Supplementary Material

1 EVALUATION WITH DIFFERENT FEATURE SETS

Four combinations of audio and video feature sets are evaluated and the key results are reported in the paper, additional results reported in this supplementary material are: eGemaps (audio) and geometric (video) in Table S1 for arousal; and Table S2 for valence; and BoAW (audio) and appearance (video) in Table S3 for arousal; and Table S4 for valence;

Table S1. Performance of **Arousal** prediction in terms of Unweighted Average Recall (UAR %) and weighted kappa k_w (reported inside parenthesis). **Audio feature: eGemaps; Video feature: Geometric**. The best performance across the mean values is indicated in bold.

		RankSVM			
		Audio	Video	Audio-Visual	Mean
OMSVM	Audio	60.0 (0.526)	51.9 (0.377)	60.2 (0.516)	$\overline{S^*_A}$ = **57.4 (0.473)**
	Video	39.1 (0.106)	37.1 (0.079)	39.5 (0.116)	$\overline{S^*_V} = 38.6(0.100)$
	Audio-Visual	53.9 (0.409)	48.7 (0.311)	51.3 (0.364)	$\overline{S^*_{AV}} = 51.3(0.361)$
	Mean	$\overline{S^A_*}$ = **51.0 (0.347)**	$\overline{S^V_*} = 45.9(0.256)$	$\overline{S^{AV}_*} = 50.3(0.332)$	-

Table S2. Performance of **Valence** prediction in terms of Unweighted Average Recall (UAR %) and weighted kappa k_w (reported inside parenthesis). **Audio feature: eGemaps; Video feature: Geometric**. The best performance across the mean values is indicated in bold.

		RankSVM			
		Audio	Video	Audio-Visual	Mean
OMSVM	Audio	41.4 (0.159)	38.9 (0.150)	40.9 (0.169)	$\overline{S^*_A} = 40.4(0.159)$
	Video	44.5 (0.202)	47.5 (0.252)	47.4 (0.264)	$\overline{S^*_V} = 46.5(0.239)$
	Audio-Visual	46.6 (0.234)	49.3 (0.288)	49.5 (0.277)	$\overline{S^*_{AV}}$ = **48.4 (0.266)**
	Mean	$\overline{S^A_*} = 44.2(0.198)$	$\overline{S^V_*} = 45.2(0.230)$	$\overline{S^{AV}_*}$ = **45.9 (0.236)**	-

2 EVALUATION WITH DIFFERENT THRESHOLDS

The thresholds utilised for interval to absolute ordinal labels (AOL) conversion in the experiments reported in the paper are $\{\theta_{a1} = -0.14, \theta_{a2} = 0.14\}$ for arousal, and $\{\theta_{v1} = 0, \theta_{v2} = 0.17\}$ for valence. These were chosen to provide a reasonably balanced distributions of the three states (low, medium, high) over the dataset. Here we report results from additional experiments conducted with slightly larger or lower thresholds to demonstrate that the inferences are not sensitive to the threshold values. These additional results are provided in Tables S7 - S10. Thresholds used in this section are: $\{\theta_{a1} = -0.13, \theta_{a2} = 0.13\}$ and $\{\theta_{a1} = -0.15, \theta_{a2} = 0.15\}$ for arousal; and $\{\theta_{v1} = -0.01, \theta_{v2} = 0.16\}$ and $\{\theta_{v1} = 0.01, \theta_{v2} = 0.18\}$ for valence. Class distributions in training and test sets based on the different thresholds are summarised in

Tables S5 - S6. The distributions based on the thresholds reported in the main paper is also repeated here for ease of comparison and indicated in bold. All experiments are carried out with the best performing feature sets based on the experimental results reported in Table 2 in the main paper. i.e., eGampes (audio) and appearance (video) are used for arousal; and BoAW (audio) and geometric (video) for valence prediction.

Table S3. Performance of **Arousal** prediction in terms of Unweighted Average Recall (UAR %) and weighted kappa k_w (reported inside parenthesis). **Audio feature: BoAW; Video feature: Appearance**. The best performance across the mean values is indicated in bold.

		RankSVM			
		Audio	Video	Audio-Visual	Mean
OMSVM	Audio	55.8 (0.461)	51.9 (0.428)	55.6 (0.455)	$\overline{S_A^*} =$ **55.5 (0.448)**
	Video	40.7 (0.189)	39.5 (0.159)	39.4 (0.181)	$\overline{S_V^*} = 39.9(0.176)$
	Audio-Visual	50.1 (0.357)	47.1 (0.322)	51.2 (0.371)	$\overline{S_{AV}^*} = 49.5(0.350)$
	Mean	$\overline{S_*^A} =$ **48.9 (0.336)**	$\overline{S_*^V} = 47.2(0.303)$	$\overline{S_*^{AV}} = 48.7(0.336)$	-

Table S4. Performance of **Valence** prediction in terms of Unweighted Average Recall (UAR %) and weighted kappa k_w (reported inside parenthesis). **Audio feature: eGemaps; Video feature: Geometric**. The best performance across the mean values is indicated in bold.

		RankSVM			
		Audio	Video	Audio-Visual	Mean
OMSVM	Audio	42.9 (0.179)	45.2 (0.214)	43.7 (0.203)	$\overline{S_A^*} = 43.9(0.199)$
	Video	46.6 (0.266)	43.3 (0.200)	44.6 (0.234)	$\overline{S_V^*} =$ **44.8 (0.233)**
	Audio-Visual	44.0 (0.214)	44.6 (0.211)	45.2 (0.222)	$\overline{S_{AV}^*} = 44.6(0.216)$
	Mean	$\overline{S_*^A} = 44.5(0.219)$	$\overline{S_*^V} = 44.3(0.208)$	$\overline{S_*^{AV}} =$ **44.5 (0.220)**	-

Table S5. Arousal absolute ordinal labels distribution on RECOLA dataset with different thresholds. Thresholds reported in the main manuscript is indicated in bold.

	θ_{a1}, θ_{a2}	Low	Medium	High
Training set	[-0.13, 0.13]	378	365	580
	[-0.14, 0.14]	**363**	**443**	**526**
	[-0.15, 0.15]	601	290	441
Test set	[-0.13, 0.13]	344	523	456
	[-0.14, 0.14]	**578**	**348**	**406**
	[-0.15, 0.15]	554	422	356

Table S6. Valence absolute ordinal labels distribution on RECOLA dataset with different thresholds. Thresholds reported in the main manuscript is indicated in bold.

	θ_{v1}, θ_{v2}	Low	Medium	High
Training set	[-0.01, 0.16]	344	578	419
	[0, 0.17]	**462**	**463**	**416**
	[0.1, 0.18]	511	437	393
Test set	[-0.01, 0.16]	443	520	378
	[0, 0.17]	**545**	**432**	**364**
	[0.1, 0.18]	586	424	311

Table S7. Performance of **Arousal** prediction in terms of Unweighted Average Recall (UAR %) and weighted kappa k_w (reported inside parenthesis). $\theta_{a1} = -0.13$ and $\theta_{a2} = 0.13$. The best performance across the mean values is indicated in bold.

		RankSVM			
		Audio	Video	Audio-Visual	Mean
OMSVM	Audio	57.2 (0.485)	54.8 (0.457)	56.2 (0.481)	$\overline{S_A^*} =$ **56.1 (0.474)**
	Video	41.6 (0.182)	39.2 (0.158)	41.0 (0.175)	$\overline{S_V^*} = 40.6(0.172)$
	Audio-Visual	53.1 (0.437)	50.4 (0.369)	51.6 (0.397)	$\overline{S_{AV}^*} = 51.7(0.401)$
	Mean	$\overline{S_*^A} =$ **50.6 (0.368)**	$\overline{S_*^V} = 48.1(0.328)$	$\overline{S_*^{AV}} = 49.6(0.351)$	-

Table S8. Performance of **Arousal** prediction in terms of Unweighted Average Recall (UAR %) and weighted kappa k_w (reported inside parenthesis). $\theta_{a1} = -0.15$ and $\theta_{a2} = 0.15$. The best performance across the mean values is indicated in bold.

		RankSVM			
		Audio	Video	Audio-Visual	Mean
OMSVM	Audio	58.6 (0.471)	53.7 (0.392)	56.7 (0.459)	$\overline{S_A^*} =$ **56.3 (0.441)**
	Video	42.1 (0.210)	39.4 (0.157)	40.7 (0.184)	$\overline{S_V^*} = 40.8(0.184)$
	Audio-Visual	49.2 (0.315)	46.3 (0.280)	48.9 (0.309)	$\overline{S_{AV}^*} = 48.1(0.301)$
	Mean	$\overline{S_*^A} =$ **50.0 (0.332)**	$\overline{S_*^V} = 46.5(0.276)$	$\overline{S_*^{AV}} = 48.8(0.317)$	-

Table S9. Performance of **Valence** prediction in terms of Unweighted Average Recall (UAR %) and weighted kappa k_w (reported inside parenthesis). $\theta_{v1} = -0.01$ and $\theta_{v2} = 0.16$. The best performance across the mean values is indicated in bold.

		RankSVM			
		Audio	Video	Audio-Visual	Mean
OMSVM	Audio	40.6 (0.121)	42.9 (0.169)	44.4 (0.172)	$\overline{S_A^*} = 42.6(0.154)$
	Video	48.1 (0.257)	45.5 (0.217)	45.6 (0.225)	$\overline{S_V^*} = 46.4(0.233)$
	Audio-Visual	50.1 (0.293)	47.6 (0.254)	48.4 (0.269)	$\overline{S_{AV}^*} =$ **48.7 (0.272)**
	Mean	$\overline{S_*^A} =$ **46.2 (0.224)**	$\overline{S_*^V} = 45.3(0.213)$	$\overline{S_*^{AV}} = 46.1(0.222)$	-

Table S10. Performance of **Valence** prediction in terms of Unweighted Average Recall (UAR %) and weighted kappa k_w (reported inside parenthesis). $\theta_{v1} = 0.01$ and $\theta_{v2} = 0.18$. The best performance across the mean values is indicated in bold.

		RankSVM			
		Audio	Video	Audio-Visual	Mean
OMSVM	Audio	39.6 (0.134)	44.7 (0.219)	45.0 (0.219)	$\overline{S_A^*} = 43.1(0.191)$
	Video	48.4 (0.273)	46.7 (0.238)	47.4 (0.257)	$\overline{S_V^*} = 47.5(0.256)$
	Audio-Visual	50.7 (0.332)	49.0 (0.282)	50.1 (0.309)	$\overline{S_{AV}^*} =$ **49.9 (0.308)**
	Mean	$\overline{S_*^A} = 46.2(0.247)$	$\overline{S_*^V} = 46.8(0.246)$	$\overline{S_*^{AV}} =$ **47.5 (0.262)**	-