems. However, creating models for such scenarios requires recording individuals of interest, which can be time-consuming and consequently may involve acting emotions. This motivated us to make an effort to improve the overall accuracy of such models by incorporating federated learning in a closed-set scenario in order to leverage different data sources and minimize biases produced due to small training sets.

The rest of this paper is organized as follows: In Section 2.1, we provide a description of the proposed audiovisual emotion recognition model. Section 2.2 provides an overview of federated averaging, which accompanies the model described in Section 2.1. The experiment is described in Section 2.3 while results are presented in Section 3. Comparison with other state-of-the-art methods is provided in Section 4. Finally, the advantages and disadvantages of the proposed model are summarized and discussed in Section 5.

## 2. Materials and Methods

### 2.1. Audiovisual Emotion Recognition Model

In our study, a multimodal model is proposed, designed to process signals from two different modalities: facial expressions, given as a sequence of images, and speech, given as a set of image spectrograms created in the pre-processing phase. To process facial expressions, a MobileNetV2 model is utilized [26], pre-trained on the ImageNet [27], a large-scale dataset widely used for image classification tasks. The idea of transfer learning is introduced due to the fact that available audiovisual emotion-capturing datasets are relatively small, and it is often impossible or at least highly impractical to record large

datasets for speaker-dependent emotion recognition tasks. To adapt the MobileNetV2 model for our specific task (described further ahead in Section 2.3), we introduce a new prediction layer consisting of six classes corresponding to different facial expressions.

While incorporating key frame detection could potentially enhance the overall accuracy of emotion recognition in video signals [28], this approach is computationally intensive, which contradicts our objective of creating a simple model suitable for deployment on edge devices. Consequently, we made the decision to proceed with a uniform extraction of frames from the video. In the pre-processing phase, we extracted frames from the original videos with a step size of 8 and resized each frame to the resolution of  $224 \times 224$  pixels, so that our inputs would be compatible with ImageNet resolution used in MobileNetV2. Considering the chosen input shape and the aforementioned output layer, we designed a neural network model for analyzing facial expressions, which comprised approximately 2.63 million parameters. The overall size of the model is approximately 10.3 MB, with 0.37 million trainable parameters. Since the number of trainable parameters is relatively small and the impact of federated learning on such a small parameter set yielded negligible changes in our experiments, we opted to further simplify the proposed system by applying federated learning to the audio (i.e., speech) modality only. By focusing on federated learning solely for the audio modality, we aimed to simplify the system and reduce the computational requirements, aligning with our objective of creating an efficient model suitable for edge devices.

For the task of emotion detection from the audio modality, we propose a model inspired by the network from [29]. However, we simplify the method by excluding one fully connected layer from the 2D CNN network, reducing its complexity. The proposed model is described in Table 1.

**Table 1.** The proposed CNN model for speech emotion recognition.

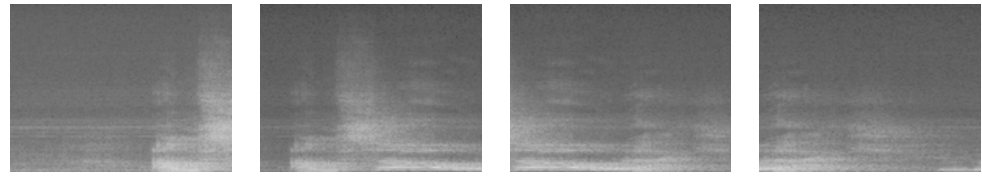
Layer	Arguments	Number of Parameters
Convolution2D	Filters = 64; kernel size = (7, 7); stride = (2, 2); input shape = (128, 170, 1)	3200
MaxPooling2D		
Convolution2D	Filters = 128; kernel size = (7, 7); stride = (2, 2)	401,536
AveragePooling2D		
Convolution2D	Filters = 256; kernel size = (3, 3); stride = (2, 2)	295,168
AveragePooling2D		
Convolution2D	Filters = 512; kernel size = (3, 3); stride = (2, 2)	1,180,160
Flatten		
Dense_1	Nodes = 4096	4,198,400
Dropout	Rate = 0.5	
Dense_2	Nodes = 6	24,582
Total number of parameters		6,103,046

Unlike the method proposed in [29], we do not utilize MFCC (Mel frequency cepstral coefficients) as inputs. Instead, we adopt the approach of generating spectrograms using short-time Fourier transform (STFT), following the methodology described in [30].

We decided to choose spectrograms, as their computing is straightforward and there are efficient algorithms and libraries for further deployment, whereas scalogram computation might be more computationally intensive. Recent studies have shown that spectrograms and scalograms lead to the similar performance on the speech emotion recognition tasks [9], so we decided to continue with the STFT approach.

In the first step of the spectrogram creation, a speech file is divided into 1.0 s long segments with an overlap of 0.5 s as shown in Figure 1. If the last segment is shorter than 0.7 s, it is expelled from the training set. If the last segment's length falls between 0.7 and 1.0 s, it is extended to a 1.0 s length by replicating samples from the end of that segment. Spectrograms are calculated using an 11.6 ms Hanning window with a time shift of 5.8 ms.

The obtained STFT coefficients are then converted to PNG image of size  $128 \times 170$  pixels. It was demonstrated in the literature that Hanning and Hamming windowing functions usually provide better performance than the other in the case of speech emotion recognition task [31]. A similar approach of using spectrograms of fixed size can be found in the literature as a common choice for processing speech in the case of speaker verification [32].

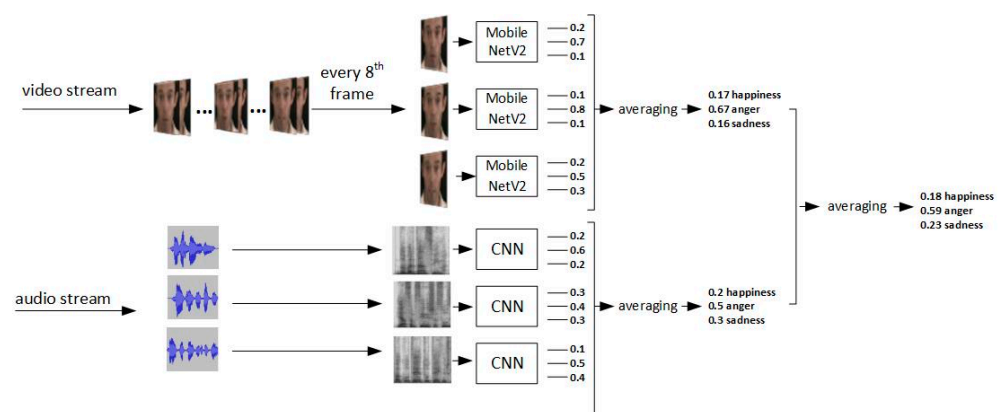


**Figure 1.** Example of spectrograms extracted from a single speech stream.

The final speech processing model consists of approximately 6.1 million trainable parameters, resulting in a size of around 23.3 MB. We also experimented with larger VGGish-like architectures [33,34], but found that their performance on the small client subsets was either similar or inferior.

In order to maximize classification accuracy at the utterance level, we apply a sequential voting procedure in the post-processing phase for both modalities, inspired by the approach described in [33]. Similar approaches for video processing are already available in the literature [35], simplifying the overall complexity, avoiding a need to utilize LSTM or 3D-CNN for a sequence of frames. In the case of facial expression analysis, after training a classifier to make decisions for individual input video frames, we buffer the outputs for all frames corresponding to a single utterance and make an utterance-level decision using average probability voting. Similarly, for audio signals, neural networks are trained to make decisions for individual 1.0 s long spectrograms. The outputs corresponding to a single utterance are buffered, and average probability voting is applied. Previously, we found that the average probability voting is a preferable choice comparing to the maximum probability voting and the majority voting for a scenario related to the CNN-based speech emotion recognition [33].

The outputs from both modalities are fused at the decision level, and the final decision is made by averaging scores of these two outputs. The decision fusion process allows us to leverage the complementary information from both facial expressions and speech for a more robust emotion recognition decision. The architecture of the proposed audiovisual emotion recognition model is presented in Figure 2, with an accompanied inference example. Without loss of generality, we present an example where three different classes are analyzed, whereas the outputs for other classes tend to be zero. In the next subsection, we explain the federated learning protocol applied to the audio modality of the proposed audiovisual emotion recognition model.



**Figure 2.** The architecture of the proposed audiovisual emotion recognition model with an inference example.



## 2.2. Federated Learning and Averaging

The purpose of federated learning is to enable the training of machine learning models in a decentralized manner while preserving data privacy. Federated learning aims to leverage the collective knowledge from multiple devices or clients without requiring them to share their raw data with a central fog or cloud server. In a typical federated learning setup, a large number of client devices, such as smartphones or IoT devices, participate in the training process [36]. Each client holds its local dataset, which may contain sensitive or private information. Instead of uploading their data to a central server, clients collaborate by sharing model updates. This approach helps to overcome data privacy concerns and reduces the need for a large-scale data transfer, as only model updates are communicated between clients and the central server.

The popularity of this technique started after the introduction of the federated averaging (FedAvg) algorithm proposed by Google's researchers in 2016 [37]. If we consider that  $N$  clients are indexed by  $i$ , the fraction of clients that perform each round is  $F$ , the local minibatch size is  $B$ , the number of local epochs is  $M$ , and the learning rate is  $\eta$ , the FedAvg algorithm could be defined using the following steps [37]:

- (1) Initialization: a global model is initialized on a central server (initialize  $w_0$ ).
- (2) Client selection: a subset  $S_t$  of  $\max(F \times N, 1)$  clients is randomly or strategically selected for participation in each round of training.
- (3) Model distribution: The current global model is sent to the selected clients in parallel.

For each client  $i \in S_t$  in parallel :

$$w_{t+1}^i \leftarrow \text{ClientUpdate}(i, w_t)$$

$$w_{t+1} \leftarrow \sum_{i=1}^N \frac{n_i}{n} w_{t+1}^i \quad (1)$$

- (4) Local training: Each client trains the model on its local dataset using the received model parameters. This training can involve multiple local iterations to improve accuracy.

ClientUpdate( $i, w$ ): //run on client  $i$

$B \leftarrow$  (split partition  $P_i$  into batches of size  $B$ )

for each local epoch  $j$  from 1 to  $M$  do

for batch  $b \in B$  do

$$w \leftarrow w - \eta \nabla \ell(w, b) \quad (2)$$

In the previous expression,  $\ell$  represents the loss term of a chosen loss function for training a neural network model, which varies based on the task the model is set up for.

- (5) Model aggregation: After the local training, updated client models are sent back to the central server, which aggregates the models' parameters by computing their average; return  $w$  to server.
- (6) Global model update: The aggregated model becomes the updated global model for the next round of training.

Iterative process: Steps 2–6 are repeated for multiple rounds until convergence is reached, or until a desired performance level is achieved.

The loss function used in both audio and video modalities is the categorical cross-entropy loss function, commonly used in image classification tasks. Since the audio data were pre-processed to visual form (spectrograms), we were able to use the same loss function. Each of our three separate clients owns local weights ( $w$ ), which are unique to the client. These weights represent all trainable model parameters (i.e., layer weights and biases) that local models use. Since the models are trained in a federated fashion, the weights of the local models are also affected by other clients' parameters (global model).

### 2.3. Experimental Setup

We consider the eNTERFACE'05 audiovisual emotion dataset as an example, given within a speaker-dependent scenario described in Section 1.1. The dataset encompasses six distinct emotions: anger, disgust, fear, happiness, sadness, and surprise [38]. It comprises recorded utterances from 42 individuals representing 14 different nationalities. Although the utterances in this dataset were recorded in the English language, the involvement of people from different cultural backgrounds increases the complexity of the experiment as it is well known that emotion semantics show both cultural variations and universal structure [39].

During the experiment, the participants were presented with short stories carefully designed to elicit specific emotions. Subsequently, the participants were required to respond with a predefined set of utterances corresponding to the emotions they experienced. The collected reactions were then evaluated by two experts, who discarded any ambiguous responses.

To facilitate the federated learning experiment and simulate different clients, we divided the eNTERFACE'05 dataset into three separate non-overlapping parts, each one containing an equal number of speakers. This means that each client consisted of utterances recorded by 14 different speakers. The aim of such an experimental setup is to emulate distinct data sources or clients in a federated learning scenario by leveraging diverse speaker data from multiple sources.

### 3. Experimental Results

In this section, we provide detailed classification results obtained using the proposed audiovisual emotion recognition model, specifically observing the influence of implementing the federated averaging described in Section 2.2. Firstly, let us observe the performance of the proposed model without the accompanied federated learning considering the experimental setup described in Section 2.3. The results of classification accuracy are presented in Table 2 whereas detailed results of unimodal video and audio classifiers, without applied sequential voting strategy, are presented in Tables 3 and 4, respectively.

**Table 2.** Classification accuracies of the proposed AVER model without federated learning.

	Classification Accuracy (%)				Multimodal
	Audio		Video		
	Single Spectrogram	Voting	Single Frame	Voting	
Client 1	45.95	57.14	82.04	88.10	89.29
Client 2	48.44	61.90	78.70	85.71	91.66
Client 3	44.08	55.95	87.94	91.66	92.86

It can be noticed that the proposed model for the video modality provides relatively good performance, achieving classification accuracy in the range of 78.70–87.94 (%) on a single frame of test video signals. Similarly, we can notice from Table 3 that the achieved average precision, recall and F1-score are relatively high, following the results of classification accuracy. However, detailed analysis leads us to the observation that there are differences among clients and among emotions. The best performance is achieved in the case of Client 3, whereas classification accuracy in the case of Client 2 is about 9% less in the case of a single spectrogram.

After applying the proposed sequential voting strategy, the accuracy significantly increased. The achieved performance is in the range of 85.71–91.66 (%). As we use a lightweight model with only 0.37 million trainable parameters, pre-trained on the ImageNet, there is not so much space for further improvement in the case of the video modality. Also, the expected tradeoff between possible improvement and an increase in complexity led us to the decision not to implement federated averaging to the video modality.

**Table 3.** Detailed performance–video modality.

Emotion	Client 1			Client 2			Client 3		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Anger	0.91	0.76	0.83	0.84	0.70	0.76	0.97	0.84	0.90
Disgust	0.75	0.92	0.83	0.79	0.78	0.79	0.69	1.00	0.82
Fear	0.85	0.81	0.83	0.89	0.75	0.81	0.97	0.73	0.83
Happiness	0.68	0.95	0.79	0.72	0.87	0.79	0.87	0.81	0.84
Sadness	0.95	0.65	0.77	0.92	0.75	0.82	0.97	0.93	0.95
Surprise	0.82	0.89	0.85	0.65	0.90	0.75	0.90	0.96	0.93
<b>Average</b>	0.83	0.83	0.82	0.80	0.79	0.79	0.89	0.88	0.88
<b>Weighted average</b>	0.84	0.82	0.82	0.80	0.79	0.79	0.90	0.88	0.88

**Table 4.** Detailed performance–audio modality.

Emotion	Client 1			Client 2			Client 3		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Anger	0.54	0.55	0.55	0.60	0.69	0.64	0.55	0.69	0.61
Disgust	0.42	0.48	0.45	0.44	0.34	0.39	0.33	0.38	0.35
Fear	0.45	0.37	0.40	0.43	0.38	0.40	0.27	0.27	0.27
Happiness	0.38	0.42	0.40	0.52	0.47	0.49	0.45	0.41	0.43
Sadness	0.56	0.59	0.57	0.47	0.53	0.50	0.67	0.53	0.59
Surprise	0.34	0.30	0.32	0.41	0.45	0.43	0.35	0.28	0.31
<b>Average</b>	0.45	0.45	0.45	0.48	0.48	0.47	0.43	0.43	0.43
<b>Weighted average</b>	0.46	0.46	0.46	0.48	0.48	0.48	0.44	0.44	0.44

The obtained results for the classification accuracy in the case of the audio modality (Table 2) indicate that the proposed classifier could not generalize well as the one in the case of the video modality. Similar behavior was also observed in the literature [15,17]. By observing Table 4, it can be noticed that the average results for precision, recall, and F1-score are in accordance with the results for classification accuracy. This motivated us to apply federated learning framework to the clients in the case of the audio modality in order to provide better generalization. These results are presented in Tables 5 and 6. The performance of the video modality in Table 6 is omitted as we do not apply federated learning for this modality. For the purpose of multimodal inference, we consider the performance of the video modality from Table 2.

In order to compare the achieved accuracy improvement using the federated learning model, we define classification accuracy gain as an improvement of the proposed federated model over the non-federated model as:

$$Gain_f[\%] = CA_f - CA_{nf}, \quad (3)$$

where  $CA_f$  represents the classification accuracy of the observed client achieved using federated learning, whereas  $CA_{nf}$  represents the classification accuracy of the same client achieved by training the model without client federation. The achieved gains in the case of the audio modality and multimodal inference are presented in Figure 3.

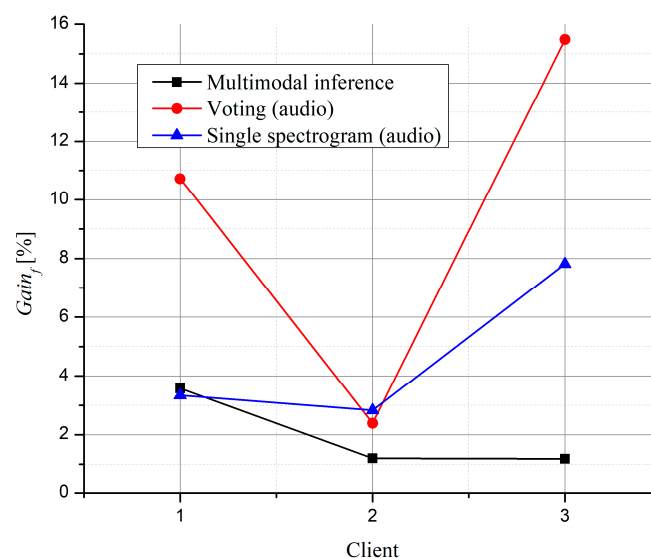


**Table 5.** Detailed performance—audio modality in federated learning setup.

Emotion	Client 1			Client 2			Client 3		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Anger	0.60	0.57	0.59	0.68	0.67	0.67	0.61	0.59	0.60
Disgust	0.44	0.57	0.50	0.45	0.36	0.40	0.50	0.35	0.41
Fear	0.41	0.38	0.39	0.39	0.44	0.41	0.44	0.48	0.46
Happiness	0.43	0.64	0.51	0.58	0.69	0.63	0.69	0.44	0.54
Sadness	0.76	0.46	0.57	0.49	0.55	0.52	0.54	0.69	0.61
Surprise	0.39	0.38	0.38	0.45	0.36	0.40	0.39	0.52	0.45
<b>Average</b>	0.51	0.50	0.49	0.51	0.51	0.51	0.53	0.51	0.51
<b>Weighted average</b>	0.52	0.49	0.50	0.51	0.51	0.51	0.53	0.52	0.52

**Table 6.** Classification accuracy of the proposed AVER model with federated learning.

	Classification Accuracy (%)				Multimodal
	Audio		Video		
	Single Spectrogram	Voting	Single Frame	Voting	
Client 1	49.29	67.86	/	/	92.86
Client 2	51.27	64.29	/	/	92.86
Client 3	51.89	71.43	/	/	94.05

**Figure 3.** Classification accuracy gain  $Gain_f$  [%] =  $CA_f - CA_{nf}$  of the proposed emotion recognition approach accompanied with federated learning, compared to the non-federated learning approach.

As already described in Section 2, we do not apply federated learning to the video modality as we exploit transfer learning and there is only a small number of trainable parameters. By observing Figure 3, we can see that the implementation of federated learning significantly improves classification accuracy in the case of the audio modality, both for making decisions using the single spectrogram (the blue line) and after the sequential voting process (the red line), demonstrating the importance of implementing such a technique. It can also be noticed that there is a gain in the range of 1.19–3.57 (%) in the case of multimodal inference, which is significant, considering that the non-federated learning audiovisual model provides a relatively high classification accuracy in the (already high) range of

89.29–92.86 (%). This is mainly due to the good performance of the classifier in the case of the video modality, whereas the audio modality behaves like a supporting modality within this experiment. However, such behavior could be different when analyzing other datasets, and federated learning application to the video modality may lead to a significant multimodal performance increase.

Finally, we provide comparison of the average values of precision, recall, and F1-score in the case of the multimodal inference in Table 7, in order to observe the influence of applied federated learning framework. It is evident that the approach incorporating federated learning clearly outperforms the case of training on single clients as well as cases of unimodal inference, comparing to the results in Tables 3–5.

**Table 7.** Detailed average performance of the proposed AVER model.

	AVER			AVER + FL		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Client 1	0.90	0.89	0.89	0.94	0.93	0.93
Client 2	0.88	0.87	0.87	0.93	0.93	0.93
Client 3	0.93	0.93	0.93	0.95	0.94	0.94

#### 4. Discussion

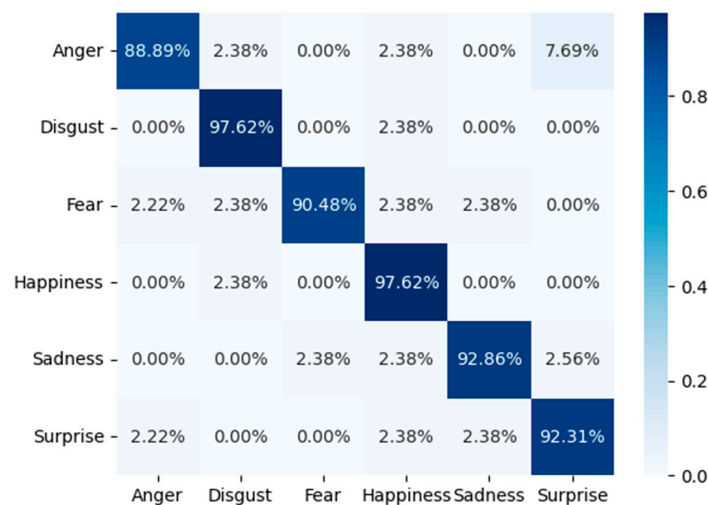
The aim of this section is to provide a performance comparison of the proposed audio-visual emotion recognition model accompanied by federated learning with other available state-of-the-art methods described in the literature and analyzed for the eNTERFACE’05 dataset. As already mentioned, there is a lack of multimodal audiovisual emotion recognition models that exploit federated learning. Consequently, similar experiments and dataset partitioning into such clients in the case of audiovisual emotion recognition are not described in the literature, which makes a comparison challenging. However, as the average performance per client in the case of a federated learning-based approach should not exceed the performance of the same non-federated model trained on the whole dataset, we compare average results with the performance of other non-federated learning-based methods available in the literature. The results are shown in Table 8. To highlight the importance of introducing federated learning techniques, we also provided the averaged results obtained by training the model for each client separately in a non-federated manner.

**Table 8.** Performance of the state-of-the-art audiovisual emotion recognition methods.

Model	Classification Accuracy (%)		
	Audio Modality	Video Modality	Multimodal
CNN + LSTM [15]	58.95	83.21	85.43
DL + attention [16]	/	/	88.11
MERML on AV [17]	55.9	77	91.50
C-GCN [18]	/	/	97.07
AV + MoBEL [19]	67.7	62	81.70
AVER	58.33	88.49	91.27
AVER + FL	67.86	/	93.29

By observing the results of the proposed model, we can conclude that the federated learning implementation increased multimodal accuracy by about 2% on average, whereas its influence is much more emphasized in the case of the audio modality. The performance of several state-of-the-art methods, tested in a non-federated manner in the case of the eNTERFACE’05 dataset is also provided.

The methods [15–19] are designed using different approaches compared to the proposed method. By analyzing the performance of methods [15–17,19], and by comparing them with the results of the proposed model, we can notice that the proposed method provides better multimodal accuracy. Also, the proposed method achieves better classification accuracy in the case of unimodal recognition compared to models [15,17,19]. Although the model given in [18] outperforms the proposed model in terms of the average classification accuracy, these observations should be considered carefully, as the performance of the proposed model is obtained by averaging the performance of each client after the federated learning procedure, unlike the procedure performed for other listed methods. In order to observe the achieved performance in detail, we provide a confusion matrix of the proposed AVER + FL model in Figure 4.



**Figure 4.** The confusion matrix for the eNTERFACE'05 dataset in the case of the proposed multimodal AVER + FL model.

By observing the results from Figure 4 and comparing them to the results from Figure 9 of the C-GCN model [18], we can observe that the proposed model provides better performance in the case of disgust and happiness, whereas the model from [18] outperforms the proposed model for other emotions. Additionally, we should highlight that the method proposed in this paper is based on the usage of lightweight CNN models following the objective of creating models capable of operating on edge devices. Therefore, an additional discussion about the tradeoff between classification accuracy and computational complexity, as well as the time required for pre-processing in the case of the graph convolutional network from [18] would be required for making any precise conclusions. Also, as the method we propose exploits transfer learning in the case of the video modality, the application of the federated learning approach in the case of C-GCN network might be more complex. Such considerations could be a part of our future research.

## 5. Conclusions

We studied the implementation of federated learning to the multimodal audiovisual emotion recognition task by designing two separate audio and video classifiers using convolutional neural networks, whose outputs are fused using late fusion. The proposed classifiers are relatively small, requiring 10.3 MB in the case of the video modality and 23.3 MB in the case of the audio modality, making them suitable for potential implementation on edge devices. Federated averaging is applied only to the audio modality since the analyzed audio models do not generalize well on the small, federated clients, whereas transfer learning provided good results in the case of the video modality, which required only 0.37 million trainable parameters. We demonstrated that implementation of federated learning significantly improves classifier performance in the case of the audio modality,

both in the case of single spectrogram classification (4.66% higher classification accuracy on average) and decision-making after sequential voting on the utterance level (9.53% higher classification accuracy on average). This further led to an overall improvement of ~2% in the case of multimodal emotion recognition compared to the non-federated scenario, making the proposed AVER model competitive and preferable to other state-of-the-art models. However, such comparison with other models has its limitations due to the fact that there are differences among federated and non-federated scenarios and there is a lack of similar experiments for comparison purposes.

In the future, we intend to extend the research from a proof of concept to the real recording scenario, involving real-time monitoring and monitoring of dialogues. Such scenario would require additional analysis involving the selection of edge devices, device connectivity and latency measurement, in order to understand limitations of a service deployment and provide audiovisual model as well as federated learning framework upgrades accordingly.

**Author Contributions:** Conceptualization and methodology, D.B. and N.S.; software, N.S., N.M., S.S. and V.S.; validation, N.S. and N.M.; formal analysis, N.S., N.M. and S.S.; investigation, N.S. and S.S.; resources, S.S., T.N. and B.P.; data curation, S.S., T.N., B.P. and N.S.; writing—original draft preparation, N.S., V.S., S.S. and N.M.; writing—review and editing, D.B., N.S., S.S., N.M., B.P. and T.N.; visualization, S.S. and N.S.; supervision, D.B.; project administration, D.B.; funding acquisition, D.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was funded by the European Union’s Horizon 2020 research and innovation program MARVEL under grant agreement No 957337. This publication reflects the authors’ views only. The European Commission is not responsible for any use that may be made of the information it contains.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data analyzed in this paper are available in a publicly accessible repository (the eINTERFACE ’05 audiovisual emotion database): <http://www.enterface.net/enterface05>, results section (accessed on 13 February 2023).

**Conflicts of Interest:** The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial neural network
AVER	Audiovisual emotion recognition model
CNN	Convolutional neural network
C-GCN	Correlation-based graph convolutional network
DL	Deep learning
DT	Decision tree
EEG	Electroencephalography
FedAvg	Federated averaging algorithm
GMM	Gaussian mixture model
HMM	Hidden Markov model
k-NN	k-nearest neighbors
LSTM	Long short-term memory
MFCC	Mel frequency cepstral coefficients
ML	Machine learning
PNG	Portable network graphics
RNN	Recurrent neural network
STFT	Short-time Fourier transform
SVM	Support vector machine

## References

1. Malik, M.; Malik, M.K.; Mehmood, K.; Makhdoom, I. Automatic speech recognition: A survey. *Multimed. Tools Appl.* **2021**, *80*, 9411–9457. [\[CrossRef\]](#)
2. Campanella, S.; Belin, P. Integrating face and voice in person perception. *Trends Cogn. Sci.* **2007**, *11*, 535–543. [\[CrossRef\]](#)
3. Wu, C.; Lin, J.; Wei, W. Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies. *APSIPA Trans. Signal Inf. Process.* **2014**, *3*, E12. [\[CrossRef\]](#)
4. Avots, E.; Sapiński, T.; Bachmann, M.; Kamińska, D. Audiovisual emotion recognition in wild. *Mach. Vis. Appl.* **2019**, *30*, 975–985. [\[CrossRef\]](#)
5. Schoneveld, L.; Othmani, A.; Abdelkawy, H. Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognit. Lett.* **2021**, *146*, 1–7. [\[CrossRef\]](#)
6. Akçay, M.B.; Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* **2020**, *116*, 56–76. [\[CrossRef\]](#)
7. Khalil, R.A.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access* **2019**, *7*, 117327–117345. [\[CrossRef\]](#)
8. Andayani, F.; Theng, L.B.; Tsun, M.T.; Chua, C. Hybrid LSTM-transformer model for emotion recognition from speech audio files. *IEEE Access* **2022**, *10*, 36018–36027. [\[CrossRef\]](#)
9. Dominic Enriquez, M.; Rudolf Lucas, C.; Aquino, A. Scalogram vs Spectrogram as Speech Representation Inputs for Speech Emotion Recognition Using CNN. In Proceedings of the 34th Irish Signals and Systems Conference (ISSC), Dublin, Ireland, 13–14 June 2023; pp. 1–6. [\[CrossRef\]](#)
10. Canal, F.Z.; Müller, T.R.; Matias, J.C.; Scotton, G.G.; de Sa Junior, A.R.; Pozzebon, E.; Sobieranski, A.C. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Inf. Sci.* **2022**, *582*, 593–617. [\[CrossRef\]](#)
11. Fraiwan, M.; Alafeef, M.; Almomani, F. Gauging human visual interest using multiscale entropy analysis of EEG signals. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *12*, 2435–2447. [\[CrossRef\]](#)
12. Fournier, Q.; Caron, G.M.; Aloise, D. A practical survey on faster and lighter transformers. *ACM Comput. Surv.* **2023**, *55*, 304. [\[CrossRef\]](#)
13. Jelčicová, Z.; Verhelst, M. Delta keyword transformer: Bringing transformers to the edge through dynamically pruned multi-head self-attention. *arXiv* **2022**, arXiv:2204.03479.
14. Bajovic, D.; Bakhtiarnia, A.; Bravos, G.; Brutti, A.; Burkhardt, F.; Cauchi, D.; Chazapis, A.; Cianco, C.; Dall’Asen, N.; Delic, V.; et al. MARVEL: Multimodal Extreme Scale Data Analytics for Smart Cities Environments. In Proceedings of the International Balkan Conference on Communications and Networking (BalkanCom), Novi Sad, Serbia, 20–22 September 2021; pp. 143–147. [\[CrossRef\]](#)
15. Ma, F.; Zhang, W.; Li, Y.; Huang, S.-L.; Zhang, L. An End-to-End Learning Approach for Multimodal Emotion Recognition: Extracting Common and Private Information. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 1144–1149. [\[CrossRef\]](#)
16. Tang, G.; Xie, Y.; Li, K.; Liang, R.; Zhao, L. Multimodal emotion recognition from facial expression and speech based on feature fusion. *Multimed. Tools Appl.* **2023**, *82*, 16359–16373. [\[CrossRef\]](#)
17. Ghaleb, E.; Popa, M.; Asteriadis, S. Metric Learning-Based Multimodal Audio-Visual Emotion Recognition. *IEEE MultiMedia* **2020**, *27*, 37–48. [\[CrossRef\]](#)
18. Nie, W.; Ren, M.; Nie, J.; Zhao, S. C-GCN: Correlation Based Graph Convolutional Network for Audio-Video Emotion Recognition. *IEEE Trans. Multimed.* **2021**, *23*, 3793–3804. [\[CrossRef\]](#)
19. Farhoudi, Z.; Setayeshi, S. Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition. *Speech Commun.* **2021**, *127*, 92–103. [\[CrossRef\]](#)
20. Chhikara, P.; Singh, P.; Tekchandani, R.; Kumar, M.; Guizani, M. Federated Learning Meets Human Emotions: A Decentralized Framework for Human–Computer Interaction for IoT Applications. *IEEE Internet Things J.* **2021**, *8*, 6949–6962. [\[CrossRef\]](#)
21. Nandi, A.; Xhafa, F. A federated learning method for real-time emotion state classification from multi-modal streaming. *Methods* **2022**, *204*, 340–347. [\[CrossRef\]](#)
22. Salman, A.; Busso, C. Privacy Preserving Personalization for Video Facial Expression Recognition Using Federated Learning. In Proceedings of the ICMI ’22: 2022 International Conference on Multimodal Interaction, Bangalor, India, 7–11 November 2022; pp. 495–503. [\[CrossRef\]](#)
23. Chang, Y.; Laridi, S.; Ren, Z.; Palmer, G.; Schuller, B.W.; Fisichella, M. Robust Federated Learning Against Adversarial Attacks for Speech Emotion Recognition. *arXiv* **2022**, arXiv:2203.04696.
24. Zhang, T.; Feng, T.; Alam, S.; Lee, S.; Zhang, M.; Narayanan, S.S.; Avestimehr, S. FedAudio: A Federated Learning Benchmark for Audio Tasks. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5. [\[CrossRef\]](#)
25. Rybka, J.; Janicki, A. Comparison of speaker dependent and speaker independent emotion recognition. *Int. J. Appl. Math. Comput. Sci.* **2013**, *23*, 797–808. [\[CrossRef\]](#)
26. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520. [\[CrossRef\]](#)



27. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [\[CrossRef\]](#)
28. Wei, J.; Yang, X.; Dong, Y. User-generated video emotion recognition based on key frames. *Multimed. Tools Appl.* **2021**, *80*, 14343–14361. [\[CrossRef\]](#)
29. Hossain, M.S.; Muhhamad, G. Emotion recognition using deep learning approach from audio–visual emotional big data. *Inf. Fusion* **2019**, *49*, 69–78. [\[CrossRef\]](#)
30. Simic, N.; Suzic, S.; Nosek, T.; Vujovic, M.; Peric, Z.; Savic, M.; Delic, V. Speaker Recognition Using Constrained Convolutional Neural Networks in Emotional Speech. *Entropy* **2022**, *24*, 414. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Madanian, S.; Chen, T.; Adeleye, O.; Templeton, J.M.; Poellabauer, C.; Parry, D.; Schneider, S.L. Speech emotion recognition using machine learning—A systematic review. *Intell. Syst. Appl.* **2023**, *20*, 200266. [\[CrossRef\]](#)
32. Nagrani, A.; Chung, J.S.; Xie, W.; Zisserman, A. Voxceleb: Large-scale speaker verification in the wild. *Comput. Speech Lang.* **2020**, *60*, 101027. [\[CrossRef\]](#)
33. Simic, N.; Suzic, S.; Nosek, T.; Vujovic, M.; Secujski, M. Impact of different voting strategies in CNN based speech emotion recognition. In Proceedings of the 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 29 August–2 September 2022; pp. 1174–1177. [\[CrossRef\]](#)
34. Hershey, S.; Chaudhur, S.; Ellis, D.P.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN architectures for large-scale audio classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 131–135. [\[CrossRef\]](#)
35. Shehu, H.A.; Browne, W.; Eisenbarth, H. Emotion Categorization from Video-Frame Images Using a Novel Sequential Voting Technique. In *Advances in Visual Computing. ISVC 2020. Lecture Notes in Computer Science*; Bebis, G., Yin, Z., Kim, E., Bender, J., Subr, K., Kwon, B.C., Zhao, J., Kalkofen, D., Baci, G., Eds.; Springer: Cham, Switzerland, 2020; Volume 12510. [\[CrossRef\]](#)
36. Brecko, A.; Kajati, E.; Koziorek, J.; Zolotova, I. Federated Learning for Edge Computing: A Survey. *Appl. Sci.* **2022**, *12*, 9124. [\[CrossRef\]](#)
37. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR 54, Ft. Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
38. Martin, O.; Kotsia, I.; Macq, B.; Pitas, I. The eNTERFACE’05 Audio-Visual Emotion Database. In Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW’06), Atlanta, GA, USA, 3–7 April 2006; pp. 1–8. [\[CrossRef\]](#)
39. Conrad, J.J.; Watts, J.; Henry, T.R.; List, J.-M.; Forkel, R.; Mucha, P.J.; Greenhill, S.J.; Gray, R.D.; Lindquist, K.A. Emotion semantics show both cultural variation and universal structure. *Science* **2019**, *366*, 1517–1522. [\[CrossRef\]](#)

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.