An Open-source Benchmark of Deep Learning Models for Audio-visual Apparent and Self-reported Personality Recognition: Supplementary Material

Rongfan Liao, Siyang Song*, and Hatice Gunes

Abstract—This paper presents the supplementary materials to the first reproducible audio-visual benchmarking framework, which provides a fair and consistent evaluation of eight existing personality computing models (e.g., audio, visual and audio-visual) and seven standard deep learning models on both self-reported and apparent personality recognition tasks. Specifically, we provide a set of additional ablation experiments, statistical significance analysis, discussions of the poor performances in self-reported personality recognition, as well as more details of benchmarked models and datasets. We make all the code and settings of this personality computing benchmark publicly available at https://github.com/liaorongfan/DeepPersonality.

Index Terms—Self-reported (true) personality recognition, Apparent personality (impression) recognition, Audio-visual personality computing benchmark, Spatio-temporal modelling, Deep Learning

1 Experimental results

1.1 10-fold cross-validation results

We further conducted 10-fold cross-validation for six best benchmarked models (i.e., two audio models: CRNet and VGGish; two visual models: HRNet and VAT; and two audio-visual models: CRNet and Amb-Fac-VGGish) for both apparent and self-reported personality recognition. The results are provided in Table 1 and Table 2. The statistical significance analysis of between these approaches are provided in Table 9 and Table 10.

1.2 Additional ablation studies

Single-target VS. Multi-target: Fig. 1, Table 3 and Table 4 show that almost all of the employed models that are trained to jointly predict five traits outperformed their corresponding variants that individually predicted each trait for both apparent and self-reported personality traits recognition (except VAT model that is trained to individually predict each trait achieved better performance for self-reported personality recognition).

Short VS. Long temporal contexts: We employ three temporal visual models (3DResNet, TPN and VAT) and three audio models (CRNet, ResNet and VGGish) from our benchmark to evaluate the impact of temporal scale on the personality recognition performance. Specifically, we experimented with four different temporal scales, including setting the length of each input segment as 0.53, 1.07, 2.13 seconds (i.e., 16, 32, and 64 visual frames), as well as using spectral representation [5] to obtain clip-level temporal information. It can be observed from Table 5 and Table 6 that the results are largely influenced by the temporal scale (especially for apparent personality recognition), where the optimal temporal scales are model-dependent (e.g., 1.07s is the optimal temporal scale for multiple visual models while in terms of audio using the full clip frequently produced the best performance).

Metadata VS. No metadata: We also evaluate the usefulness of the metadata. As shown in Fig. 2 and Table 7, while the ChaLearn Impression dataset does not provide metadata (the additional gender and ethnicity labels are no longer available in the website), the UVIDA dataset provides four types of metadata (i.e., gender, age, country and education level). Thus, we only evaluate the usefulness of metadata on self-reported personality recognition based on UDIVA. In particular, we integrate the metadata as additional features for the last fully connected layer for each model. Fig. 2 and Table 7 also show that adding metadata did not clearly improve the self-reported personality recognition results in general despite these metadata slightly increasing some models' performance. We attribute this lack of improvement to the already poor performance of behaviour-based self-reported personality recognition, i.e., the extracted features contain a lot of task-unrelated noises, and thus additional consideration of metadata of subjects can not compensate for the negative impacts of such noises.

Influence of the task contents: Table 8 presents a comparison of the self-reported personality recognition results achieved by different tasks on the UDIVA dataset. Each deep learning model is trained with behaviours expressed under

- *Rongfan Liao is with SONY China Software Center. E-mail: rongfan.liao@sony.com*
- *Siyang Song and Hatice Gunes are with the AFAR Lab, Department of Computer Science and Technology, University of Cambridge, Cambridge, CB3 0FT, United Kingdom. E-mail: ss2796@cam.ac.uk, Hatice.Gunes@cl.cam.ac.uk (* Corresponding Author: Siyang Song, E-mail: ss2796@cam.ac.uk)*

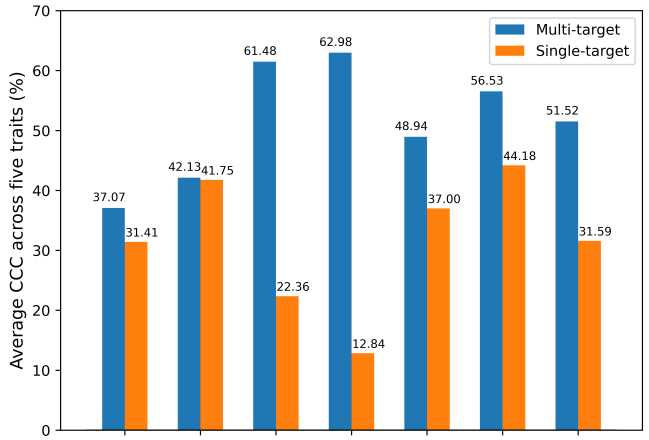
Manuscript received September 18, 2022.

	Traits	Open	Consc	Extrav	Agree	Neuro	Avg.
Audio	CRNet	0.3729 ± 0.0728	0.2797 ± 0.0704	0.3407 ± 0.0943	0.2522 ± 0.0674	0.3901 ± 0.0574	0.3271 ± 0.0724
	VGGish	0.4292 ± 0.0476	0.4505 ± 0.0794	0.4177 ± 0.0685	0.2828 ± 0.0599	0.4419 ± 0.0445	0.4044 ± 0.0477
Visual	HRNet	0.3451 ± 0.0092	0.4268 ± 0.0593	0.3809 ± 0.0153	0.2465 ± 0.0272	0.3129 ± 0.0269	0.3424 ± 0.0271
	VAT	0.3973 ± 0.0219	0.4063 ± 0.0161	0.4268 ± 0.0281	0.2397 ± 0.0338	0.3805 ± 0.0275	0.3701 ± 0.0211
Aud-vis	CRNet	0.4740 ± 0.0140	0.4627 ± 0.0106	0.4692 ± 0.0357	0.3637 ± 0.0477	0.4709 ± 0.0153	0.4481 ± 0.0232
	Amb-Fac-VGGish	0.4009 ± 0.0057	0.5140 ± 0.0061	0.4239 ± 0.0085	0.2945 ± 0.0416	0.3981 ± 0.0150	0.4063 ± 0.0054

TABLE 1
The 10-fold cross-validation CCC results achieved for the apparent personality recognition on the ChaLearn 2016 First impression dataset. This table summaries the **average CCC values**.

	Traits	Open	Consc	Extrav	Agree	Neuro	Avg.
Audio	CRNet	-0.0075 ± 0.0282	-0.0050 ± 0.0030	0.0006 ± 0.0024	-0.0180 ± 0.0160	-0.0009 ± 0.0168	-0.0062 ± 0.0049
	VGGish	0.0293 ± 0.0399	0.0584 ± 0.0136	0.0532 ± 0.1046	0.0507 ± 0.0333	0.0226 ± 0.0610	0.0428 ± 0.0065
Visual	HRNet	0.3074 ± 0.3791	-0.0633 ± 0.2614	0.1834 ± 0.2631	0.1267 ± 0.1628	-0.0763 ± 0.1924	0.0956 ± 0.0310
	VAT	0.0024 ± 0.0339	0.0528 ± 0.1255	0.0047 ± 0.0077	-0.0228 ± 0.0362	0.0243 ± 0.0184	0.0123 ± 0.0164
Aud-vis	CRNet	0.0347 ± 0.1427	0.1655 ± 0.1477	-0.0907 ± 0.0220	-0.0228 ± 0.0541	-0.0268 ± 0.0649	0.0120 ± 0.0802
	Amb-Fac-VGGish	0.0735 ± 0.0717	0.0600 ± 0.0534	0.0275 ± 0.0469	0.0316 ± 0.0319	0.0069 ± 0.0195	0.0399 ± 0.0030

TABLE 2
The 10-fold cross-validation CCC results achieved for the self-reported personality recognition on the UDIVA dataset. This table summaries the **average CCC values**.



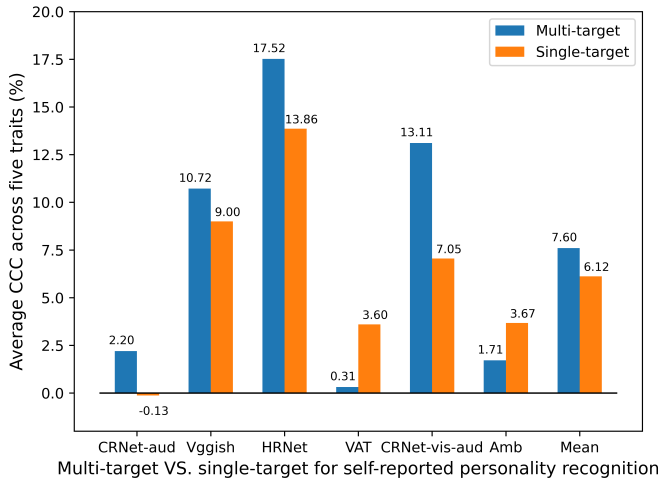


Fig. 1. The average CCC results of the five traits achieved from both single-target and multi-target systems on ChaLearn First Impression and UDIVA datasets

	Traits	Open	Consc	Extrav	Agree	Neuro	Avg.
Audio	CRNet [1]	0.4122	0.3406	0.3846	0.2857	0.4306	0.3707
	CRNet [1]	0.3846	0.1504	0.4136	0.2041	0.4177	0.3141
	VGGish [2]	0.4516	0.4493	0.4429	0.3127	0.4500	0.4213
	VGGish [2]	0.4441	0.4563	0.4384	0.3097	0.4389	0.4175
Visual	HRNet [3]	0.5923	0.6912	0.6436	0.5195	0.6273	0.6148
	HRNet [3]	0.2332	0.3223	0.2589	0.0806	0.2230	0.2236
	VAT [4]	0.6216	0.6753	0.6836	0.5228	0.6456	0.6298
	VAT [4]	0.1391	0.1636	0.1549	0.0526	0.1321	0.1284
Aud-vis	CRNet [1]	0.5193	0.5106	0.5024	0.4026	0.5119	0.4894
	CRNet [1]	0.3991	0.4243	0.3952	0.2018	0.4298	0.3700
	Amb-Fac-VGGish [2]	0.5618	0.6421	0.5921	0.4620	0.5734	0.5663
	Amb-Fac-VGGish [2]	0.3968	0.5318	0.4667	0.2821	0.3814	0.4118

TABLE 3
The CCC results achieved for the apparent personality recognition on the ChaLearn First Impression dataset. The model names in bold format denote they are trained from single traits and those in plain format are models trained for jointly predicting 5 traits together.

	Traits	Open	Consc	Extrav	Agree	Neuro	Avg.
Audio	CRNet [1]	0.0005	0.0290	0.0201	0.0335	0.0267	0.0220
	CRNet [1]	0.0031	0.0121	0.0058	-0.0280	0.0007	-0.0013
	VGGish [2]	0.0688	0.1882	0.1310	0.1069	0.0412	0.1072
	VGGish [2]	-0.0062	0.1473	0.0669	0.1422	0.1003	0.0900
Visual	HRNet [3]	0.2175	0.2998	-0.0039	0.1680	0.1945	0.1752
	HRNet [3]	0.1450	0.3307	-0.0007	-0.0112	0.2294	0.1386
	VAT [4]	-0.0139	-0.0016	0.0016	-0.0013	-0.0006	-0.0031
	VAT [4]	0.0912	0.0277	0.0571	0.0009	0.0032	0.0360
Aud-vis	CRNet [1]	0.0998	0.1780	0.1158	0.2168	0.0449	0.1311
	CRNet [1]	0.0000	0.0000	0.0001	0.3527	0.0000	0.0705
	Amb-Fac-VGGish [2]	-0.0348	0.0468	0.0302	0.0397	0.0041	0.0171
	Amb-Fac-VGGish [2]	0.0311	0.0371	0.0068	0.0837	0.0247	0.0367

TABLE 4
The CCC results achieved for self-reported personality recognition on the UDIVA dataset. The model names in bold format denote they are trained from single traits and those in plain format are models trained for jointly predicting 5 traits together.

a specific task. The results show that there are fluctuations across the results achieved for different tasks, while the combination of behaviours expressed from all tasks does not offer significant advantages for self-reported personality recognition models on the UDIVA dataset.

	Length	Open	Consc	Extrav	Agree	Neuro	Avg.
3DResNet (C)	0.53 s (16 frames)	0.3031	0.3290	0.3185	0.1871	0.2725	0.2820
	1.07 s (32 frames)	**0.3248**	**0.3601**	**0.3601**	**0.2120**	**0.3352**	**0.3185**
	2.13 s (64 frames)	0.1123	0.1333	0.1234	0.0451	0.1065	0.1041
	Full clip	0.0168	0.0298	-0.0017	-0.009	0.0096	0.0091
TPN (C)	0.53 s (16 frames)	0.1680	0.1909	0.1572	0.0552	0.1447	0.1432
	1.07 s (32 frames)	**0.4427**	**0.4767**	**0.4998**	**0.3230**	**0.4675**	**0.4420**
	2.13 s (64 frames)	0.0079	-0.0304	0.0079	-0.0116	-0.0172	-0.0087
	Full clip	0.4088	0.4280	0.4537	0.2936	0.4164	0.4001
VAT (C)	0.53 s (16 frames)	0.2293	0.2826	0.2671	0.1415	0.2397	0.2320
	1.07 s (32 frames)	**0.6216**	**0.6753**	**0.6836**	**0.5228**	**0.6456**	**0.6298**
	2.13 s (64 frames)	0.5598	0.5819	0.6413	0.4728	0.5891	0.5690
	Full clip	0.5491	0.6109	0.6024	0.4301	0.5623	0.5510
3DResNet (U)	0.53 s (16 frames)	-0.1033	-0.0316	-0.0054	-0.0940	**0.0580**	-0.0352
	1.07 s (32 frames)	-0.0478	0.0102	0.0478	**0.0499**	-0.0240	0.0072
	2.13 s (64 frames)	**0.0375**	**0.0137**	**0.2120**	-0.1245	0.0400	**0.0357**
	Full clip	0.0245	-0.0164	-0.0055	-0.0204	0.0084	-0.0019
TPN (U)	0.53 s (16 frames)	-0.0136	-0.0490	**0.0585**	-0.1084	0.0538	-0.0117
	1.07 s (32 frames)	**0.0448**	0.0348	0.0287	-0.0177	-0.0281	**0.0125**
	2.13 s (64 frames)	-0.0007	0.0067	-0.0328	**0.0388**	0.0078	0.0040
	Full clip	-0.0494	**0.0355**	-0.0028	0.0352	**0.0125**	0.0062
VAT (U)	0.53 s (16 frames)	-0.0018	0.0097	0.0035	0.004	0.0031	0.0037
	1.07 s (32 frames)	-0.0139	-0.0016	0.0016	-0.0013	-0.0006	-0.0031
	2.13 s (64 frames)	-0.0072	**0.0142**	0.0056	**0.0254**	0.0092	**0.0095**
	Full clip	**0.0469**	-0.0483	**0.0212**	-0.0682	**0.0894**	0.0082

TABLE 5
The CCC results of visual models achieved for apparent personality recognition on the ChaLearn First Impression dataset ("(C)") and for self-reported personality recognition on the UDIVA dataset ("(U)") based on spatio-temporal visual models, where different length settings for the input video segment are evaluated.

	Length	Open	Consc	Extrav	Agree	Neuro	Avg.
VGGish (C)	0.53 s	0.3251	0.3102	0.3427	0.1777	0.3151	0.2942
	1.07 s	0.3783	0.3547	0.3894	0.2189	0.356	0.3395
	2.13 s	0.3947	0.3846	0.4035	0.2577	0.3947	0.3670
	Full clip	**0.4516**	**0.4493**	**0.4429**	**0.3127**	**0.4500**	**0.4213**
CRNet (C)	0.53 s	0.2202	0.1071	0.2180	0.1000	0.1939	0.1678
	1.07 s	0.2449	0.1320	0.2513	0.1076	0.2226	0.1917
	2.13 s	0.3136	0.2015	0.2998	0.1654	0.2832	0.2527
	Full clip	**0.4122**	**0.3406**	**0.3846**	**0.2857**	**0.4306**	**0.3707**
ResNet (C)	0.53 s	0.1026	0.0703	0.073	0.0565	0.1211	0.0847
	1.07 s	0.1504	**0.0831**	0.1145	0.0700	**0.1562**	0.1148
	2.13 s	**0.1570**	0.0792	**0.1382**	0.0887	0.1514	**0.1229**
	Full clip	0.1293	0.0830	0.0458	**0.1101**	0.1548	0.1046
VGGish (U)	0.53 S	0.0388	0.0847	0.1128	0.0218	0.0409	0.0598
	1.07 S	-0.0502	0.0966	0.1028	**0.1105**	0.1505	0.0820
	2.13 S	-0.0801	**0.2605**	0.0847	0.0438	**0.1875**	0.0993
	Full clip	**0.0688**	0.1882	**0.1310**	0.1069	0.0412	**0.1072**
CRNet (U)	0.53 S	-0.0024	-0.0218	0.0040	-0.008	-0.0577	-0.0172
	1.07 S	0.0017	0.0050	0.0079	0.0024	0.0140	0.0055
	2.13 S	**0.3042**	**0.1939**	**0.2952**	**0.1554**	**0.2823**	**0.2462**
	Full clip	0.0005	0.0290	0.0201	0.0335	0.0267	0.0220
ResNet (U)	0.53 S	**-0.0011**	0.0066	-0.0003	0.0009	0.0005	0.0013
	1.07 S	-0.0216	0.0209	**0.0065**	0.0035	-0.0001	0.0018
	2.13 S	-0.0075	0.0567	0.0048	0.0017	-0.0002	0.0111
	Full clip	-0.0459	**0.1045**	-0.0416	**0.0429**	**0.0015**	**0.0123**

TABLE 6
The CCC results of audio models achieved for apparent personality recognition on the ChaLearn First Impression dataset ("(C)") and for self-reported personality recognition on the UDIVA dataset ("(U)") based on spatio-temporal visual models, where different length settings for the input video segment are evaluated.

1.3 Statistical significance analysis

In this section, we provide statistical significance analyses for the following experiments, where we found that the differences in most settings brought much more impacts on apparent personality recognition performances over the self-reported personality recognition performances. This can be explained by the fact that the benchmarked models can hardly infer self-reported personality traits from human non-verbal behaviours, and thus no matter what the settings are, they always achieved very unreliable predictions.

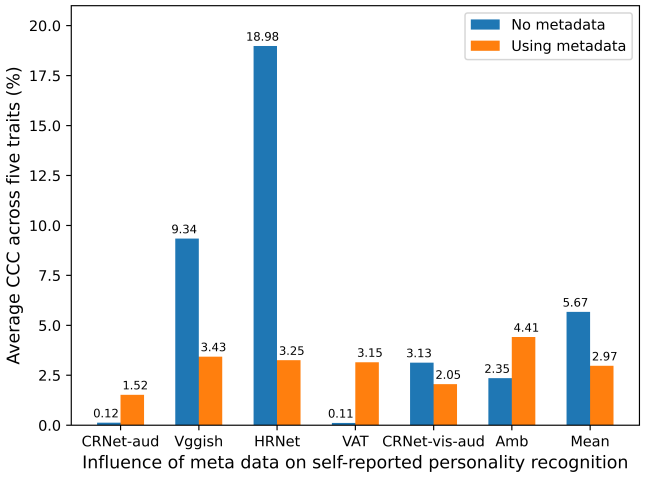


Fig. 2. The average CCC results of the five traits achieved from self-reported personality recognition models using or without using metadata on UDIVA dataset.

	Traits	Open	Consc	Extrav	Agree	Neuro	Avg.
Audio	**CRNet [1]**	-0.0588	0.0777	0.0027	0.03	0.0245	0.0152
	CRNet [1]	0.0012	0.0091	0.0045	-0.0062	-0.0027	0.0012
	Vggish [2]	-0.0505	0.105	0.0316	0.0474	0.0379	0.0343
	VGGish [2]	0.0124	0.1351	0.1162	0.1353	0.0682	0.0934
Visual	**HRNet [3]**	0.0566	0.0401	0.0226	0.0288	0.0146	0.0325
	HRNet [3]	0.3715	0.3551	-0.1155	0.1266	0.2115	0.1898
	VAT [4]	-0.1511	0.1823	0.0168	0.0379	0.0716	0.0315
	VAT [4]	-0.0069	0.0045	0.0019	-0.0003	0.0064	0.0011
Aud-vis	**CRNet [1]**	0.016	0.0455	0.0049	0.0141	0.0223	0.0205
	CRNet [1]	0.1291	0.0164	0.0758	0.0402	-0.105	0.0313
	Amb-Fac-VGGish [2]	-0.0778	0.1931	0.0086	-0.0173	0.114	0.0441
	Amb-Fac-VGGish [2]	-0.0577	0.1475	0.0062	0.0066	0.0148	0.0235

TABLE 7
The CCC results achieved for self-reported personality recognition on the UDIVA dataset. The model names in bold format denote they are using metadata and those in plain format are models do not use metadata.

- (i) Table 9 and Table 10 compare the performances of the six best benchmarked models in terms of their 10-fold cross-validation results, where we individually compare the performances achieved by top-2 models corresponding to each modality. It can be seen that the apparent personality recognition results achieved by top-2 visual and audio systems for almost all traits are significantly different, while the apparent personality recognition results achieved by audio-visual CRNet and Amb-VGGish are similar on all traits. Meanwhile, the self-reported personality recognition results achieved by unimodal systems are significantly different on recognising the Agreeableness trait.
- (ii) Table 11 and Table 12 compare the results achieved by models of different modalities (i.e., audio VS. visual VS. audio-visual models), where we use the CCC results achieved on standard test set to represent each modality. It is clear that apparent personality recognition results achieved between audio and other settings (visual and audio-visual) are

	Traits	Open	Consc	Extrav	Agree	Neuro	Avg.
Audio	CRNet (A)	-0.0006	0.0017	-0.0018	0.0002	0.0	-0.0001
	CRNet (L)	-0.0002	0.0004	-0.0004	0.0000	-0.0002	-0.0001
	CRNet (G)	-0.0172	0.0445	0.0233	0.0935	0.1945	0.0677
	CRNet (T)	0.0199	0.0696	0.0594	0.0404	-0.0875	0.0204
	CRNet (All)	0.0012	0.0091	0.0045	-0.0062	-0.0027	0.0012
	VGGish (A)	0.1097	0.2187	0.1272	0.1827	0.0949	0.1467
	VGGish (L)	-0.0841	0.1474	0.0804	0.0454	0.0	0.0378
	VGGish (G)	0.3098	0.2403	0.2324	0.1882	0.0698	0.2081
	VGGish (T)	-0.0601	0.1463	0.084	0.0112	0.0	0.0363
	VGGish (All)	0.0124	0.1351	0.1162	0.1353	0.0682	0.0934
Visual	HRNet (A)	0.0443	0.2446	-0.0621	0.1355	0.0964	0.0918
	HRNet (L)	0.2741	0.2739	-0.0384	0.0743	0.1481	0.1464
	HRNet (G)	0.3780	0.3317	0.0064	0.2090	0.3315	0.2513
	HRNet (T)	0.1735	0.349	0.0785	0.2534	0.202	0.2113
	HRNet (All)	0.3715	0.3551	-0.1155	0.1266	0.2115	0.1898
	VAT(A)	-0.0079	0.0342	-0.0103	-0.0398	0.0039	-0.004
	VAT(L)	-0.008	0.0095	-0.01	0.0058	0.0136	0.0022
	VAT (G)	0.0117	0.0453	-0.0093	-0.0031	-0.0112	0.0067
	VAT (T)	-0.0031	0.0073	-0.0082	0.0	0.0019	-0.0004
	VAT (All)	-0.0069	0.0045	0.0019	-0.0003	0.0064	0.0011
Aud-vis	CRNet(A)	-0.0517	0.1452	0.2303	0.4313	-0.0341	0.1442
	CRNet(L)	0.322	0.056	-0.0255	0.2547	0.065	0.1345
	CRNet(G)	0.1416	0.1284	0.1104	0.1827	0.1386	0.1403
	CRNet (T)	-0.0126	0.3825	0.1482	-0.0014	0.01	0.1053
	CRNet (All)	0.1291	0.0164	0.0758	0.0402	-0.105	0.0313
	Amb-VGGish (A)	-0.0741	-0.1113	0.0513	0.0258	0.0004	-0.0216
	Amb-VGGish (L)	0.0	0.0002	0.0	0.0032	0.0	0.0007
	Amb-VGGish (G)	-0.0001	0.055	0.0	0.0	0.0	0.011
	Amb-VGGish (T)	-0.1267	0.082	0.025	0.1336	0.0008	0.0229
	Amb-VGGish (All)	-0.0577	0.1475	0.0062	0.0066	0.0148	0.0235

TABLE 8
The CCC results achieved for the self-reported personality recognition on UDIVA dataset from "Animal", "Lego", "Ghost" and "Talk" sessions, where we use "(A)", "(L)", "(G)" and "(T)" to denote the four sessions, respectively. The indicator "(All)" represents the four sessions are concatenated as a single clip to train a target model.

Modality	Open	Consc	Extrav	Agree	Neuro	Mean
(i) CRNet-aud VS. VGGish	$+(***)$	$+(***)$	$+(***)$	$-$	$+(*)$	$+(***)$
(ii) HRNet VS. VAT	$+(***)$	$+(***)$	$+(***)$	$+(***)$	$+(***)$	$+(***)$
(iii) CRNet VS. Amb-VGGish	$-$	$-$	$-$	$-$	$-$	$-$

TABLE 9
Statistical significance testing results achieved for apparent personality recognition task in terms of CCC achieved by: (i) top-2 audio systems; (ii) top-2 visual systems; and (iii) top-2 audio-visual systems. Here, **+** / **-** denotes that there is / there is no statistically significant difference between a pair of systems (The significance level of $* P < 0.05$, $** P < 0.01$, $*** P < 0.001$). To conduct the T-Test, we used the 10-fold cross-validation results on the ChaLearn First Impression dataset.

Modality	Open	Consc	Extrav	Agree	Neuro	Mean
(i) CRNet-aud VS. VGGish	$-$	$+(**)$	$-$	$+(***)$	$-$	$+(***)$
(ii) HRNet VS. VAT	$-$	$-$	$-$	$+(**)$	$-$	$+(*)$
(iii) CRNet VS. Amb-VGGish	$-$	$-$	$+(*)$	$-$	$-$	$-$

TABLE 10
Statistical significance testing results achieved for self-reported personality recognition task in terms of CCC achieved by: (i) top-2 audio systems; (ii) top-2 visual systems; and (iii) top-2 audio-visual systems. Here, **+** / **-** denotes that there is / there is no statistically significant difference between a pair of systems (The significance level of $* P < 0.05$, $** P < 0.01$, $*** P < 0.001$). To conduct the T-Test, we used the 10-fold cross-validation results on the UDIVA dataset.

Models	Open	Consc	Extrav	Agree	Neuro	Mean
Audio VS. Visual	$+(**)$	$+(***)$	$+(**)$	$+(***)$	$+(**)$	$+(**)$
Audio VS. Audio-visual	$+(**)$	$+(**)$	$+(**)$	$+(**)$	$+(**)$	$+(**)$
Visual VS. Audio-visual	$-$	$+(*)$	$-$	$+(*)$	$-$	$+(*)$

TABLE 11
Statistical significance testing results achieved for apparent personality recognition task in terms of CCC, where **+** / **-** denotes that there is / there is no statistically significant difference between a pair of systems (The significance level of $* P < 0.05$, $** P < 0.01$, $*** P < 0.001$). To conduct the T-Test, we compared results achieved by each pair of modalities (their Top-5 systems) on the ChaLearn First Impression dataset.

significantly different, as visual/audio-visual models generally achieved much better apparent personality recognition performances than audio models. Meanwhile, models of all three settings achieved less promising results on self-reported personality recognition, and thus they have similar performances on recognising most self-reported personality traits.

- (iii) Table 13 compares models that take full frame vs. models that take cropped face frame as the input, where we found that the influence of this setting is very low on both apparent and self-reported personality recognition results.
- (iv) Table 14 compares the results achieved between spatial visual models and spatio-temporal visual models, where we surprisingly found that this setting has a large impact on self-reported personality recognition but much smaller impact on apparent personality recognition.
- (v) Table 15 compares short segment-level VS. clip-level personality modelling (models that use SFP VS. without SFP for aggregating frame/segment-level predictions), where we found that with SFP, the recognition performance of all five apparent personality traits have been significantly increased. In contrast, the SFP did not significantly influence the self-reported personality recognition.
- (vi) Table 16 and Table 17 compare the advantage of the employed SFP over other clip-level modelling strategies (i.e., clip-level predictions aggregation strategies). It can be observed that the SFP brings more advantages over other strategies on apparent personality recognition than self-reported personality recognition.
- (vii) Table 18 compares systems that individually predict each trait with the systems that jointly predict five traits. Again, this setting brought significant differences in results achieved for apparent personality recognition on all traits while having very limited impact on self-reported personality recognition.
- (viii) Table 19 compares systems that additionally consider metadata of subjects with systems that do not use metadata on UDIVA dataset (i.e., ChaLearn First Impression dataset did not provide metadata). It can be seen that the metadata only brought a significant difference in recognising the Openness trait.
- (ix) Table 20 compares the systems trained on behaviours expressed by each single task with the systems trained on behaviours expressed by multiple tasks on UDIVA dataset. As we can see, this factor has a very small impact on recognising most traits.

Models	Open	Consc	Extrav	Agree	Neuro	Mean
Audio VS. Visual	–	+(**)	–	–	–	+(*)
Audio VS. Audio-visual	–	–	–	–	–	–
Visual VS. Audio-visual	–	+(**)	–	–	–	+(*)

TABLE 12
Statistical significance testing results achieved for self-reported personality recognition task in terms of CCC, where **+** / **-** denotes that there is / there is no statistically significant difference between a pair of systems (The significance level of $* P < 0.05$, $** P < 0.01$, $*** P < 0.001$). To conduct the T-Test, we compared results achieved by each pair of modalities (their Top-5 systems) on the UDIVA dataset.

	Open	Consc	Extrav	Agree	Neuro	Mean
APR	–	–	–	–	–	–
SPR	–	–	–	–	–	–

TABLE 13
Statistical significance testing results achieved for visual models that take full frames or cropped face frames, where **+** / **-** denotes that there is / there is no statistically significant difference between a pair of systems (The significance level of $* P < 0.05$, $** P < 0.01$, $*** P < 0.001$). **APR** represents the results achieved for apparent personality recognition models while **SPR** represents the results achieved for self-reported personality recognition models.

	Open	Consc	Extrav	Agree	Neuro	Mean
APR	–	–	–	–	–	–
SPR	+(**)	+(***)	–	+(**)	+(*)	+(**)

TABLE 14
Statistical significance testing results achieved for static visual models that take still face images and spatio-temporal visual models that take face image sequences, where **+** / **-** denotes that there is / there is no statistically significant difference between a pair of systems (The significance level of $* P < 0.05$, $** P < 0.01$, $*** P < 0.001$). **APR** represents the results achieved for apparent personality recognition models while **SPR** represents the results achieved for self-reported personality recognition models.

	Open	Consc	Extrav	Agree	Neuro	Mean
APR	+(*)	+(*)	+(*)	+(*)	+(*)	+(*)
SPR	–	–	–	–	–	–

TABLE 15
Statistical significance testing results achieved for models using or without using spectral representation encoding strategy (i.e., short segment-level modelling VS. clip-level modelling), where **+** / **-** denotes that there is / there is no statistically significant difference between a pair of systems (The significance level of $* P < 0.05$, $** P < 0.01$, $*** P < 0.001$). **APR** represents the results achieved for apparent personality recognition models while **SPR** represents the results achieved for self-reported personality recognition models.

1.4 Poor performances in self-reported personality recognition

Correlation between each AU and personality traits: The detailed figures illustrating the correlation between each AU and each personality trait are provided in Fig. 3.

Results achieved for modified ChaLearn Impression and UDIVA datasets: Table 23 compares the results achieved on original ChaLearn First Impression and UDIVA datasets with the results achieved on the modified datasets (i.e., each dataset contains 232 training clips of 99 candidates and 36 validation clips of 20 candidates).

1.5 Data hungry experiments

Since personality is a complex construct, it is an open question of how much audio-visual material and what are the best recording scenarios required for its recognition. Table21 and Table22 provide the 'data hungry' experiments. The results show that the performances of APR keep improving with the increasing number of training samples, suggesting that 4000 (even 6000) training clips may still not be enough for APR. Meanwhile, the performances of SPR are fluctuated, which may be caused by the fact that these models are unreliable for inferring self-reported personality traits. In addition, we also found that if we select different numbers of frames from each video for developing models, the performances of benchmarked personality recognition models highly depend on the selected frames, i.e., the same number of frames selected from different video segments lead to large performance differences. However, due to the limited number of training samples for both datasets as well as the unreliable SPR models, it is still an open question what is the minimum amount of data that is needed for developing APR and SPR models.

1.6 Visualisation of model predictions

We visualise added scatter the predictions of the best audio, visual and audio-visual models vs. ground truths for both APR and SPR tasks in Fig. 4 to Fig. 9.

1.7 Other analysis

This section displays additional results to provide insightful understanding of the benchmarked models in apparent and self-reported personality recognition.

Identifiable information (subjects' identities) or behaviours/expressions: Fig. 10 illustrates whether the benchmarked personality recognition models focused on identifiable information (subjects' identities) or behaviours/expressions.

Correlation between personality traits: Fig. 11 illustrates the substantial dependencies and correlations among ground-truth apparent personality traits. Similarly, the predicted apparent personality traits also exhibit high interdependence, where individually predicting each trait did not eliminate correlations and dependencies among apparent personality traits. In contrast, there only exists very low correlations between ground-truth self-reported personality traits. However, in comparison to the systems that jointly predict five traits, individually predicting each trait even largely increased the dependency between self-reported personality traits predictions.

2 MODEL LOW-LEVEL DETAILS

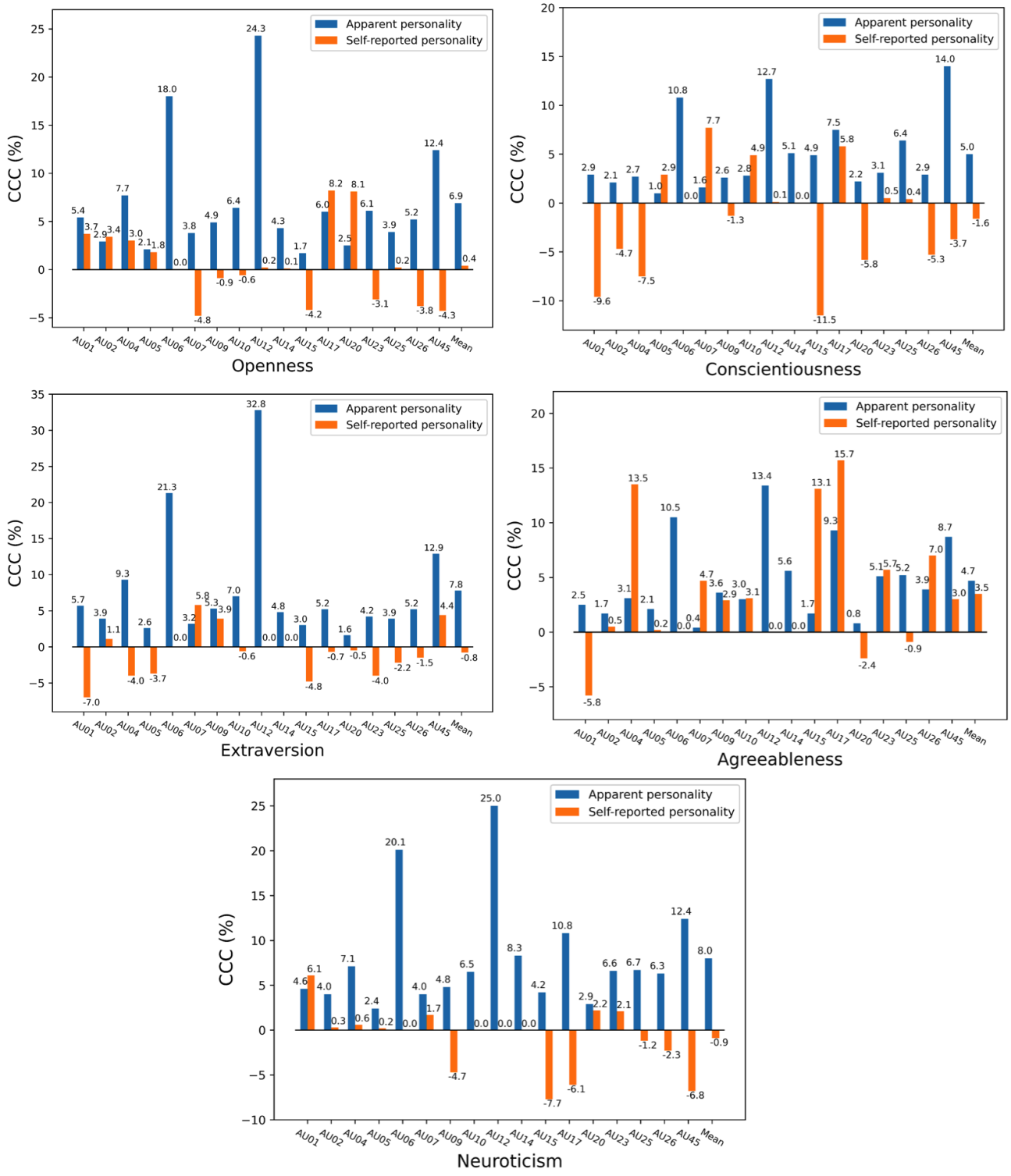


Fig. 3. The CCC between each AU and each personality trait

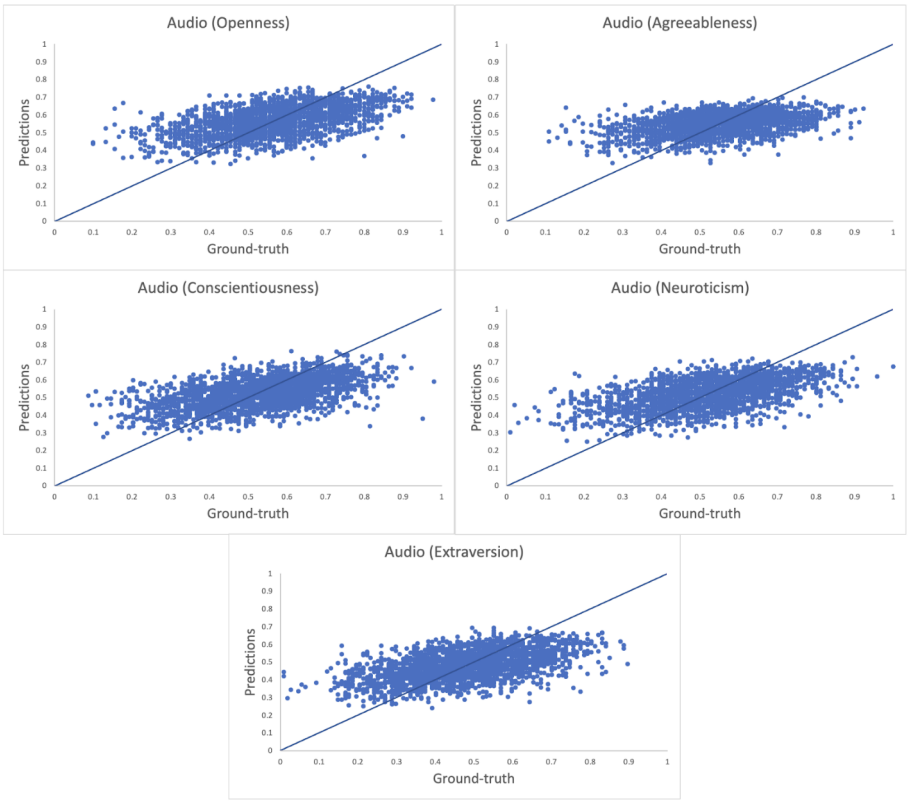


Fig. 4. Visualisation of the predictions made by the VGGish model on the ChaLearn First Impression dataset.

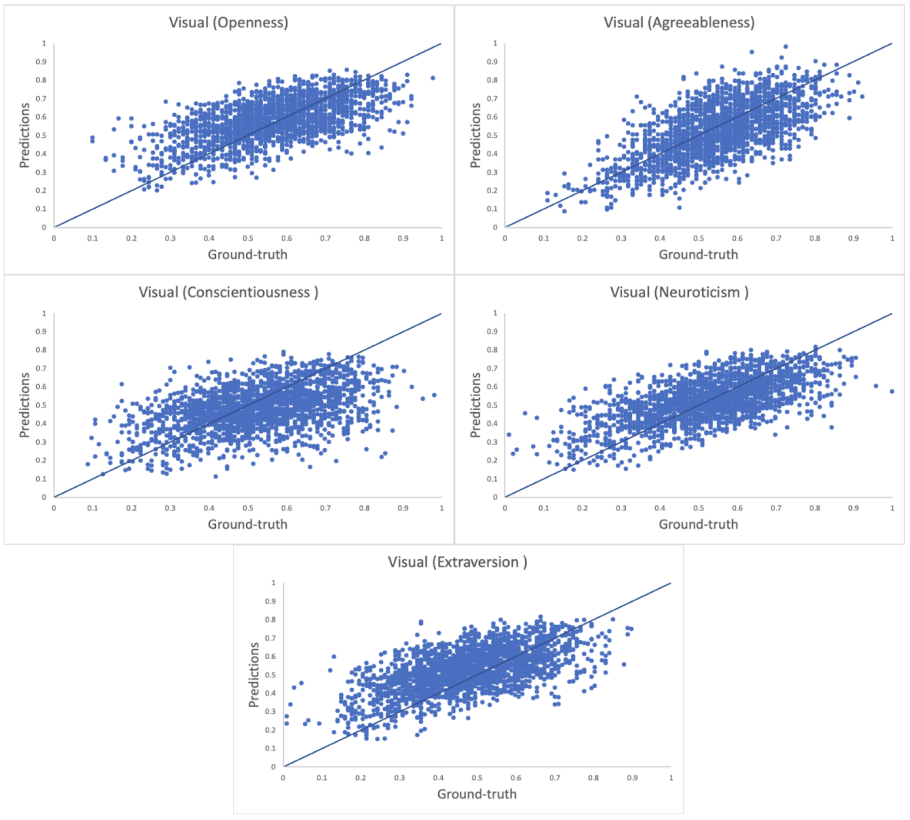


Fig. 5. Visualisation of the predictions made by the VAT model on the ChaLearn First Impression dataset.

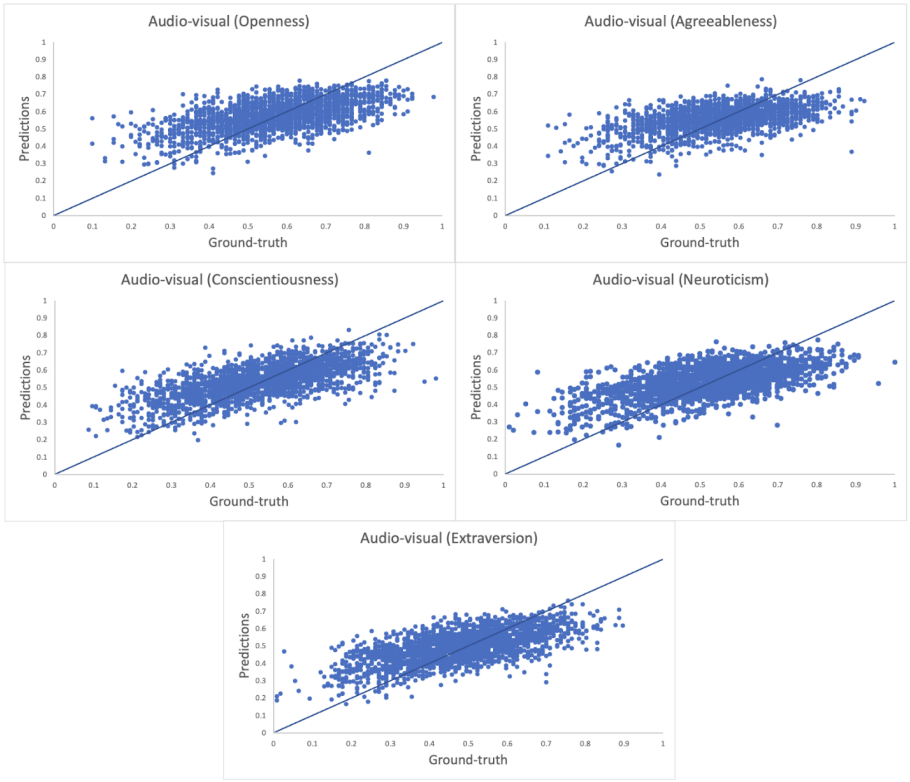


Fig. 6. Visualisation of the predictions made by the Amb-Fac-VGGish model on the ChaLearn First Impression dataset.

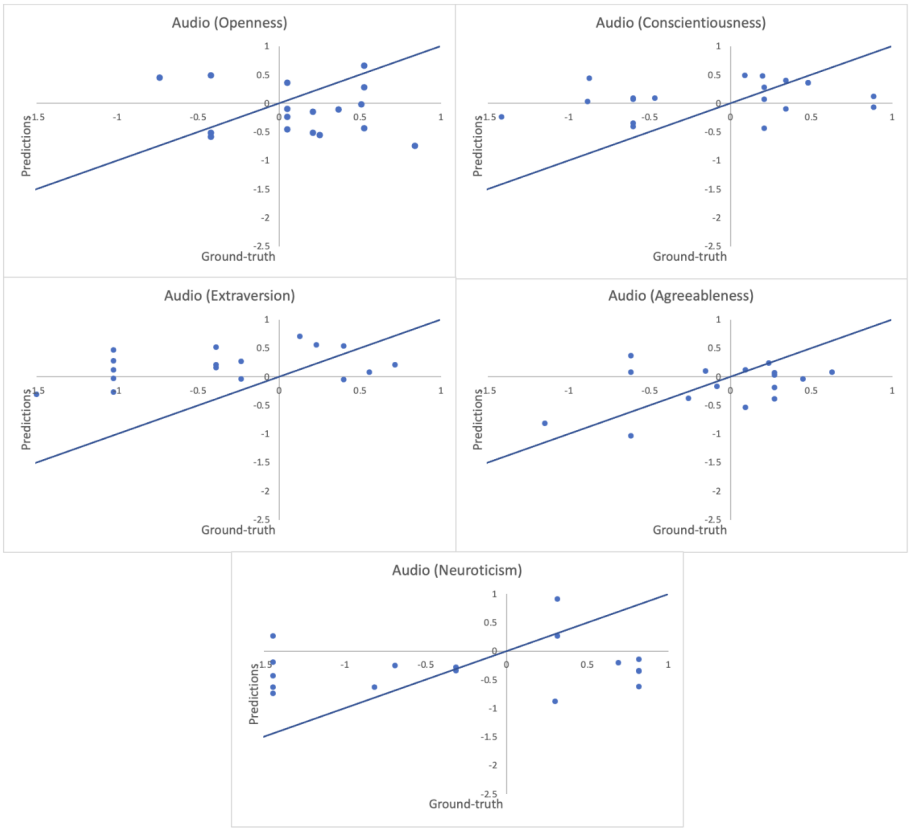


Fig. 7. Visualisation of the predictions made by the VGGish model on the UDIVA dataset.

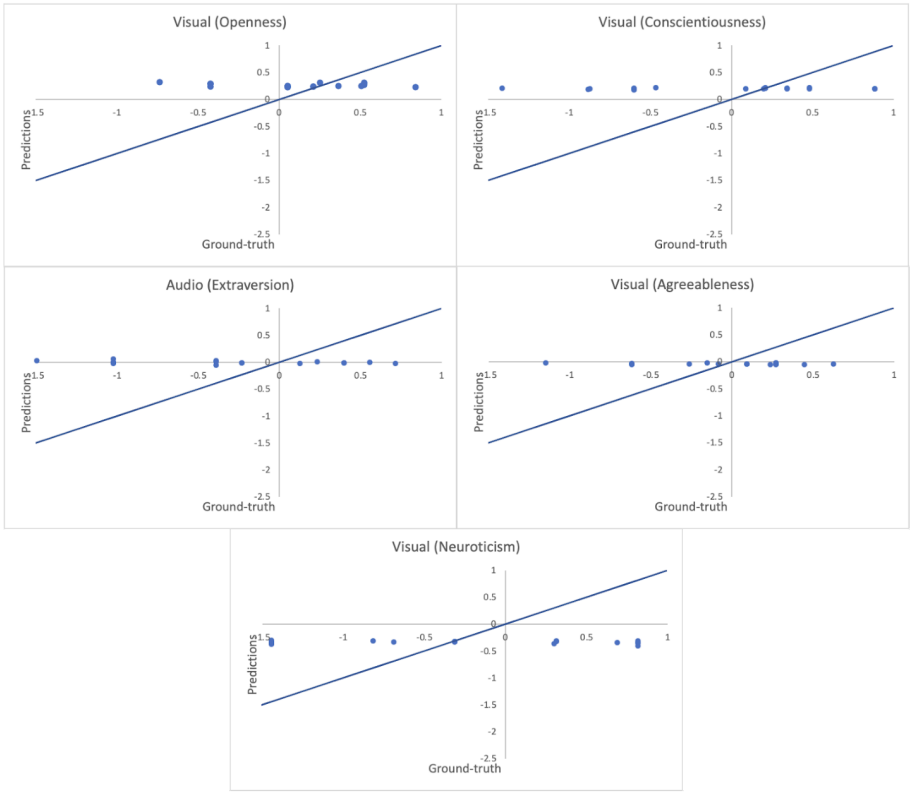


Fig. 8. Visualisation of the predictions made by the VAT model on the UDIVA dataset.

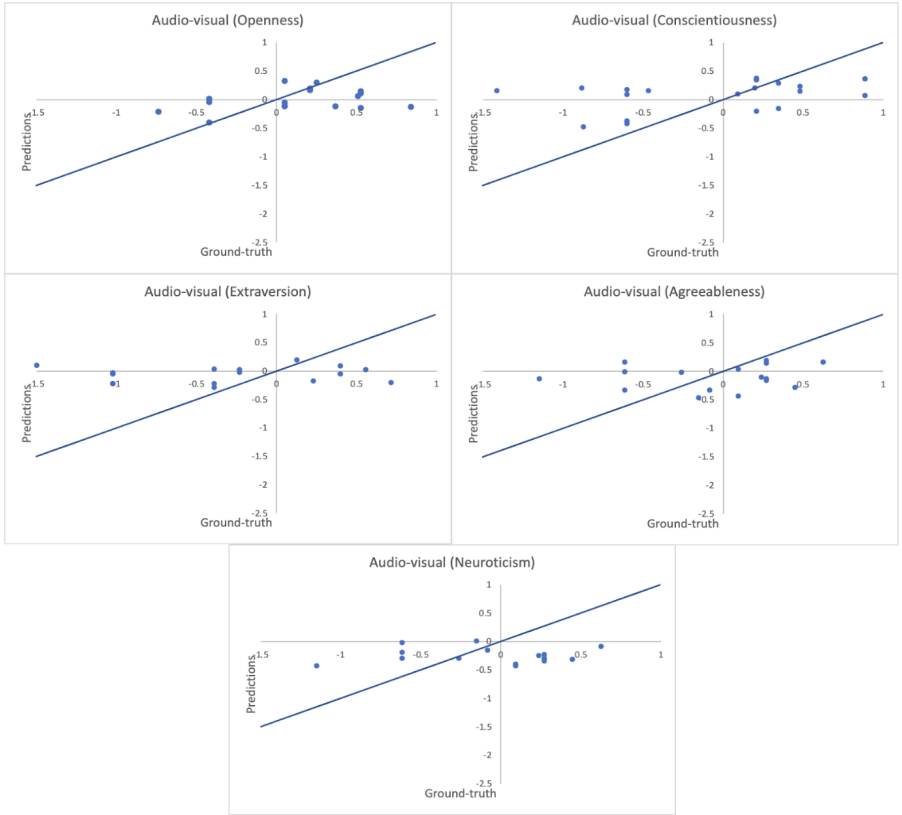


Fig. 9. Visualisation of the predictions made by the Amb-Fac-VGGish model on the UDIVA dataset.

Methods	Open	Consc	Extrav	Agree	Neuro	Mean
AFP	$-$	$+(**)$	$+(**)$	$-$	$+(**)$	$+(**)$
LRP	$-$	$+(**)$	$-$	$+(**)$	$-$	$-$
MLP	$+(**)$	$+(**)$	$+(*)$	$+(*)$	$-$	$+(**)$

TABLE 16
Statistical significance testing results achieved between the spectral representation of frame/segment-level predictions (SFP) and other different clip-level aggregation strategies for apparent personality recognition on the ChaLearn First Impression dataset, where **+** / **-** denotes that there is / there is no statistically significant difference between a pair of systems (The significance level of $* P < 0.05$, $** P < 0.01$, $*** P < 0.001$).

SFP	Open	Consc	Extrav	Agree	Neuro	Mean
AFP	$+(*)$	$-$	$-$	$-$	$-$	$-$
LRP	$-$	$-$	$-$	$-$	$-$	$-$
MLP	$-$	$-$	$-$	$-$	$-$	$-$

TABLE 17
Statistical significance testing results achieved between the spectral representation of frame/segment-level predictions (SFP) and other different clip-level aggregation strategies for self-reported personality recognition on the UDIVA dataset, where **+** / **-** denotes that there is / there is no statistically significant difference between a pair of systems (The significance level of $* P < 0.05$, $** P < 0.01$, $*** P < 0.001$).

	Open	Consc	Extrav	Agree	Neuro	Mean
APR	$+(*)$	$+(**)$	$+(*)$	$+(**)$	$+(*)$	$+(*)$
SPR	$-$	$-$	$-$	$-$	$-$	$-$

TABLE 18
Statistical significance testing results achieved by models that jointly predict five traits and models that individually predict each trait, where **+** / **-** denotes that there is / there is no statistically significant difference between a pair of systems (The significance level of $* P < 0.05$, $** P < 0.01$, $*** P < 0.001$).

	Open	Consc	Extrav	Agree	Neuro	Mean
SPR	$+(**)$	$-$	$-$	$-$	$-$	$-$

TABLE 19
Statistical significance testing results achieved by self-reported personality recognition models that additionally use metadata as the input and models that do not use metadata, where **+** / **-** denotes that there is / there is no statistically significant difference between a pair of systems (The significance level of $* P < 0.05$, $** P < 0.01$, $*** P < 0.001$).

All	Open	Consc	Extrav	Agree	Neuro	Mean
Animal	$-$	$-$	$-$	$-$	$-$	$-$
Lego	$-$	$-$	$-$	$-$	$-$	$-$
Ghost	$-$	$-$	$-$	$-$	$-$	$+(**)$
Talk	$+(*)$	$-$	$-$	$-$	$-$	$-$

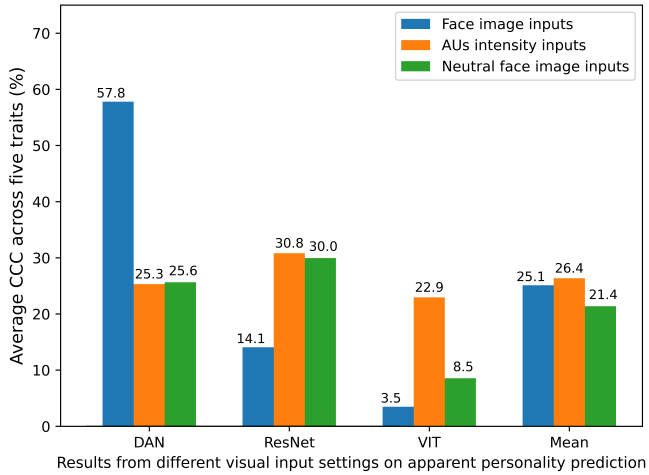
TABLE 20
Statistical significance testing results achieved by self-reported personality recognition models that conducted on the combined task and models that conducted on each individual task based on the UDIVA dataset, where **+** / **-** denotes that there is / there is no statistically significant difference between a pair of systems (The significance level of $* P < 0.05$, $** P < 0.01$, $*** P < 0.001$).

	Clips	Open	Consc	Extrav	Agree	Neuro	Avg.
Audio	58	-0.0873	0.0825	0.0656	0.1147	0.0300	0.04110
	116	-0.0339	0.1420	0.0415	0.1230	0.0642	0.06736
	174	0.0054	0.1426	0.1573	0.2213	0.0973	0.12478
	232	-0.0139	-0.0016	0.0016	-0.0013	-0.0006	-0.00316
Visual	58	-0.0082	0.0059	-0.0008	-0.0087	0.0079	-0.00078
	116	0.0082	0.0061	0.0017	-0.0052	0.0069	0.00354
	174	-0.0305	0.0007	-0.0140	0.0003	0.0068	-0.00734
	232	-0.0348	0.0468	0.0302	0.0397	0.0041	0.01720
Aud-vis	58	0.0139	0.0103	0.0103	-0.0076	0.0418	0.01374
	116	0.0056	-0.0123	-0.0123	0.0097	0.0013	-0.00160
	174	0.0086	-0.0063	-0.0063	0.0087	-0.0062	-0.00030
	232	0.0688	0.1882	0.1310	0.1069	0.0412	0.10722

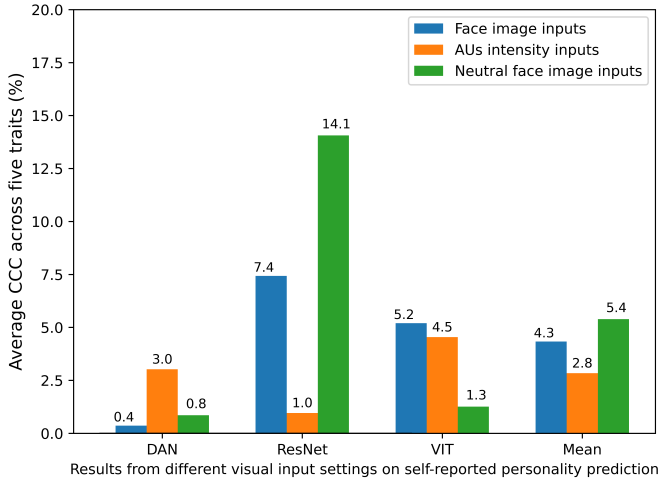
TABLE 21
Self-reported personality recognition results achieved on UDIVA with different number of training clips

	Clips	Open	Consc	Extrav	Agree	Neuro	Avg.
Audio	232	0.0974	0.1252	0.2902	0.1360	0.1003	0.14982
	1000	0.2391	0.1936	0.2550	0.1280	0.1845	0.20004
	2000	0.3610	0.2990	0.3632	0.1983	0.3380	0.31190
	3000	0.3784	0.3564	0.3780	0.2214	0.3789	0.34262
	4000	0.4071	0.3747	0.4020	0.2408	0.3938	0.36368
	6000	0.4516	0.4493	0.4429	0.3127	0.4500	0.42130
Visual	232	0.1754	0.2186	0.2524	0.2187	0.2317	0.21936
	1000	0.1812	0.2512	0.2145	0.1469	0.2228	0.20332
	2000	0.3142	0.3479	0.3778	0.2435	0.3387	0.32442
	3000	0.3778	0.4097	0.4526	0.2999	0.3948	0.38696
	4000	0.4207	0.4812	0.4869	0.3364	0.4419	0.43342
	6000	0.6216	0.6753	0.6836	0.5228	0.6456	0.62978
Aud-vis	232	0.0768	0.1397	0.2999	0.0624	0.0015	0.11606
	1000	0.2492	0.2829	0.2891	0.1681	0.2155	0.24096
	2000	0.3842	0.4779	0.4399	0.2950	0.3691	0.39322
	3000	0.4010	0.5152	0.4615	0.3176	0.4052	0.42010
	4000	0.4083	0.5189	0.4696	0.3195	0.4103	0.42532
	6000	0.5618	0.6421	0.5921	0.4620	0.5734	0.56628

TABLE 22
Apparent personality recognition results achieved on the ChaLearn First Impression dataset with different number of training clips

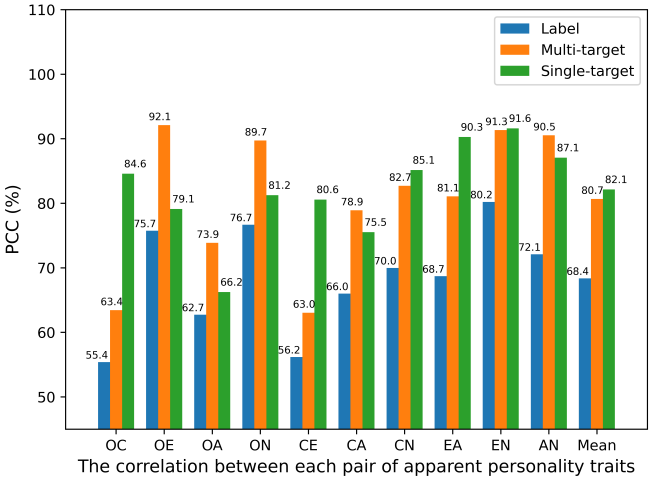


(a) CCC results achieved for the ChaLearn First Impression dataset.

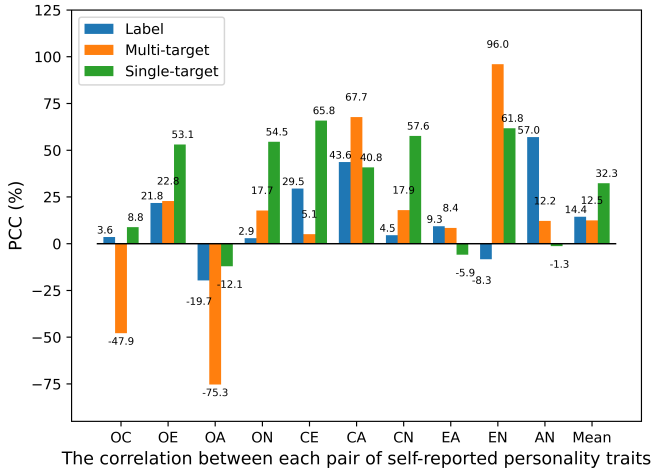


(b) CCC results achieved for the UDIVA dataset.

Fig. 10. Comparison between the average CCC results achieved by different visual inputs.



(a) Results achieved for the ChaLearn First Impression dataset.



(b) Results achieved for the UDIVA dataset.

Fig. 11. Correlation between each pair of personality traits, where the description of each bar can be interpreted as the first character for a pair of traits (e.g., OA denotes the correlation between Openness and Agreeableness).

	Traits	Open	Consc	Extrav	Agree	Neuro	Avg.
Audio	CRNet (C)	0.4122	0.3406	0.3846	0.2857	0.4306	0.3707
	CRNet (C)	-0.1174	-0.2467	-0.1273	-0.2423	-0.124	-0.1715
	CRNet (U)	0.0005	0.0290	0.0201	0.0335	0.0267	0.0220
	CRNet (U)	0.0016	0.0098	0.0076	0.015	-0.0221	0.0024
	VGGish (C)	0.4516	0.4493	0.4429	0.3127	0.4500	0.4213
	VGGish (C)	0.2571	0.0066	0.1022	-0.0337	0.056	0.0776
	VGGish (U)	0.0688	0.1882	0.1310	0.1069	0.0412	0.1072
	VGGish (U)	0.0464	0.1618	0.1115	0.1312	0.0379	0.0978
Visual	HRNet (C)	0.5923	0.6912	0.6436	0.5195	0.6273	0.6148
	HRNet (C)	0.2121	-0.0152	0.2248	0.0947	0.1718	0.1376
	HRNet (U)	0.2175	0.2998	-0.0039	0.1680	0.1945	0.1752
	HRNet (U)	0.0769	0.036	-0.0111	0.011	0.2476	0.0721
	VAT(C)	0.6216	0.6753	0.6836	0.5228	0.6456	0.6298
	VAT (C)	-0.0004	0.0037	0.0093	0.0042	0.0026	0.0039
	VAT (U)	-0.0139	-0.0016	0.0016	-0.0013	-0.0006	-0.0031
	VAT (U)	-0.0584	0.0058	0.0101	0.0169	-0.0187	-0.0089
Aud-vis	CRNet(C)	0.5193	0.5106	0.5024	0.4026	0.5119	0.4894
	CRNet (C)	0.082	-0.0791	0.0471	-0.0762	-0.0197	-0.0092
	CRNet (U)	0.0998	0.1780	0.1158	0.2168	0.0449	0.1311
	CRNet (U)	0.0287	0.1224	0.0166	0.1898	0.0511	0.0817
	Amb-VGGish (C)	0.5618	0.6421	0.5921	0.4620	0.5734	0.5663
	Amb-VGGish(C)	0.2696	0.1955	0.2546	0.074	0.1274	0.1842
	Amb-VGGish (U)	-0.0348	0.0468	0.0302	0.0397	0.0041	0.0171
	Amb-VGGish(U)	-0.0239	0.0916	0.0394	-0.0039	-0.0085	0.0189

TABLE 23
The CCC results achieved for the apparent personality recognition on the ChaLearn First Impression and self-reported personality recognition UDIVA dataset. To make the two datasets having the consistent size, we select a subset of Chalearn 2016 that contains 232 training clips, 36 valid clips, and 22 test clips and a subset of UDIVA dataset where a segment of 15s long is extracted from each clip. Here, the models in the bold format denote they are trained/evaluated using the subset and those in plain format are models trained/evaluated using using the original dataset. The ”(C)” and ”(U)” indicates the results achieved on ChaLearn 2016 dataset (C) or UDIVA ”(U)” dataset.

Models	Modality	Pre-processing / Feature extraction	Model Description	Modality fusion	Post-processing
DAN	Static full visual frame	Downsampling each video into 100 frames by randomly selecting one frame from every 6 frames and resizing each frame to 224×224	DAN network consists of 13 convolution -ReLU blocks, and an additional block that is equipped with both average-pooling and max-pooling layers, is added between the last convolution layer and the final FC layer	Frame-wise decision -level fusion	The clip-level visual personality prediction is obtained by averaging all frame-level predictions from the selected static frames
	Audio	Resampling audio data at a frequency of 44100 Hz by FFMpeg, and concatenating all frame-level MFCC features as a 39767-D vector	1-layer MLP		
CAM-DAN$^+$	Static full frame Static Face frame	10 frames are selected from each video and a face region is cropped from each frame, which is resized to 224×224	Based on DAN, this model applies a max pooling and an average pooling as two parallel branches in addition to convo- -lution layers. This model is trained using full frames and fine-tuned based on face frames	-	The clip-level visual personality prediction is obtained by averaging all frame-level predictions from the selected static frames
CNN-LSTM	Static full visual frame	Splitting each video into 6 segments and randomly selecting one frame from each segment. Face regions are extracted from the selected frame, which are resized to 112×112	The model stacks three 2D convolutional layers and two fully connected layer.	Frame-wise feature-level fusion, which is then fed to a LSTM model	The clip-level personality prediction is obtained by by averaging all frame-level predictions.
	Audio	Splitting the audio signal into 6 partitions and extracting 68-D hand-crafted features using pyAudioAnalysis [6] for each part	1-layer MLP		
ResNet	Static full visual frame	Resampling each video to the resolution of 456×256 and 25 fps. Then, randomly cropping a 224×224 patch from each frame	17-layer ResNet (ResNet-17)	Frame-wise feature-level fusion	The clip-level personality prediction is obtained by by averaging all frame-level predictions.
	Audio	Resampling each audio signal at the frequency of 16000 Hz and then randomly cropping a 50176-D feature	17-layer ResNet (ResNet-17)		
CRNet	Full visual frame sequence	Resampling each video into 32 frames and resizing them to 112×112	34-layer ResNet (ResNet-34)	Combining features output from full frame, face frame and audio branches based on the CR-block	The clip-level personality prediction is obtained by feeding combined clip-level feature to an ETR
	Face frame sequence	Resampling a video to 32 frames and extracting face regions from them	34-layer ResNet (ResNet-34)		
	Audio	Resampling the audio data at the frequency of 16000 Hz and then converting it into fixed-length vectors using librosa	34-layer ResNet (ResNet-34)		
PersEmoN	Static face frame	Randomly selecting 10 frames from each video coupled with randomly selected 10 images from an emotion dataset, where face regions are extracted from them	SphereFace network [7] with 4 MLPs	Weighted sum of emotion and personality features generated from two MLPs	–
Amb-Fac	Static full visual frame	Evenly downsampling each video into 6 frames and resizing each frame to 224×224.	ResNet-101	Frame-wise feature-level fusion	The clip-level personality prediction is obtained by by averaging all frame-level predictions.
	static face frame	Evenly downsampling each video into 6 frames, and extract face regions from them	ResNet-101		
	Audio	Resampling the audio data at the frequency of 16000 Hz	VGGish		

TABLE 24
Low-level details of the reproduced existing apparent personality recognition models that have been conducted on the ChaLearn First Impression dataset.

2.1 Model training and inference time

This section provides the training and inference time of the most time-consuming visual models, including their convergence time (Table 25), average training time for each epoch (Table 26 and Table 27). Here, the employed machine is equipped with an A4000 GPU of 16G memory as well as a 32-core AMD EPYC 7551p CPU.

3 More details of the employed datasets

The mean, variance, and median values of each trait's annotations for the three subsets of the UDIVA dataset are provided in Table28 and Fig. 12, where it is clear that the annotation distributions of three subsets are different. In contrast, Table 29 and Fig. 13 demonstrate that personality trait annotations of the three subsets defined by the ChaLearn First Impression dataset are more consistent. For both datasets, the annotations of all traits are unbalanced, as for each trait, most participants are annotated as the 'neutral' status (values close to 0 for the UDIVA and 0.5 for the ChaLearn First Impression dataset). Also, we provide the age and gender distribution of the UDIVA participants in Fig.14 and Fig 15, where the dataset is relatively balanced with respect to gender, i.e., the training dataset has 56% female versus 44% male, the validation dataset has 45% female versus 55% male, and the test dataset has 47% female versus 53% male. Meanwhile, the age distributions are not consistent for the three subsets. The training subset has the widest age distribution with some participants in their 70s, while all participants of the other two subsets are younger than 60. Furthermore, the test dataset has the most even and compact distribution. On the other hand, the training dataset has the most uneven age distribution. This age distribution gap may also challenge the personality computing model development. For the ChaLearn First Impression dataset, we are not able to obtain their meta data. Please refer to the original dataset paper for details.

4 Examples of the excluded personality computing approaches

Table 30 lists a set of example automatic personality recognition approaches that did not included in our benchmark, where the reason are also provided.

References

[1] Y. Li, J. Wan, Q. Miao, S. Escalera, H. Fang, H. Chen, X. Qi, and G. Guo, "Cr-net: A deep classification-regression network for multimodal apparent personality analysis." *International Journal of Computer Vision*, vol. 128, no. 12, 2020.

[2] C. Suman, S. Saha, A. Gupta, S. K. Pandey, and P. Bhattacharyya, "A multi-modal personality prediction system," *Knowledge-Based Systems*, vol. 236, p. 107715, 2022.

[3] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," 2020.

[4] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," 2019.

[5] S. Song, S. Jaiswal, L. Shen, and M. Valstar, "Spectral representation of behaviour primitives for depression analysis," *IEEE Transactions on Affective Computing*, 2020.

[6] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, 2015.

[7] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[8] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions-dataset and results," in *European conference on computer vision*. Springer, 2016, pp. 400–418.

[9] C. Palmero, J. Selva, S. Smeureanu, J. C. J. Junior, A. Clapés, A. Moseguí, Z. Zhang, D. Gallardo, G. Guilera, D. Leiva *et al.*, "Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset." in *WACV (Workshops)*, 2021, pp. 1–12.

[10] S. Song, Z. Shao, S. Jaiswal, L. Shen, M. Valstar, and H. Gunes, "Learning person-specific cognition from facial reactions for automatic personality recognition," *IEEE Transactions on Affective Computing*, 2022.

[11] S. Song, S. Jaiswal, E. Sanchez, G. Tzimiropoulos, L. Shen, and M. Valstar, "Self-supervised learning of person-specific facial dynamics for automatic personality recognition," *IEEE Transactions on Affective Computing*, 2021.

[12] Z. Shao, S. Song, S. Jaiswal, L. Shen, M. Valstar, and H. Gunes, "Personality recognition by modelling person-specific cognitive processes using graph representation," in *proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 357–366.

[13] K. Ilmini and T. Fernando, "Persons' personality traits recognition using machine learning algorithms and image processing techniques," *Advances in Computer Science: an International Journal*, vol. 5, no. 1, pp. 40–44, 2016.

[14] H. Salam, V. Manoranjan, J. Jiang, and O. Celiktutan, "Learning personalised models for automatic self-reported personality recognition," in *Understanding Social Behavior in Dyadic and Small Group Interactions*. PMLR, 2022, pp. 53–73.

[15] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74–79, 2017.

[16] N. Lemos, K. Shah, R. Rade, and D. Shah, "Personality prediction based on handwriting using machine learning," in *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 2018, pp. 110–113.

[17] V. Varshney, A. Varshney, T. Ahmad, and A. M. Khan, "Recognising personality traits using social media," in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 2017, pp. 2876–2881.

[18] F. Yang, X. Quan, Y. Yang, and J. Yu, "Multi-document transformer for personality detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 16, 2021, pp. 14 221–14 229.

[19] A. Kumar, R. Beniwal, and D. Jain, "Personality detection using kernel-based ensemble model for leveraging social psychology in online networks," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 5, pp. 1–20, 2023.

TABLE 25
The model training convergence time (h)

Method	DAN	CRNet	SENet	HRNet	Swin	VAT	Slow-Fast	TPN
ChaLearn [8]	1.8	4.2	2.2	4.3	5.3	24.2	21.0	19.0
UDIVA [9]	5.2	15.5	23.6	22.5	23.0	21.4	21.8	17.0

TABLE 26
The model training time (s) under the batch size of 2 on the UDIVA dataset.

Method	DAN	CRNet	SENet	HRNet	Swin	VAT	Slow-Fast	TPN
Each batch	0.2	1.9	0.4	0.8	0.9	0.8	1.2	1.6
Each Epoch	11600	110200	23200	46400	52200	368	556	742
Each Inference	0.2	1.9	0.4	0.8	0.9	0.8	1.2	1.6

TABLE 27
The model training time (s) with the batch size of 2 on the ChaLearn dataset

Method	DAN	CRNet	SENet	HRNet	Swin	VAT	Slow-Fast	TPN
Each video	0.2	1.9	0.35	1.2	0.8	0.75	1.1	1.5
Each Epoch	600	5700	1050	3600	2400	2250	3300	4500
Each Inference	0.2	1.9	0.35	1.2	0.8	0.75	1.1	1.5

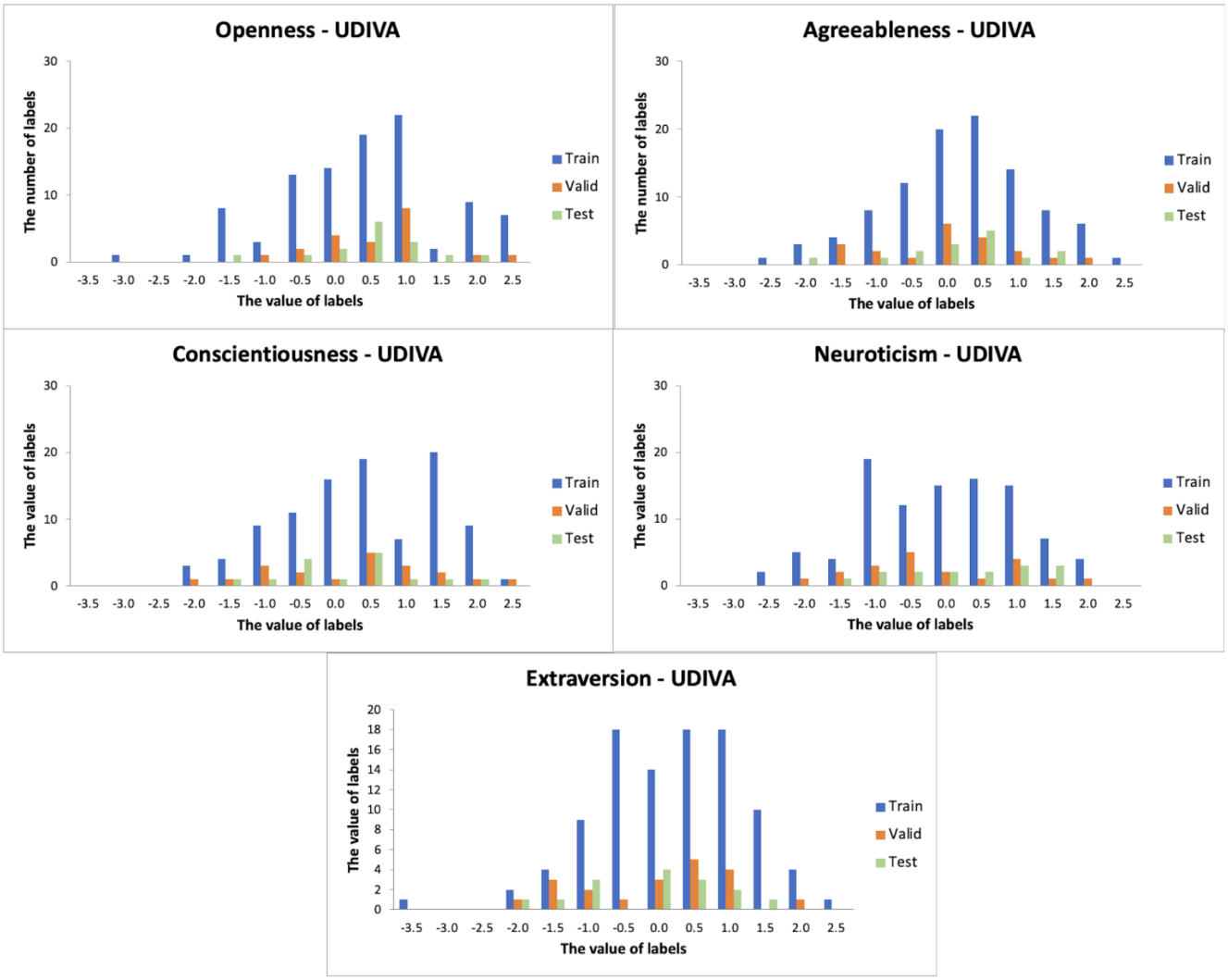


Fig. 12. Self-reported personality traits annotation distribution of the UDIVA dataset.

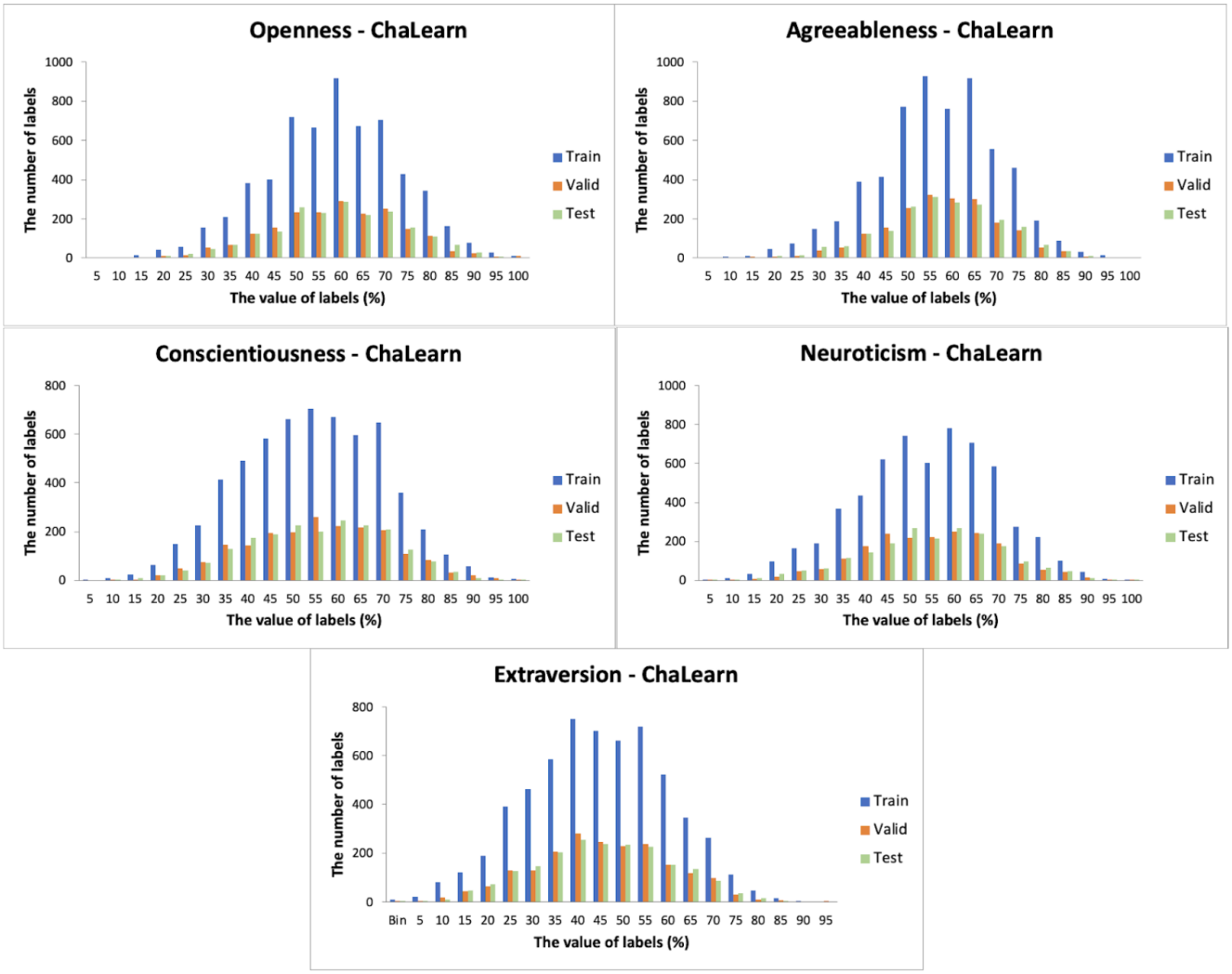


Fig. 13. Apparent personality traits annotation distribution of the ChaLearn First Impression dataset.

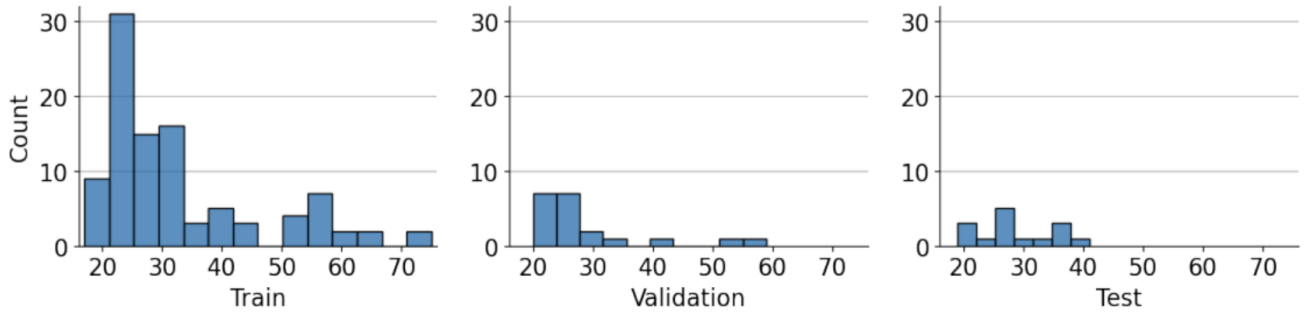


Fig. 14. Age distribution across train, validation and test splits of the UDIVA dataset.

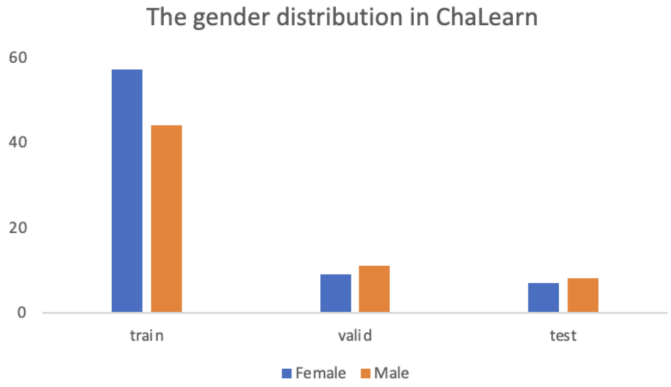


Fig. 15. Genda distribution across train, validation and test splits of the UDIVA dataset.

	Statistic	Open	Consc	Extrav	Agree	Neuro
Train	Mean	0.191	0.165	-0.052	-0.011	-0.306
	Variance	1.302	1.142	0.995	0.972	1.132
	Median	0.209	0.075	0.085	0.095	-0.312
Valid	Mean	0.359	-0.075	-0.287	-0.198	-0.388
	Variance	0.695	1.450	1.237	0.989	1.238
	Median	0.446	0.210	0.006	-0.171	-0.626
Test	Mean	0.160	-0.088	-0.360	-0.034	0.039
	Variance	0.776	0.872	0.960	0.721	0.721
	Median	0.209	0.091	-0.231	0.095	0.095

TABLE 28
Statistics of the self-reported personality traits annotations of the UDIVA dataset.

	Statistic	Open	Consc	Extrav	Agree	Neuro
Train	Mean	0.566	0.523	0.476	0.548	0.520
	Variance	0.022	0.024	0.023	0.019	0.024
	Median	0.578	0.524	0.477	0.560	0.531
Valid	Mean	0.566	0.528	0.477	0.551	0.522
	Variance	0.021	0.024	0.022	0.016	0.022
	Median	0.567	0.534	0.477	0.560	0.521
Test	Mean	0.568	0.525	0.478	0.552	0.522
	Variance	0.021	0.023	0.023	0.018	0.024
	Median	0.578	0.534	0.477	0.560	0.531

TABLE 29
Statistics of the apparent personality traits annotations of the ChaLearn First Impression dataset.

TABLE 30
Example personality computing approaches that were excluded in our benchmark.

Method	Excluding reasons
[10]:	1. the personality representation learning step and personality recognition step are separately conducted; 2. it hasn't been evaluated on the ChaLearn First Impression dataset.
[11]:	1. the personality representation learning step and personality recognition step are separately conducted.
[12]:	1. the personality representation learning step and personality recognition step are separately conducted; 2. it hasn't been evaluated on the ChaLearn First Impression dataset.
[13]:	1. the personality representation learning step (i.e., achieved by a hand-crafted method) and personality recognition step (i.e., SVM is employed) are separately conducted; 2. personality recognition is treated as a multi-class classification task rather than a regression task.
[14]	1. it is developed for the self-reported personality recognition task.
[15]:	1. the text-based personality recognition approach; 2. the personality recognition step is treated as a binary classification task rather than a regression task.
[16]:	1. the handwriting-based personality recognition approach; 2. the pipeline is not end-to-end, where a hand-crafted signal processing algorithm is combined with a CNN for personality representation learning.
[17]:	1. social media data such as the number of likes, groups, tags and events are used for personality recognition; 2. personality recognition is treated as a multi-class classification task rather than a regression task.
[18]:	1. social media posts are used for training data; 2. personality recognition is treated as a multi-class classification task rather than a regression task.
[19]	1. personality detection from natural language; 2. multiple machine-learning methods, such as Support Vector Machines with voting techniques and ensemble methods, are proposed to form a personality detection pipeline.