# Multimodal Emotion Classification Supported in the Aggregation of Pre-Trained classification models*

Pedro J. S. Cardoso[1,2][0000−0003−4803−7964], João M. F. Rodrigues[1,2][0000−0002−3562−6025], and Rui Novais[2][0000−0002−6720−9234]

[1] LARSyS – Laboratory for Robotics and Engineering Systems, ISR-Lisbon, 1049-001 Lisboa, Portugal
[2] Instituto Superior de Engenharia, Universidade do Algarve, 8005-129 Faro, Portugal
{pcardoso,jrodrig}@ualg.pt

**Abstract.** Human-centric artificial intelligence struggles to build automated procedures that recognize emotions which can be integrated in artificial systems, such as user interfaces or social robots. In this context, this paper researches on building an Emotion Multi-modal Aggregator (EMmA) that will rely on a collection of open-source single source emotion classification methods aggregated to produce an emotion prediction. Although extendable, tested solution takes a video clip and divides into its frames and audio. Then a collection of primary classifiers are applied to each source and their results are combined in a final classifier utilizing machine learning aggregator techniques. The aggregator techniques that have been put to the test were Random Forest and k-Nearest Neighbors which, with an accuracy of 80%, have demonstrated superior performance over primary classifiers on the selected dataset.

**Keywords:** Affective Computing · Multimodal Ensembles · Facial Emotions · Speech Emotions.

## 1 Introduction

While part of the humans' communication is verbal, the truth is that a big part of our communication is nonverbal. Facial expressions, the tone of the voice (vocalization), body movements and gestures, posture, all contribute to how we communicate and understand each other. Often, humans are not even aware of that nonverbal part of what they transmit or receive, because this is inherent to them, from the day they are born. Communication is therefore achieved by using multiple sources, received by multiple "sensors", which implies that using a single source of information, such as face expressions, will provide limited information to an automated emotion detection process.

* We thank the Portuguese Foundation for Science and Technology (FCT) under Project UIDB/50009/2020—LARSyS.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_35

Human-centric artificial intelligence (HCAI) struggles to build automated procedures that recognize emotions which can, and in many cases should, be integrated in artificial systems, such as user interfaces or social robots [5,17,53]. For example, research on the latter subject is fundamental, as the worldwide elderly population is set to be more than double by 2050 and robots are expected to assume new roles in health and social care, to meet that higher demand [1,21]. A robot can only really interact with a person, if it achieves some degree of emotional recognition in interaction, i.e., if it understands the person's emotions and sentiments in a way to, on the fly, adjust its behavior in function of it. Emotion and sentiment analysis are therefore fundamental in the development of socially assistive robot (SAR) technologies for people care. Recent studies analyze emotional intelligence in SAR for elders [2] or which aspects may influence human-robot interaction in assistive scenarios [52]. Poria et al. [45] address the multi-modal sentiment and emotion prediction, living open to development several problems such as aspect-level sentiment analysis, sarcasm analysis, multimodal sentiment analysis, sentiment aware dialogue generation, and others. Also, Birjali et al. [10] present a study of sentiment analysis approaches, challenges, and trends, to give researchers a global survey on sentiment analysis and its related fields.

So, to implement SAR technologies, state-of-the-art results in emotions and sentiments classification are needed, being machine learning (ML) algorithms the more promising methods at the moment. Several ways are known to improve algorithms' results, being the more usual way to train them repeatedly, with available data, with different settings, until the best possible result is achieved (fine-tuning the algorithm). Training might be extremely time-consuming, as well as it implies spending a lot of energy during the training phase, also increasing the model's "carbon footprint". A solution to mitigate this is applying ensemble techniques, i.e., using the results from various algorithms previously thought and available in the community [35,34]. This use of hybridization / ensemble techniques allows empowering computation, functionality, robustness, and accuracy aspects of modelling [6], as well as it allows to reduce the referred "carbon footprint" of the models.

In this context, this paper is part of a series of studies to build a framework for emotion classification based on multiple sources (e.g., facial, speech, text, and body expression), the Emotion Multi-modal Aggregator (EMmA). The EMmA is to be supported on an ensemble of open-source code, retrieved from off-the-shelf available methods. The previous studies already presented the idea associated with single sources, namely, faces [35] and speech [34]. Here, the authors propose to integrate both sources in a single prediction, consistent with the emotions presented by the system's user. In more detail, given as input a video clip, the process starts by splitting it in its set of frames / images and its soundtrack. Then, a set of primary classifiers is applied to each source (images and speech), returning a set of probability associated to each emotion. A ML aggregator method is then fitted with those probabilities to build a final classifier. Then, to make inference over new samples, those samples will pass through the

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_35

primary classifiers and the predicted probabilities are injected in the aggregator, to estimate a final prediction. Details are presented in Section 3.

The tested aggregator method includes two well know ML algorithm, namely: Random Forest (RF) and k-Nearest Neighbors (kNN). For faces and for speech, used independently, the proposed single source aggregators formerly proved their effectiveness over the primary classifiers, improving the individual classifiers' accuracy over the Facial Expression Recognition 2013 (FER-2013) Dataset [22], the Real-world Affective Faces Database (RAF-DB) [30], the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [32], the Toronto emotional speech set (TESS) [41], the Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) [15], and the Surrey Audio-Visual Expressed Emotion (SAVEE) database [23] for speech. Further, a baseline aggregator was defined using a voting methodology, i.e., each primary classifier "votes" in one emotion and the one with more votes is selected as the estimation.

The best configuration was achieved with the kNN, with an accuracy of 80.6%, 9.7% better than the best result for the individual classifiers, and 4.9% better than the voting method. Besides accuracy, the recall, the precision and the F1 scores also corroborate the attained results.

Some main contributions of the paper are the proof that the previously proposed methods in [34,35] can be extended to a multi-modal scenario, it is possible to build a high accuracy emotion detector combining off-the-shelf methods, which may allow reducing costs (e.g., computational, financial, $CO_2$ emissions etc.). Further, it is possible to see that the primary classifiers can be trained in different datasets than the ones on which the aggregator is to be applied, allowing a generalization of the proposed method.

The paper is structured as follows. The next section presents a brief state of the art in emotion detection from static images, speech, and video clips. The following section presents the EMmA and voting methods, along with some consideration about the used dataset. The fourth section presents the experimental results. The final section present the conclusion and future work.

## 2 Related Work

Recognizing expressions in order to predict interpersonal relations requires input from various sources, including sound, body language, and facial expressions, as well as factors such as age and cultural environment. For instance, Zhang et al. [56] proposed an effective multitask network that is able to learn from various auxiliary attributes, such as gender, age, and head pose, in addition to just facial expression data. Noroozi et al. [33] explored the topic of emotional body gesture classification through a comprehensive survey, concluding that, despite a plethora of research on facial expressions and speech, the recognition of the impact of body gestures remains a less explored area. The work intended to increase interest in this field by providing a new survey and highlighting the importance of emotional body gestures as a component of "body language", discussing things such as gender differences and cultural dependence, within the

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_35

context of emotional body gesture recognition. A complete framework for automatic emotional body gesture recognition was also presented. Other solutions were also proposed, such as combining body posture with facial expressions for classifying affect in child-robot interaction [20]. More recent examples can be found in the literature, such as studies on mood estimation based on facial expressions and postures [14], or in other works such as the ones from Ahmed et al. [3] or Liang et al. [31].

Ekman and Friesen [19], concentrating on face expression, showed that facial displays of emotion are universal, demonstrating that the human ability to express an emotion is an evolutionary, biological fact that is independent of any particular culture. Nevertheless, even given the same input, several approaches for facial emotion classification produce varying outcomes. So, facial expression recognition utilizing a group of classifiers is not a novel concept. For instance, a pool of base classifiers developed utilizing two feature sets – Gabor filters and Local Binary Patterns – was described by Zavaschi et al. [55]. The accuracy and size of the ensemble were employed as objective functions in a multi-objective genetic algorithm that was used to find the best ensemble. Ali et al. [4] presented an ensemble method for evaluating multicultural facial expressions, proposing a set of computational strategies to manage those variations. They make use of facial photos taken from participants in the multicultural dataset, who belong to four distinct ethnic groups, namely, "Caucasians", Japan, Taiwan, and Morocco. Wang et al. [54] presented the Oriented Attention Ensemble for Accurate Facial Expression Recognition. An oriented attention pseudo-Siamese network that utilizes both global and local face information was employed by the authors. The network consists of two branches: an attention branch with a UNet-like architecture to gather local highlight information and a maintenance branch with various convolutional blocks to exploit high-level semantic features. To output the results of the classification, the two branches are combined. Benamara et al. [8] present a facial emotion recognition system that deals with automated facial detection and facial expression classification separately. The latter is carried out by a limited ensemble of only four deep convolutional neural networks, and a label smoothing technique is employed to deal with the training data that has been incorrectly labeled. The Local (Multi) Head Channel (Self-Attention) method, or LHC for short, is founded on two primary concepts [38]. First, convolution will not be replaced by attention modules like recurrent networks were in NLP (natural language processing); and second, a local approach has the potential to overcome convolutions' limitations more effectively than global attention. This is because local attention is more focused on the local region of interest than global attention, which is where the self-attention paradigm has been most extensively studied in computer vision. With LHC, the authors were able to surpass the previous state-of-the-art for the FER-2013 dataset, with a substantially reduced level of complexity and impact on the "host" architecture in terms of computational cost. An open-source Python toolbox named Py-Feat [25] supports the detection, pre-processing, analysis, and visualization of facial expression data. Py-Feat allows end users to swiftly process, analyze, and visualize face expres-

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_35

sion data while also enabling experts to share and benchmark computer vision models. For further studies in face emotion classification, please refer to, e.g., the works of Banerjee et al. [7] and Revina & Emmanuel [47] who assess multiple deep learning algorithms for effective facial expression classification and human face recognition techniques.

Popova et al. [44] described a method in which the classification of a sound fragment is reduced to an image recognition issue. The waveform and spectrogram are used by the authors to represent the sound. When they combine a Mel-spectrogram with a convolutional neural network (VGG-16), they test their method with RAVDESS and get an accuracy of 71%. The Mel-spectrogram with deltas and delta-deltas is utilized as input by Chen et al. [16] in a 3D attention-based convolutional recurrent neural network to learn discriminative features for speech emotion recognition. Experiments on the Interactive emotional dyadic motion capture database (IEMOCAP) [13] and Berlin Database of Emotional Speech (Emo-DB) corpus [11] provided cutting-edge results. By demonstrating 92.89% validation accuracy on the ESC-50 dataset and 87.42% validation accuracy on the UrbanSound8K dataset, Palanisamy et al. [37] assert that ImageNet pre-trained standard deep convolutional neural network (CNN) models can be employed as powerful baseline networks for audio categorization. De Pinto et al. [43] presented a CNN-based classification model of the emotions produced by speeches (using the RAVDESS dataset). The neutral and calm emotions, as well as those described by Ekman in 1992 [18], have also been taught to the model. They received a weighted average F1 score of 0.91. Using the RAVDESS, Emo-DB, and CaFE databases, El Seknedy and Fawzi [48]presented their findings on speech emotion classification. The key speech features are prosodic 7 features, spectral features, and energy. They employ four machine learning classifiers (Multi-Layer Perceptron, Support Vector Machine – SVM, Random Forest, and Logistic Regression). The models' accuracy was 70.56% on RADVESS, 85.97% on Emo-DB, and 70.61% on CaFE. A deep continuous recurrent neural network (C-RNN) method was presented by Kumaran et al. [29] to classify the efficiency of learning emotion changes in the classification stage. To begin with, they extract high-level spectral features using a combination of Mel-Gammatone filter in convolutional layers. The long-term temporal context is then learned from the high-level features using recurrent layers. In RADVESS, the authors had an accuracy of 80%.

Siddiqui and Javaid [50] created a framework for classifying facial and vocal emotions. Three CNNs and two detection layers make up the proposed structure. Two CNNs are trained separately utilizing visible and infrared images in the first layer, and the features they produce are then given to an SVM for classification. Another CNN was used to learn the emotions in speech using information extracted from audio spectrograms. Ankur Bhatia suggested a system that can extract sentiment and emotion from text, facial and sound [9,46]. The technique was developed over MELD (Multimodal Emotion-Lines Dataset) dataset. MELD is a dataset for spoken language emotion recognition. It is designed to be used for training and assessing models for multimodal emotion recognition and in-

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_35

cludes audio, text transcriptions, and annotations for the emotions expressed in conversations. MELD, which was taken from the Friends TV series, has more than 13,000 dialogues, more than 100,000 utterances, 7 fundamental emotions, and 2 feelings (positive and negative). On this subject, several other works can be found such as the ones from Pandeya & Lee [28], Heredia et al. [24] or Ortega el al. [36].

Despite the fact that the methodologies described have partially or globally the same overall objective as this work, they do it in different ways. The aforementioned authors concentrate on creating a single model, or on teaching a brand-new model, involving a sizable amount of data to learn from. Instead, the presented approach aims to build on the successes of earlier approaches, concentrating on employing primary classifiers that have already been built and making use of them as part of a final classification technique. I.e., the approach here presented is intended to learn from the outcomes of previously developed models, simplifying the learning phase and reducing the time needed to teach the classification model as well as the computing power that is needed for that. The proposed framework is further examined in the next sections.

## 3 Multi-Source Aggregator and Sentiment Classifier

The proposed Emotion Multi-modal Aggregator (EMmA) is a multi-modal extension of the aggregators previously presented by Novais et al. [34,35] where face and speech were treated separately. Figure 1 illustrates the model's architecture and flow, where the primary classifiers are used in the following manner. First, Fig. 1(a), the primary classifiers models were pre-trained with some dataset proper for their objective (in the present case, face emotion dataset or speech emotion dataset). Then, Fig. 1(b), the obtained models use inference over a new dataset which includes at least faces and speech, to define the training and testing dataset that the aggregator will use to train its own model. I.e., in the present case, for each sample, each primary classifier model predicts the probability of each emotion. Since we are considering 7 emotions, following Ekman and Friesen [19] plus neutral (namely: neutral, calm, happy, sad, angry, fearful, surprised, and disgusted), and 6 primary classifiers, the output will be 42 "emotions-probabilities" values associated to a target. This 42 "emotions-probabilities" associated to a ground-truth values build the aggregator's dataset. Finally, with the primary classifiers' and aggregator's models trained, inference can be done by "passing" the new sample through the primary classifiers, which will return the emotion-probabilities, which are then fed to the aggregator, which will infer an emotion, Fig. 1 (c). Overall, one of the most cost consuming step is the first one, (a), which in this architecture can be simplified from the moment a trained model is available.

If some source (e.g., faces or speech) is not available, the aggregator's model can be retrained without requiring any change to the primary classifiers models. In this context, a solution to activate/deactivate primary classifiers models is being thought. Further, if a new primary classifier, from the same or new

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_35

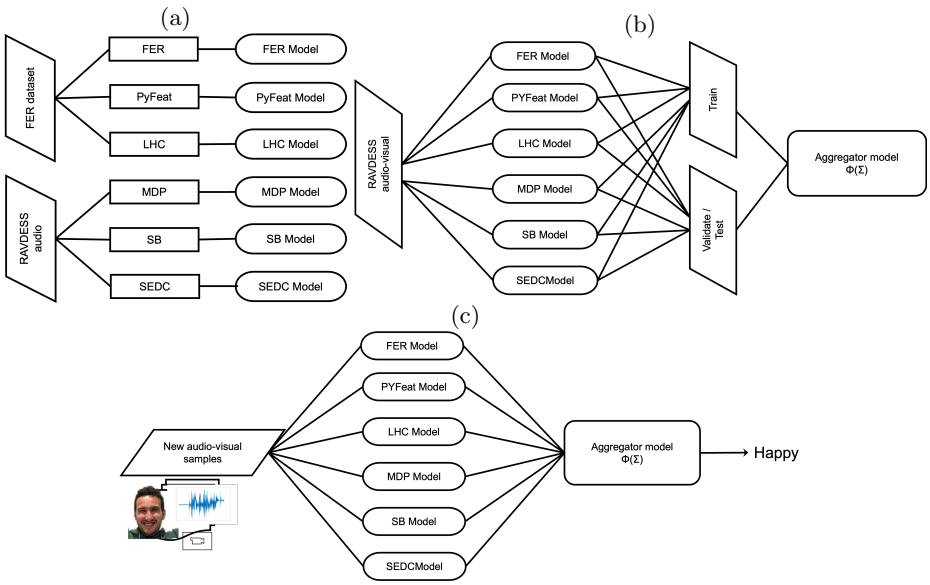


**Fig. 1.** Architecture and flow of the Emotion Multi-modal Aggregator (EMmA) classifier framework: (a) training of the primary classifiers' model, (b) building the dataset to be used by the aggregator for training and training of the aggregator's model, and (c) samples inference using the primary classifiers and aggregator's models.

sources (e.g., text or body posture), becomes available, this model can be added requiring, as expected, the retraining of the aggregator's model. For this latter condition, continual learning solution are being thought. Finally, as already mentioned, (i) the primary classifiers can be off-the-shelf models (see Sec. 3.1) and (ii) the aggregator method was implemented using common ML methods, namely: RF and kNN classifiers (see Sec. 3.4). Besides, complementing the threshold performance given by the individual methods, a baseline for the aggregator method was defined using a voting methodology (see Sec. 3.3).

### 3.1 Primary classifiers

Three primary classifiers for faces and three for speech were taken into consideration. So, the following methods were employed in relation to face detection classifiers: (i) Local (Multi) Head Channel (Self-Attention) (LHC)[38], whose source code is accessible at [39]; (ii) Py-Feat [26], whose source code is available at [27]; and (iii) FERjs, a free implementation created by Justin Shenk, whose source code is available at [49]. It is essential to reiterate the fact that, as was already mentioned, there are other potential approaches [7,47]. On the speech side, the primary emotion classifiers were: (i) MDP [43], with its code available at [42]; (ii) SB, which is a free implementation done by Shivam Burnwal and has its code available at [12]; and SEDC (Speech Emotion Detection Classifier), which is an implementation done by the authors [34]. The six classifiers

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_35

were chosen because they offer cutting-edge results, are recent implementations, represent various architecture, and have publicly accessible code.

The primary classifiers for face emotion classification were trained using FER dataset (see [35]) and the primary classifiers for speech were trained using (audio) RAVDESS dataset (see [34]). It is also important to stress that data used from RAVDESS to train the speech was not used to validate and test the multi-modal aggregator. Nevertheless, the EMmA aggregator was trained using audiovisual-RAVDESS data (image and speech).

Furthermore, since the methods for facial emotion recognition were trained for static images, the facial emotion classification previously developed had to be prepared to deal with videos, as the audiovisual-RAVDESS is composed by movie clips (see Sec. 3.2). The process included the following steps. For each clip, the (i) first 30 frames (1 second) and the (ii) last 30 frames (1 second) were discarded. Then, for the (iii) remaining frames it was applied the primary classifiers to each one, followed by a (iv) non-maximum suppression technique, i.e., over the results of the clip a sliding neighborhood window with similar emotions are considered as candidate classes, which leads to several proposals. It was considered the proposal/emotion with the highest count.

Table 1 shows the baseline results for the primary classifiers methods, as explained above[3]. The results from the face classifiers seem reasonable due to the approach that applies static faces emotion detection methods to video clips, as explained in the previous paragraph, and the use of different datasets. On the other side, the speech classifiers seem obviously overfitted, since the difference between the metric values attained with the train data set are significantly better than the ones attained with the validation and test data sets. This later fact, was somehow considered as acceptable since we are using off-the-shelf methods. So, it was decided to keep them and proceed to the following steps.

### 3.2 Dataset for the aggregators

The three primary classifiers for face emotion classification were trained using FER dataset (see [35]) and the three primary classifiers for speech were trained using audio-RAVDESS (audio files, see [34]). Then, a randomly selected part of the audiovisual speech files of the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [32] (720p H.264, AAC 48kHz, MP4) were used to build a dataset for the aggregator. I.e., the primary classifiers were applied to the audiovisual-RAVDESS' files to build a set of 1437 for samples for training, 308 samples for validation and 309 samples for testing (see Tab. 2). The database contains 24 professional actors (12 female and 12 male), vocalizing two lexically matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprised, and disgusted emotions. Each expression is produced at two levels of emotional intensity (normal and strong), with an additional neutral expression. Data used from RAVDESS to train the speech were not used to validate or test the final multi-modal aggregators.

[3] See, e.g., [51] for the definitions of Accuracy, Precision, Recall, and F1-score.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_35

**Table 1.** Primary classifiers' individual performance.

| | | FER | PyFeat | LHC | MDP | SB | SEDC |
|---|---|---|---|---|---|---|---|
| Train | Accuracy | 0.555 | 0.460 | 0.531 | **0.996** | 0.990 | 0.944 |
| | F1-score | 0.552 | 0.461 | 0.516 | **0.996** | 0.990 | 0.944 |
| | Recall | 0.555 | 0.460 | 0.531 | **0.996** | 0.990 | 0.944 |
| | Precision | 0.631 | 0.553 | 0.613 | **0.996** | 0.990 | 0.945 |
| Validation | Accuracy | 0.506 | 0.425 | 0.500 | 0.682 | **0.705** | 0.636 |
| | F1-score | 0.499 | 0.442 | 0.490 | 0.688 | **0.707** | 0.637 |
| | Recall | 0.506 | 0.425 | 0.500 | 0.682 | **0.705** | 0.636 |
| | Precision | 0.569 | 0.575 | 0.588 | 0.706 | **0.714** | 0.642 |
| Test | Accuracy | 0.531 | 0.450 | 0.528 | **0.709** | 0.673 | 0.612 |
| | F1-score | 0.530 | 0.449 | 0.520 | **0.713** | 0.666 | 0.597 |
| | Recall | 0.531 | 0.450 | 0.528 | **0.709** | 0.673 | 0.612 |
| | Precision | 0.611 | 0.517 | 0.614 | **0.727** | 0.667 | 0.607 |

**Table 2.** Emotions classes' distribution on the used audio-visual RAVDESS subset.

| | angry | disgust | fear | happy | neutral | sad | surprise | **total** |
|---|---|---|---|---|---|---|---|---|
| **Train** | 263 | 134 | 263 | 263 | 127 | 252 | 135 | **1437** |
| **Validation** | 57 | 29 | 55 | 57 | 28 | 54 | 28 | **308** |
| **Test** | 56 | 29 | 57 | 56 | 28 | 54 | 29 | **309** |

### 3.3 Voting aggregator's baseline

As a very simple baseline for the aggregator method, it was decided to implement a voting method, as detailed in this section. Each of the primary classifiers return a "vote", predicting an emotion, supported on the emotion with the highest probability it produced. Then, the voting aggregator predicts an emotion as the one with most votes. In the case were two or more emotions are tie (with same amount of votes), and one of the tied emotions is the real one, for metrics purposes, the voting was accounted as correct. Table 3 shows the metrics attained with the voting aggregator, showing an increase in the accuracy over the test dataset (comparing with the best primary classifiers' results), improving the accuracy from 70.9% to 75.7%. For the test dataset, the remaining metrics (Recall, Precision and F1-score) were also improved.

### 3.4 ML aggregator methods

Two well-known machine learning methods were used to implement ML aggregator, namely: RF and kNN classifiers. The two ML methods were fitted using grid search cross validation, considering 6 training/testing cases: (i) the training data set was the dataset obtained by running the primary classifiers over the audiovisual-RADVESS samples defined for training ($D$); (ii) the train dataset was the previous one but standard scaled ($D_{scaled}$), i.e., removing the mean and

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_35

**Table 3.** Voting aggregator metrics over the audiovisual-RAVDESS dataset.

| | Accuracy | F1-score | Recall | Precision |
|---|---|---|---|---|
| train | 0.981 | 0.981 | 0.981 | 0.982 |
| val | 0.740 | 0.748 | 0.740 | 0.786 |
| **test** | **0.757** | **0.756** | **0.757** | **0.789** |

scaling to unit variance, $z = (x - \mu_f)/\sigma_f$, where $\mu_f$ in the mean value and $\sigma_f$ the standard deviation of the values observed for each feature $f$; (iii) the training dataset was set as the scaled version of the union of the train and validation datasets ($DV_{scaled}$); and (iv)–(vi) the training dataset was the result of applying a polynomial feature transformation of degree 2 to the previous datasets (respectively, $D_{poly}$, $D_{scaled,poly}$, $DV_{scaled,poly}$). In this context, since cross validation was being used, it was possible to skip validation, going directly to testing. Further, this allowed to observe if adding new samples, which were not direct part of training of the primary classifiers, and did not seem to have major liking to the primary classifiers fitting given the metrics values, could improve the aggregator's performance.

To summarize, aggregators trained with datasets (i)–(iii) receive $7 \times 6 = 42$ values, where 7 is the number of expressions (neutral, calm, happy, sad, angry, fearful, surprised, and disgusted) per primary classifier and 6 is the number of primary classifiers (3 for faces and 3 for speech emotion recognition). Similar, doing all combinations of features, 946 features are fed to the aggregator for cases (iv)–(vi).

## 4 Tests and Results

The experimental process was conducted using version 1.0.2 of the Scikit–Learn framework [40]. In this context, a grid-search cross validation was developed considering 5 folders and the following parameters. To fit the RF it were considered number of estimators in the set $\{50, 100, 200, 400\}$, criterion in {"gini", "entropy"}, maximum depth in $\{2, 5, 10, \infty\}$, minimum number of samples required to split an internal node in $\{2, 5, 10\}$, the minimum number of samples required to be at a leaf node in $\{1, 2, 5, 10\}$, the minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node in $\{0, .25, .5\}$, and the maximum number of features to consider when looking for the best split in {"sqrt", "log2"}. For kNN case, it was considered number of neighbors in $\{3, 5, 7, 9\}$, the weights in {"uniform", "distance"}, the algorithm in {"ball_tree", "kd_tree", "brute"}, and power parameter for the Minkowski metric ($p$) in $\{1, 2\}$. In both cases, the remaining parameters were the default of the Scikit–Learn library.

Table 4 summarizes the results achieved over the test set. The best results were achieved using kNN (configured with a ball_tree, 3 neighbors, $p = 1$, and uniform weights, returned from the grid search cross validation process) using

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_35

**Table 4.** Accuracy, F1-score, Recall and Precision for the aggregators methods.

| ML method | Training dataset | Accuracy | F1-score | Recall | Precision |
|---|---|---|---|---|---|
| RF | $D$ | 0.735 | 0.732 | 0.740 | 0.727 |
| kNN | $D$ | 0.767 | 0.762 | 0.766 | 0.763 |
| RF | $D_{poly}$ | 0.731 | 0.722 | 0.725 | 0.722 |
| kNN | $D_{poly}$ | 0.764 | 0.761 | 0.766 | 0.761 |
| RF | $D_{scaled}$ | 0.735 | 0.732 | 0.740 | 0.727 |
| kNN | $D_{scaled}$ | 0.796 | 0.789 | 0.791 | 0.793 |
| RF | $D_{scaled,poly}$ | 0.738 | 0.727 | 0.733 | 0.724 |
| kNN | $D_{scale,poly}$ | 0.803 | 0.794 | **0.794** | **0.805** |
| RF | $DV_{scaled}$ | 0.770 | 0.767 | 0.768 | 0.771 |
| kNN | $DV_{scaled}$ | 0.796 | 0.786 | 0.784 | 0.794 |
| RF | $DV_{scaled,poly}$ | 0.793 | 0.786 | 0.788 | 0.792 |
| kNN | $DV_{scaled,poly}$ | **0.806** | **0.796** | **0.794** | **0.805** |

scaled and polynomial features over the training and validation data, with an accuracy of 80.6%, more 9,7% than the best result for the individual classifiers (MDP achieved an accuracy of 70.9%), and 4.9% higher accuracy than the voting method (which achieved an accuracy of 75.7%). Moreover, without using validation as part of the training of the aggregator, the best result was also achieved by the kNN method, with an accuracy of 80.3%, just 0.3% less than the best case. This latter case, attained the same recall and precision than the best case, and the F1-score was just 0.002 points different (0.796 to 0.794). Relatively to the RF, it was found a big difference when considering the training and validation data (scaled and with polynomial features) to train the aggregator, passing from 73.8% accuracy ($D_{scaled,poly}$) to 79.3% accuracy.

## 5 Conclusion

This paper presented a framework based on an Emotion Multi-modal Aggregator (EMma), which aggregates the results extracted from the primary emotion classifications from different sources, namely facial and speech. The framework was tested using the audiovisual-RAVDESS dataset, somehow validating the initial concept: it is possible to build a state-of-the-art emotion detection system supported on methods (primary classifiers) available as open-source, trained with distinct datasets, with a minimum training of an aggregator. Another advantage of this solution is the possible speed-up in the development of an integrated solution for human emotion classification.

In conclusion, there are still a considerable number of questions that remain unanswered and represent opportunities for future research in this area. One area that could benefit from further exploration is the relationship between the number of primary classifiers used and the overall computational complexity. By delving deeper into this issue, researchers can optimize the classifier's efficiency

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_35

while maintaining high levels of accuracy. Additionally, investigating the impact of incorporating primary classifiers that are significantly different in terms of accuracy compared to those already used should be addressed. Future work will also focus on the inclusion of other sources to the framework (e.g., text and body posture) and in the definition of a proper dataset that will allow testing all the strands of the model. This includes the research on the influence of other dimensions, such as gender, ethnicity, and age. By considering a broader range of factors, the model can be refined and improved to better address real-world applications and challenges. Further, if some source (e.g., faces or speech, at the moment) is not available, the aggregator model should have a solution to activate/deactivate the corresponding primary classifiers models, without retraining the former one. Also in this context, including new/removing primary classifiers, from the same or new sources (e.g., text or body posture), without retraining of the aggregator model is an objective, i.e., a procedure commonly designated as continual learning. Overall, the paper provides a foundation for further exploration and development of automated classification techniques, and future research in this area holds promises for advancing the field.

## References

1. Abdi, J., Al-Hindawi, A., Ng, T., Vizcaychipi, M.P.: Scoping review on the use of socially assistive robot technology in elderly care. BMJ Open **8**(2), e018815 (feb 2018). https://doi.org/10.1136/bmjopen-2017-018815
2. Abdollahi, H., Mahoor, M., Zandie, R., Sewierski, J., Qualls, S.: Artificial emotional intelligence in socially assistive robots for older adults: A pilot study. IEEE Transactions on Affective Computing pp. 1–1 (2022). https://doi.org/10.1109/taffc.2022.3143803
3. Ahmed, F., Bari, A.S.M.H., Gavrilova, M.L.: Emotion recognition from body movement. IEEE Access **8**, 11761–11781 (2020). https://doi.org/10.1109/ACCESS.2019.2963113
4. Ali, G., Ali, A., Ali, F., Draz, U., Majeed, F., Yasin, S., Ali, T., Haider, N.: Artificial neural network based ensemble approach for multicultural facial expressions analysis. IEEE Access **8**, 134950–134963 (2020). https://doi.org/10.1109/ACCESS.2020.3009908
5. Alonso-Martín, F., Malfaz, M., Sequeira, J., Gorostiza, J.F., Salichs, M.A.: A multimodal emotion detection system during human–robot interaction. Sensors **13**(11), 15549–15581 (2013). https://doi.org/10.3390/s131115549, `https://www.mdpi.com/1424-8220/13/11/15549`
6. Ardabili, S., Mosavi, A., Várkonyi-Kóczy, A.R.: Advances in machine learning modeling reviewing hybrid and ensemble methods pp. 215–227 (2020). https://doi.org/10.1007/978-3-030-36841-8_21
7. Banerjee, R., De, S., Dey, S.: A survey on various deep learning algorithms for an efficient facial expression recognition system. International Journal of Image and Graphics (dec 2021). https://doi.org/10.1142/S0219467822400058
8. Benamara, N.K., Val-Calvo, M., Álvarez-Sánchez, J.R., Díaz-Morcillo, A., Ferrández-Vicente, J.M., Fernández-Jover, E., Stambouli, T.B.: Real-time facial expression recognition using smoothed deep neural network ensemble. Integrated Computer-Aided Engineering **28**(1), 97–111 (dec 2020). https://doi.org/10.3233/ICA-200643

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_35

9. Bhatia, A., Rathee, A.: Multimodal emotion recognition. `https://github.com/ankurbhatia24/multimodal-emotion-recognition` (accessed 2023.01.31) (2020)
10. Birjali, M., Kasri, M., Beni-Hssane, A.: A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems **226**, 107134 (2021). https://doi.org/https://doi.org/10.1016/j.knosys.2021.107134, `https://www.sciencedirect.com/science/article/pii/S095070512100397X`
11. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., et al.: A database of german emotional speech. In: Interspeech. vol. 5, pp. 1517–1520 (2005)
12. Burnwal, S.: Speech emotion recognition. `https://www.kaggle.com/code/shivamburnwal/speech-emotion-recognition/notebook` (accessed 2023/01/31) (2020)
13. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation **42**, 335–359 (2008)
14. Canedo, D., Neves, A.: Mood estimation based on facial expressions and postures. In: Proceedings of the RECPAD. pp. 49–50 (2020)
15. Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: CREMA-d: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing **5**(4), 377–390 (oct 2014). https://doi.org/10.1109/TAFFC.2014.2336244
16. Chen, M., He, X., Yang, J., Zhang, H.: 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Processing Letters **25**(10), 1440–1444 (oct 2018). https://doi.org/10.1109/LSP.2018.2860246
17. Cheng, B., Wang, Y., Shao, D., Arora, C., Hoang, T., Liu, X.: Edge4emotion: An edge computing based multi-source emotion recognition platform for human-centric software engineering. In: 2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid). pp. 610–613 (2021). https://doi.org/10.1109/CCGrid51090.2021.00071
18. Ekman, P.: Facial expressions of emotion: New findings, new questions. Psychological Science **3**(1), 34–38 (jan 1992). https://doi.org/10.1111/j.1467-9280.1992.tb00253.x
19. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. Journal of Personality and Social Psychology **17**(2), 124–129 (1971). https://doi.org/10.1037/h0030377
20. Filntisis, P.P., Efthymiou, N., Koutras, P., Potamianos, G., Maragos, P.: Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction. IEEE Robotics and Automation Letters **4**(4), 4011–4018 (oct 2019). https://doi.org/10.1109/LRA.2019.2930434
21. Getson, C., Nejat, G.: Socially assistive robots helping older adults through the pandemic and life after COVID-19. Robotics **10**(3), 106 (sep 2021). https://doi.org/10.3390/robotics10030106
22. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., et al.: Challenges in representation learning: A report on three machine learning contests. In: International conference on neural information processing. pp. 117–124. Springer (2013)
23. Haq, S., Jackson, P.: Machine Audition: Principles, Algorithms and Systems, chap. Multimodal Emotion Recognition, pp. 398–423. IGI Global, Hershey PA (Aug 2010)
24. Heredia, J., Lopes-Silva, E., Cardinale, Y., Diaz-Amado, J., Dongo, I., Graterol, W., Aguilera, A.: Adaptive multimodal emotion detec-


ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_35

tion architecture for social robots. IEEE Access **10**, 20727–20744 (2022). https://doi.org/10.1109/ACCESS.2022.3149214
25. Jolly, E., Cheong, J.H., Xie, T., Byrne, S., Kenny, M., Chang, L.J.: Py-feat: Python facial expression analysis toolbox. arXiv preprint arXiv:2104.03509 (2021)
26. Jolly, E., Cheong, J.H., Xie, T., Byrne, S., Kenny, M., Chang, L.J.: Py-feat: Python facial expression analysis toolbox (2021). https://doi.org/https://doi.org/10.48550/arXiv.2104.03509
27. Jolly, E., Cheong, J.H., Xie, T., Byrne, S., Kenny, M., Chang, L.J.: Py-feat: Python facial expression analysis toolbox. `https://pythonrepo.com/repo/cosanlab-py-feat-python-deep-learning` (accessed 2023.01.31) (2023)
28. Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: A survey. IEEE Transactions on Affective Computing **4**(1), 15–33 (jan 2013). https://doi.org/10.1109/T-AFFC.2012.16
29. Kumaran, U., Rammohan, S.R., Nagarajan, S.M., Prathik, A.: Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep c-RNN. International Journal of Speech Technology **24**(2), 303–314 (jan 2021). https://doi.org/10.1007/s10772-020-09792-x
30. Li, S., Deng, W.: Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Transactions on Image Processing **28**(1), 356–370 (2019)
31. Liang, G., Wang, S., Wang, C.: Pose-aware adversarial domain adaptation for personalized facial expression recognition. arXiv preprint arXiv:2007.05932 (2020)
32. Livingstone, S.R., Russo, F.A.: 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 **13**(5), e0196391 (may 2018). https://doi.org/10.1371/journal.pone.0196391
33. Noroozi, F., Corneanu, C.A., Kamińska, D., Sapiński, T., Escalera, S., Anbarjafari, G.: Survey on emotional body gesture recognition. IEEE transactions on affective computing **12**(2), 505–523 (2018)
34. Novais, R., Cardoso, P.J.S., Rodrigues, J.M.F.: Emotion classification from speech by an ensemble strategy. In: ACM (ed.) 10th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2022) (2022)
35. Novais, R., Cardoso, P.J.S., Rodrigues, J.M.F.: Facial emotions classification supported in an ensemble strategy pp. 477–488 (2022). https://doi.org/10.1007/978-3-031-05028-2_32
36. Ortega, J.D.S., Cardinal, P., Koerich, A.L.: Emotion recognition using fusion of audio and video features. In: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). pp. 3847–3852 (2019). https://doi.org/10.1109/SMC.2019.8914655
37. Palanisamy, K., Singhania, D., Yao, A.: Rethinking cnn models for audio classification (2020). https://doi.org/10.48550/arXiv.2007.11154
38. Pecoraro, R., Basile, V., Bono, V.: Local multi-head channel self-attention for facial expression recognition. Information **13**(9), 419 (2022)
39. Pecoraro, R., Basile, V., Bono, V., Gallo, S.: Lhc-net: Local multi-head channel self-attention (code). `https://github.com/bodhis4ttva/lhc_net` (accessed 2023/01/29)
40. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_35

41. Pichora-Fuller, M.K., Dupuis, K.: Toronto emotional speech set (tess) (2020). https://doi.org/10.5683/SP2/E8H2MF
42. de Pinto, M.G.: Audio emotion classification from multiple datasets. `https://github.com/marcogdepinto/emotion-classification-from-audio-files` (accessed 2023/01/31) (2020)
43. de Pinto, M.G., Polignano, M., Lops, P., Semeraro, G.: Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients (may 2020). https://doi.org/10.1109/EAIS48028.2020.9122698
44. Popova, A.S., Rassadin, A.G., Ponomarenko, A.A.: Emotion recognition in sound pp. 117–124 (aug 2017). https://doi.org/10.1007/978-3-319-66604-4_18
45. Poria, S., Hazarika, D., Majumder, N., Mihalcea, R.: Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. IEEE Transactions on Affective Computing (2020)
46. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: MELD: A multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 527–536 (2019)
47. Revina, I., Emmanuel, W.S.: A survey on human face expression recognition techniques. Journal of King Saud University - Computer and Information Sciences **33**(6), 619–628 (jul 2021). https://doi.org/10.1016/j.jksuci.2018.09.002
48. Seknedy, M.E., Fawzi, S.: Speech emotion recognition system for human interaction applications (dec 2021). https://doi.org/10.1109/ICICIS52592.2021.9694246
49. Shenk, J., CG, A., Arriaga, O., Owlwasrowk: justinshenk/fer: Zenodo (Sep 2021). https://doi.org/10.5281/zenodo.5362356, `https://doi.org/10.5281/zenodo.5362356`
50. Siddiqui, M.F.H., Javaid, A.Y.: A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images. Multimodal Technologies and Interaction **4**(3), 46 (aug 2020). https://doi.org/10.3390/mti4030046
51. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information processing & management **45**(4), 427–437 (2009)
52. Sorrentino, A., Mancioppi, G., Coviello, L., Cavallo, F., Fiorini, L.: Feasibility study on the role of personality, emotion, and engagement in socially assistive robotics: A cognitive assessment scenario. Informatics **8**(2), 23 (mar 2021). https://doi.org/10.3390/informatics8020023
53. Stock-Homburg, R.: Survey of emotions in human–robot interactions: Perspectives from robotic psychology on 20 years of research. International Journal of Social Robotics **14**(2), 389–411 (jun 2021). https://doi.org/10.1007/s12369-021-00778-6
54. Wang, Z., Zeng, F., Liu, S., Zeng, B.: OAENet: Oriented attention ensemble for accurate facial expression recognition. Pattern Recognition **112**, 107694 (apr 2021). https://doi.org/10.1016/j.patcog.2020.107694
55. Zavaschi, T.H.H., Koerich, A.L., Oliveira, L.E.S.: Facial expression recognition using ensemble of classifiers (may 2011). https://doi.org/10.1109/ICASSP.2011.5946775
56. Zhang, F., Zhang, T., Mao, Q., Xu, C.: Joint pose and expression modeling for facial expression recognition (jun 2018). https://doi.org/10.1109/CVPR.2018.00354

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_35