Supplementary material

This supplementary material presents the baseline results for *fear* detection based on the intelligence engine proposed in [1] by using the physiological and speech data publicly available in WEMAC. In this baseline, 88 volunteers are employed and 12 of them are discarded due to the high unbalance of the fear labels' distribution, following a similar criterion as in [1]. Note that the fear detection results in [1] comprise the use of only 42 volunteers of the WEMAC database.

Detailed Results of Fear Classification

Based on the same mono-modal and data fusion architectures presented in [1], the same time arrangements used for the alignment of the physiological and speech signals (Bindi 1.0, Bindi 2.0a, Bindi 2.0b) are employed in this case. Figure 2 represents the F1-score (1) results by considering the 88 volunteers in WEMAC for the mono-modal fear detection systems (only based on physiological or speech data) in addition to the fusion strategies for the merging of both modalities. Both performance metrics are given to get more interpretable results from the slight imbalance between the binary *fear* labels.

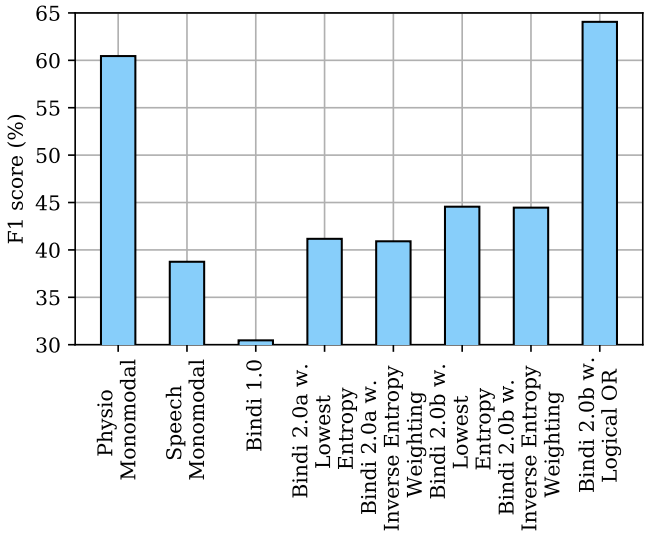

Figure 1

Figure 2. F1 score average performance analysis predicting over the 87 testing groups for the different architecture configurations. From left to right, the configurations are: physiological monomodal subsystem, the speech monomodal subsystem, Bindi 1.0, Bindi 2.0a with lowest entropy data fusion, Bindi 2.0a with inverse entropy weighting data fusion, Bindi 2.0b with lowest entropy data fusion, Bindi 2.0b with inverse entropy weighting data fusion, and Bindi 2.0b with logical OR data fusion. Note that Bindi 2.0a was not combined with logical OR data fusion because it is equivalent to Bindi 1.0.

		Physiological Monomodal	**Speech Monomodal**	**BINDI 1.0**	**Bindi 2.0a Lowest Entropy**	**Bindi 2.0a Inverse Entropy Weighting**	**Bindi 2.0b Lowest Entropy**	**Bindi 2.0b Inverse Entropy Weighting**	**Bindi 2.0b Logical OR**
F1-score	Mean	60.45	38.75	30.46	41.17	40.91	44.56	44.46	64.06
	Std	17.70	29.26	30.06	29.47	29.60	29.95	30.01	14.94

Table 1. Average performance analysis predicting over the 88 testing groups. Mean and standard deviations (Std).

Compared to the results in [1], we now double the amount of user data, adding 45 volunteers in these experiments. We use a *LASO* (Leave hAlf Subject Out) approach in which we train 88 models, one per each group or volunteers, using as fine-tuning data half of the data belonging to each user, and using as blind testing data the other half. Note that this is done with the intention of developing a general *fear* detection model personalized to each particular user.

As for the performances, these results are very similar to the ones achieved in [1], even slightly lower in some cases. This leads us to think that the addition of more data by doubling the number of users is still far from achieving higher and

more reliable rates for the detection of *fear*, leaving the door open for the research community to test new fusion methods, personalisation strategies for each user, study the correlation of temporal alignment of the two data modalities available, and test other suitable methodologies for this particular dataset.

References

1. Miranda, J. A. *et al.* Bindi: Affective internet of things to combat gender-based violence. *IEEE Internet Things J.* (2022).