Feature-wise Optimization and Performance-weighted Multimodal Fusion for Social Perception Recognition

ABSTRACT

Automatic social perception recognition is a new task to mimic the measurement of human traits, which was previously done by humans via questionnaires. We evaluated unimodal and multimodal systems to predict agentive and communal traits from the LMU-ELP dataset. We optimized variants of recurrent neural networks from each feature from audio and video data and then fused them to predict the traits. Results on the development set show a consistent trend that multimodal fusion outperforms unimodal systems. The performance-weighted fusion also consistently outperforms mean and maximum fusions. We found two important factors that influence the performance of performance-weighted fusion. These factors are normalization and a number of models.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**; **Multimedia and multimodal retrieval**; • **Human-centered computing → Empirical studies in HCI**; • **Social and professional topics → User characteristics**.

KEYWORDS

multimodal fusion, sentiment analysis, social perception, parameter optimization

ACM Reference Format:
. 2018. Feature-wise Optimization and Performance-weighted Multimodal Fusion for Social Perception Recognition. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (MM'24)Proceedings of the 32nd ACM International Conference on Multimedia (MM'24), October 28-November 1, 2024, Melbourne, Australia.* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Agentive and communality traits are important predictors of leadership potential [21]. The predictability is more important from the perspective of female leadership aspirants [22]. Although it is debatable which one is more important than the other, men mainly prefer competent (agentive) to communal, while for women leaders, communal leadership has a more substantial impact [10]. Communal, regardless of gender, is primary among the fundamental dimensions of content in social cognition [1].

Instead of measuring agentive and communal values manually via question-answer paper surveys (as in [20]), both traits can be measured automatically by capturing and analyzing multimodal data, such as voice, gesture, face expression, and verbal contents. This automatic measurement is more convenient, reducing time and the burden of manual labor. However, it is essential to perform precise and accurate measurements of both traits to obtain reliable results.

Until recently, there has been no dataset to analyze the social perception of agentive and communal traits for leadership aspirants. Reference [3] provided the LMU-ELP dataset in a challenge to promote research and development of automatic agentive and communal traits measurement by providing multimedia data, including audio and video data with their corresponding features. This social perception dataset is a great start to analyzing the social perception of agentive and communal traits despite its small number of samples.

The fusion of several models, particularly from different modalities, with different performances is important to improve the performance of the fusion model. This ensemble learning method has proven to be effective in speech emotion recognition [5, 7, 17], acoustic-linguistic emotion recognition [5], cough screening [16], and multimodal forgery detection [14]. On the other side, adding ensemble learning doubles the computation cost and the time to train the model. Hence, implementing ensemble learning without significant performance improvements over single models could be a waste of time and resources.

This study contributes to the research of automatic agentive and communal trait measurement by optimizing unimodal data and multimodal data. First, we optimize the deep learning model for both agentive and communal traits based on the specific feature. Second, we evaluate performance-weighted fusion to combine several audio and video models with different performances. These performances are incorporated into a weighted fusion model. We found two important factors that affect the performance of the performance-weighted fusion model: (1) the normalization of the prediction scores before weighting and (2) the number of models used in the fusion.

2 DATASET

The LMU-ELP dataset [3] was evaluated on unimodal and multimodal fusion. The LMU-ELP dataset consists of audio-visual recordings of US executives, specifically the highest-ranking executives of listed firms - chief executive officers (CEOs) - presenting their firms to potential investors before taking them public. The dataset consists of 8 agentive and eight communal traits. Each CEO's agency and communality has been rated on a 16-dimensional Likert scale ranging from 1 to 7. For the purpose of this challenge, these scores have been normalized to a range of [0,1]. The dataset was split into 33% training, 33% for development, and 33% for test set. The total duration of the dataset is about 1 hour and 30 minutes. Pearson Correlation Coefficient (PCC) is the evaluation metric used to measure the similarity between the predicted and ground truth labels.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM'24, October 28 - November 1, 2024, Melbourne, Australia.

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

3 METHODS

3.1 Audio and Video Features

Six features are evaluated: three audio features and three video features. The audio features include normalized egemaps [11, 12], deepspectrum (ds) [4], and wav2vec2 robust version trained on affective dataset (w2v-msp) [23]. The w2v-msp, which is finetuned from the MSP-PODCAST dataset [15], is known to be informative for speech emotion recognition [8]. The video features include facenet512 [18], Facial activation unit (FAU), and VitFER [9]. Details of audio and video feature extractions can be referred to the previous study [3].

3.2 Baseline and Non-Optimized Models

For the **baseline** model, we utilized a PyTorch binary model (in PTH format), in which the URL for checkpoints is available at MuSe-2024 repository [1]. We have no information about the architecture and training parameters to build the model. We evaluated that model on the development set of the LMU-ELP dataset.

For the non-optimized (**non-optim**) model, we evaluate a GRU model with two RNN layers. Each layer has 256 nodes. The model was trained with a learning rate of 0.0005, 5 patiences for early stopping, and linear dropout of 0.4 through five different seeds. The best model from one of five seeds was selected as the final model for a feature.

As a variant of non-optimized models, we train the same models but with a learning rate scheduler. We reduced the learning rate on factor 0.5 using the reduce on plateau method. We called this model the **non-optim-2** model.

3.3 Optimized Models

For the optimized (**optim**) model, we used the Optuna toolkit [2] to optimize models for specific features. We optimized fifteen parameters to optimize. These parameters are model dimension (model_dim), number of layers (n_layers), whether it uses bidirectional (rnn_bi), number of nodes in fully connected layer (d_fc_out), linear dropout rate (linear_dropout), batch size (bs), type of loss function (choices: mse, mae, ccc loss [6], pcc loss), regularization (reg.), rnn type (choices: RNN (Elman), GRU, and LSTM), number of epochs, learning rate (lr), and the whether it uses residual connection if it uses more than a single RNN layer. The consideration for choosing parameters is based on empirical results and previous studies (e.g., number of patience for early stopping [19], and choice of loss functions [6]).

As a variant of the optimized model, we train the same model but with an additional data augmentation. We augmented the extracted features with noise injection, time warping, magnitude warping, mixup, and cutmix. We called it **optim-2** model.

4 PERFORMANCE-WEIGHTED FUSION

Suppose we have a different model from each feature. Each model has a prediction score. We can calculate the weighted average of the prediction scores from each model. The weights are calculated based on the performance of each model. Let $P \in \mathbb{R}^{n \times m}$ be the prediction array, where n is the number of samples and m is the number of prediction models.

Weight computation for each model i (where $i = 1, \ldots, m$) is conducted as follows. First, we calculate the performance (as weights) of each model using PCC,

$$PCC_i = PCC(P_i, labels). \tag{1}$$

We discarded negative weights and only used positive weights (models with positive PCC). For all non-zeros weights, we normalized them to sum up to 1.

$$w = \frac{PCC_i}{\sum_{i=1}^{m} PCC_i} \tag{2}$$

We then normalized each score in predictions ($P_{i,j}$) before multiplying them with weights,

$$P_i = \frac{(P_{i,j} - \min P_i)}{(\max P_i - \min P_i)}. \tag{3}$$

The weighted prediction for each model can be calculated,

$$P_{weighted} = P_{i,j} \odot \mathbf{w}, \tag{4}$$

where $\odot$ denotes element-wise multiplication. Finally, we sum up the weighted predictions from all models as the fused prediction (P_f),

$$P_f = \sum_{j=1}^{m} P_{weighted,k}, \tag{5}$$

where $P_{weighted,k}$ is the k-th column of $P_{weighted}$.

The research methods above to build the models are available in the open repository [1].

5 RESULTS AND DISCUSSION

We explain the results and discuss the findings in the following forms: feature-wise optimization, unimodal and multimodal fusion, effect of normalization, effect of the number of the models and test results.

6 FEATURE-WISE OPTIMIZATION

We evaluated 15 parameters and found ten important parameters as listed in Table 1. Among the three features, vanilla RNN is the classifier that gained top performance in feature-wiser optimization. We also found that our effort to add bidirectional RNN, as well as adding residual connection between RNN layers, did not improve the performance of social perception recognition.

Figure 1 shows an example of a hyperparameter optimization result for the VitFER feature. In this VitFER case, the most important parameter is the number of layers, followed by learning rate and regularization. Not all features have the same order of parameter importance; Table 1 shows the common important parameters for all features.

[1] https://github.com/amirip/MuSe-2024

[1] https://github.com/XXX/MuSe-2024

Table 1: Optimal training parameters for the unimodal models.

Feature	rnn_type	n_layers	model_dim	d_fc_out	rnn_dropout	lr	loss	reg.	patience	# epochs
ds	rnn	4	72	60	0.48	0.00066	ccc	1.36E-05	22	224
egemaps	gru	2	108	54	0.49	0.00042	mse	0.00028	23	234
wav2vec	rnn	3	74	96	0.40	0.00069	mse	7.29E-05	21	176
fau	rnn	2	123	57	0.16	0.00022	mse	0.00080	16	585
facenet	lstm	2	121	51	0.28	0.00029	ccc	9.99E-05	14	396
vitfer	gru	1	123	124	0.48	0.00073	ccc	9.98E-05	29	832

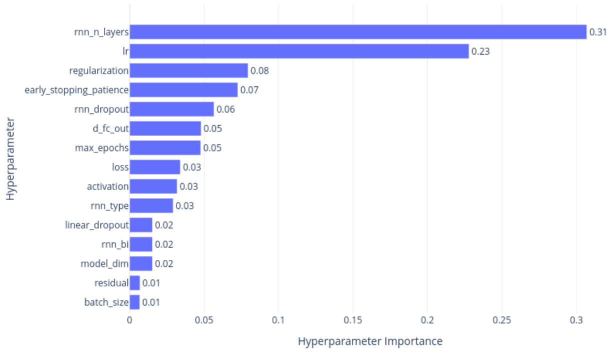

Figure 1: Hyperparameter importance for vit-fer

6.1 Unimodal and multimodal fusion results

Tables 2 - 6 show results of unimodal modal and multimodal fusion in PCC scores for each social perception attribute for the development set. The order of performance based on the highest average scores across attributes is the same for unimodal and multimodal fusion. From the highest to the lowest scores is `optim`, `optim-2`, `non-optim`, `non-optim-2`, and `baseline`.

In the unimodal evaluation, the vit-fer model consistently performs the best across all traits for optim model, with the highest average score of 0.44276. The w2v-msp model performs well for traits like independent, risk-taking, and confident, suggesting it might be particularly good at capturing assertive or leadership-related characteristics. Facenet512 shows high performance for traits like dominant and aggressive, indicating it might be effective at detecting more forceful personality traits. The deepspectrum model performs relatively well for traits like enthusiastic and collaborative, suggesting it might be good at capturing positive social interactions. The egemaps model shows consistent performance across traits but does not stand out as the top performer for any particular trait. The fau model generally performs in the middle range across most traits. Some traits, like aggressive and arrogant, show higher predictability across models, while others, like kind and sincere, are generally harder to predict.

The data from unimodal results (Tables 2, 3, 4, 5, and 6, columns 2 – 6) suggests that different models have strengths in predicting different types of personality traits, which could be useful for creating ensemble models or choosing specific models for particular applications. The trait that appears to be best predicted using unimodal approaches is “arrogant". Specifically, the vit-fer model shows the highest unimodal performance for predicting arrogance, with a PCC score of 0.6385 in the optimized model (Table 5) and 0.5952 in the optimized model with augmentation (Table 6).

The “aggressive" trait also shows strong unimodal prediction in addition to “arrogant," specifically with the vit-fer model (0.5132 in the optimized model and 0.5022 in the optimized model with augmentation). It's worth noting that the vit-fer model consistently performs well across most traits, but its performance is particularly strong for the agentive personality than communal characteristics. Another trait that shows strong prediction from the unimodal feature is dominant with the facenet512 feature, which obtains a score above 0.4 in all five data.

The performance-weighted fusion (perf.) consistently shows the highest values among the mean and max (maximum) fusions in all five evaluations (Tables 2 – 6, right side). The max fusion generally has the lowest values, indicating that individual models or features might not perform as well on their own compared to combined or weighted approaches. The performance-weighted approach seems to be particularly effective for traits like arrogant, aggressive, and risk-taking, with values above 0.6 in the optim model (Table 5). In that model, the trait “kind" shows the lowest values across all three columns; it might be the most challenging characteristic to predict or measure for the optim model. Interestingly, certain traits like “arrogant, “risk-taking," and “good-natured" consistently achieve higher PCC scores across different fusion strategies (perf., mean, and max), indicating that these traits may be more reliably predicted using multimodal approaches. There is a noticeable gap between the performance-weighted and mean columns (second best) for most traits, indicating that the weighted approach provides significant improvements over simple averaging. Our hypothesis that the weighted approach might be more effective in capturing the complex interactions between different features has proved to be effective, backed by these results.

It can be justified that the recognition rate of agentive traits is higher than communal traits on both development sets (e.g., in optim-2 perf., the average score for agentive vs. communal is 0.5875 vs. 0.5367). This empirical result on computation-based models is not the same as the results from the human-based models [24]. A recent study in psychology also found that self-ratings of agency are better predictors of self-esteem than self-ratings of communion [13].

Although we tried to improve the PCC scores of non-optim and optim models by adding a learning rate scheduler and augmentations, the development scores were not improved (but improved in the test set). We suspect that these phenomena are due to the fact that the number of samples is very small (total data is 1 hour

Table 2: PCC scores from the baseline model

		Unimodal						Multimodal fusion		
	Target	ds	egemaps	w2v-msp	facenet512	fau	vit-fer	perf.	mean	max
agentive	aggressive	0.0960	0.0360	0.3079	0.3226	0.0735	0.3898	**0.4679**	0.3657	0.0821
	arrogant	0.0576	0.0359	0.3237	0.2964	-0.1646	0.3559	**0.4578**	0.3073	0.0605
	assertive	0.1013	0.0999	0.3061	0.2885	0.0805	0.3139	**0.4615**	0.3740	0.2456
	confident	0.0423	0.1111	0.2789	0.2637	0.0707	0.2490	**0.3513**	0.1629	0.0986
	dominant	0.1504	0.0855	0.2005	0.4442	0.2147	0.3052	**0.4964**	0.4416	0.3466
	independent	0.0953	0.1167	0.2743	0.2348	0.0756	0.0198	**0.3586**	0.2301	0.2142
	risk-taking	0.2130	0.1525	0.2456	0.4757	0.0200	0.3692	**0.5381**	0.5290	0.4581
	leader-like	0.0603	0.0985	0.1614	0.2716	0.1197	0.2218	**0.3624**	0.3178	0.1382
communal	collaborative	0.0725	0.1555	0.1257	0.2048	0.0256	-0.0409	**0.2634**	0.1794	0.0185
	enthusiastic	0.1581	0.0812	0.1148	0.2719	0.0457	0.2335	**0.3513**	0.2928	0.1059
	friendly	0.0835	0.1779	0.1138	0.2128	0.0441	-0.1029	**0.2829**	0.0703	-0.1851
	good-natured	0.0437	0.1486	0.1638	0.0379	0.1668	-0.0955	**0.2365**	0.0708	-0.0246
	kind	0.0647	0.1337	0.1187	0.0268	0.0371	-0.2277	**0.1751**	-0.0125	-0.2455
	likeable	0.0589	0.1717	0.1695	0.2363	0.0661	-0.0727	**0.1751**	0.1137	0.0117
	sincere	0.0612	0.1556	-0.0181	0.2426	0.0265	-0.2094	**0.3034**	0.1128	-0.1368
	warm	0.0580	0.1507	0.0471	0.0072	0.0488	-0.0931	**0.1499**	0.0432	-0.0711
	avg.	0.0886	0.1194	0.1834	0.2399	0.0594	0.1010	**0.3395**	0.2249	0.0698

Table 3: PCC scores from the non-optimized model on the development set.

		Unimodal						Multimodal fusion		
	Target	ds	egemaps	w2v-msp	facenet512	fau	vit-fer	perf.	mean	max
agentive	aggressive	0.2090	0.0171	0.1269	0.3108	0.1470	0.5409	**0.5528**	0.4063	0.4063
	arrogant	0.0933	0.1270	0.3552	0.3355	0.1453	0.6311	**0.6330**	0.5898	0.5898
	assertive	0.2815	0.0432	0.3114	0.2885	0.0805	0.3139	**0.5227**	0.3817	0.3817
	confident	0.1830	0.2505	0.2722	0.3062	0.1727	0.2891	**0.5788**	0.4792	0.4792
	dominant	0.0971	0.1333	0.2450	0.5062	0.0639	0.4512	**0.5729**	0.4696	0.4696
	independent	0.2643	0.2052	0.3606	0.3125	0.0896	0.2518	**0.5436**	0.4380	0.4380
	risk-taking	0.1830	0.2517	0.3882	0.4635	0.2465	0.4033	**0.6315**	0.5555	0.5555
	leader-like	0.1348	0.1390	0.2592	0.4076	0.1907	0.3847	**0.5837**	0.5540	0.5540
communal	collaborative	0.1171	0.3451	0.0426	0.1306	0.1266	0.2310	**0.4121**	0.3859	0.3859
	enthusiastic	0.2352	0.3046	0.2478	0.1902	0.3162	0.3206	**0.4967**	0.3994	0.3994
	friendly	0.1300	0.2334	0.2275	0.0465	0.3066	0.3740	**0.4998**	0.4152	0.4152
	good-natured	0.1511	0.3433	0.2676	0.1115	0.3254	0.4506	**0.5732**	0.4523	0.4523
	kind	0.1910	0.2402	0.1586	0.1284	0.3004	0.4290	**0.4802**	0.3584	0.3584
	likeable	0.1489	0.3185	0.2001	0.0418	0.3526	0.3968	**0.5614**	0.3891	0.3891
	sincere	0.2073	0.3580	0.1713	0.0515	0.2435	0.2086	**0.5051**	0.4934	0.4934
	warm	0.1980	0.1964	0.2338	-0.0343	0.3219	0.2091	**0.4651**	0.1506	0.1506
	avg.	0.1765	0.2192	0.2417	0.2248	0.2143	0.3679	**0.5383**	0.4324	0.4324

and 30 minutes), and the ratio of training, development, and test (33:33:33) is not ideal for machine/deep learning.

6.2 Effect of Normalization

The table 7 compares PCC scores with and without prediction normalization (before weighting, eq. 3). The values in parentheses represent the differences from scores with normalization, consistently showing negative values. This indicates that normalization generally improves prediction performance across all models and traits. This factor is reasonable since the weights itself are normalized (eq. 2). The only trait that does not follow this trend is likable with the only baseline model. Hence, these phenomena can be ignored. It can be concluded that normalization before applying weighting is an important factor in performance-weighted fusion.

6.3 Effect of The Number of Models

Table 8 demonstrates the impact of using different numbers of top-performing models in late fusion approaches for personality trait prediction. Notably, the use of top-4 models consistently outperforms top-2 models across all fusion strategies (baseline, non-optim, non-optim-2, optim, and optim-2), with average PCC scores improving by 0.0254 to 0.0576. However, all results from top-4 are still

Table 4: PCC scores from non-optimized with learning rate scheduler (non-optim-2) model on development set.

		Unimodal						Multimodal fusion		
	Target	ds	egemaps	w2v-msp	facenet512	fau	vit-fer	perf.	mean	max
agentive	aggressive	0.2090	-0.0084	0.1005	0.3108	0.1456	0.5409	**0.5519**	0.4044	0.3610
	arrogant	0.0933	0.1270	0.3426	0.3355	0.1453	0.6311	**0.6308**	0.5852	0.1091
	assertive	0.2815	0.0432	0.3114	0.2885	0.0805	0.3139	**0.5227**	0.3817	0.2095
	confident	0.1830	0.2505	0.2722	0.3062	0.1727	0.2803	**0.5634**	0.4526	0.2739
	dominant	0.0971	0.1058	0.2450	0.5098	0.0246	0.4512	**0.5790**	0.5099	0.4028
	independent	0.2220	0.2052	0.3606	0.3123	0.0896	0.0409	**0.4683**	0.3876	0.3253
	risk-taking	0.1830	0.2517	0.3882	0.4635	0.2465	0.3955	**0.6285**	0.5510	0.4058
	leader-like	0.1348	0.1390	0.2592	0.4076	0.1907	0.3687	**0.5782**	0.5458	0.3347
communal	collaborative	0.1171	0.3486	0.0457	0.1306	0.1266	0.2310	**0.4073**	0.3806	0.3771
	enthusiastic	0.2352	0.3046	0.2478	0.1902	0.3162	0.3111	**0.4956**	0.3945	0.2737
	friendly	0.1300	0.2334	0.2275	0.0465	0.3066	0.3342	**0.4990**	0.3693	0.1707
	good-natured	0.1511	0.3433	0.2676	0.1050	0.3254	0.3087	**0.5635**	0.4021	0.2089
	kind	0.1910	0.2402	0.1586	0.0915	0.3004	0.3305	**0.4569**	0.2879	0.1604
	likeable	0.1489	0.3185	0.1658	0.0418	0.3526	0.2857	**0.5547**	0.3476	0.0337
	sincere	0.2073	0.3580	0.1839	0.0515	0.2435	0.2086	**0.5078**	0.4996	0.2073
	warm	0.1980	0.1964	0.2338	-0.0343	0.3219	0.2091	**0.4651**	0.1506	-0.0600
	avg.	0.1739	0.2161	0.2382	0.2223	0.2118	0.3276	**0.5295**	0.4157	0.2371

Table 5: PCC scores from optimized (optim) model on the development set.

		Unimodal						Multimodal fusion		
	Target	ds	egemaps	w2v-msp	facenet512	fau	vit-fer	perf.	mean	max
agentive	aggressive	0.2742	0.2181	0.1160	0.3116	0.2585	0.5132	**0.6175**	0.4357	0.3916
	arrogant	0.2122	0.2541	0.2364	0.3336	0.2639	0.6385	**0.6787**	0.4736	0.6205
	assertive	0.3201	0.1789	0.3528	0.3336	0.1292	0.4181	**0.5918**	0.3432	0.0462
	confident	0.2765	0.2156	0.3630	0.1238	0.2067	0.4865	**0.5902**	0.4043	0.2765
	dominant	-0.0253	0.2535	0.3281	0.5015	0.1291	0.4506	**0.6161**	0.5294	0.5091
	independent	0.2297	0.1065	0.4117	0.2806	0.1876	0.4153	**0.5843**	0.4048	0.2476
	risk-taking	0.1615	0.2796	0.4156	0.3898	0.3473	0.4590	**0.6293**	0.4630	0.4099
	leader-like	0.1853	0.2049	0.3346	0.3584	0.2285	0.4838	**0.6056**	0.4343	0.3132
communal	collaborative	0.3197	0.2783	0.1734	0.1089	0.2025	0.3607	**0.4859**	0.3822	0.3233
	enthusiastic	0.3869	0.1194	0.3404	0.2606	0.3091	0.4104	**0.5567**	0.4401	0.3869
	friendly	0.2314	0.2650	0.3247	0.1004	0.2780	0.3972	**0.5650**	0.4543	0.3247
	good-natured	0.2187	0.2708	0.3157	0.1704	0.3344	0.4897	**0.5978**	0.4179	0.2497
	kind	0.1869	0.1607	0.0920	0.1726	0.2570	0.4519	**0.4956**	0.2314	0.1037
	likeable	0.1969	0.2842	0.2333	0.0625	0.3488	0.3711	**0.5659**	0.4597	0.1963
	sincere	0.1490	0.3137	0.2675	0.0377	0.2480	0.3819	**0.5455**	0.4282	0.2817
	warm	0.2946	0.3332	0.3333	0.0275	0.2687	0.3563	**0.5459**	0.4543	0.2565
	avg.	0.2261	0.2335	0.2899	0.2233	0.2498	0.4427	**0.5795**	0.4223	0.3086

lower than the previous results using all six models. This result suggests that incorporating more diverse information from multiple models enhances prediction accuracy.

6.4 Test Results

Although we tried to improve the scores of the optim and non-optim models, we still could not achieve higher scores on the development set; however, the scores from the improved models (optim-2 and non-optim-2) show improvements in the test set. In the test set, we submitted four predictions: one from unimodal (optim) and three from performance-weighted multimodal fusions (optim, optim-2 and non-optim-2). The results are shown in Table 9. The results show that the performance-weighted fusion models outperform the unimodal model. The best score in this test is from the optim-2 model (optim model with augmentations), with a PCC of 0.3464. The worst predicted trait with that fusion model, i.e., confident with a PCC of 0.0867, is better to predict with a single model (PCC 0.3283). This finding can also be incorporated for future research and development, i.e., the use of a specific model to predict specific traits.

Table 6: Unimodal model performance from the optimized model with augmentation (optim-2) on the development set.

		Unimodal						Multimodal fusion		
	Target	ds	egemaps	w2v-msp	facenet512	fau	vit-fer	perf.	mean	max
agentive	aggressive	0.1442	0.2023	0.1031	0.3131	0.2592	0.5022	**0.5511**	0.4203	0.3253
	arrogant	0.1491	0.2734	0.1123	0.3499	0.2287	0.5952	**0.6343**	0.5913	0.4039
	assertive	0.3677	0.1654	0.3316	0.2587	0.1057	0.3527	**0.5681**	0.3436	0.0373
	confident	0.2777	0.1065	0.3479	0.2496	0.2201	0.4908	**0.5993**	0.4717	0.2201
	dominant	0.1113	0.2069	0.2842	0.4844	0.1253	0.4292	**0.5708**	0.4713	0.3855
	independent	0.1753	0.1860	0.3938	0.2890	0.1849	0.3598	**0.5825**	0.4012	0.3198
	risk-taking	0.1944	0.2414	0.4588	0.3955	0.2598	0.4703	**0.6443**	0.4430	0.3642
	leader-like	0.2084	0.0921	0.4069	0.3577	0.2081	0.4200	**0.5497**	0.3958	0.3323
comunal	collaborative	0.2372	0.2912	0.1214	0.1576	0.1982	0.2920	**0.4403**	0.3893	0.2372
	enthusiastic	0.3965	0.2601	0.3786	0.1582	0.2967	0.4210	**0.6139**	0.3220	0.2512
	friendly	0.2604	0.2442	0.1694	0.0550	0.2976	0.4433	**0.5475**	0.1271	0.0410
	good-natured	0.2693	0.2807	0.3217	0.1748	0.3130	0.4552	**0.5772**	0.4084	0.3050
	kinds	0.2455	0.1794	0.1771	0.1720	0.2430	0.4733	**0.5038**	0.2596	0.1416
	likeable	0.3090	0.2467	0.2413	-0.0066	0.3292	0.4059	**0.5509**	0.1119	-0.0715
	sincere	0.2287	0.2821	0.2651	0.1402	0.2508	0.3728	**0.5744**	0.5040	0.3012
	warm	0.2325	0.2007	0.3455	-0.0568	0.2406	0.3455	**0.4853**	0.0803	-0.1152
	avg.	0.2379	0.2162	0.2787	0.2183	0.2351	0.4268	**0.5621**	0.3588	0.2174

Table 7: PCC score of late fusions without predictions normalization; the values in parentheses are the differences from scores with normalization.

	Target	**baseline**	**non-optim**	**non-optim-2**	**optim**	**optim-2**
agentive	aggressive	**0.4151 (-0.0528)**	**0.5163 (-0.0365)**	**0.5163 (-0.0356)**	**0.4917 (-0.1258)**	**0.5099 (-0.0412)**
	arrogant	**0.3814 (-0.0764)**	**0.6274 (-0.0056)**	**0.6245 (-0.0063)**	**0.4957 (-0.1830)**	**0.6238 (-0.0105)**
	assertive	**0.3810 (-0.0805)**	**0.3856 (-0.1371)**	**0.3856 (-0.1371)**	**0.3406 (-0.2512)**	**0.3532 (-0.2149)**
	confident	**0.3456 (-0.0057)**	**0.4706 (-0.1082)**	**0.4491 (-0.1143)**	**0.5001 (-0.0901)**	**0.5371 (-0.0622)**
	dominant	**0.4799 (-0.0165)**	**0.5541 (-0.0188)**	**0.5571 (-0.0219)**	**0.5238 (-0.0923)**	**0.5270 (-0.0438)**
	independent	**0.2789 (-0.0797)**	**0.4559 (-0.0877)**	**0.4140 (-0.0543)**	**0.4080 (-0.1763)**	**0.3899 (-0.1926)**
	risk-taking	**0.5228 (-0.0153)**	**0.5359 (-0.0956)**	**0.5324 (-0.0961)**	**0.4663 (-0.1630)**	**0.4351 (-0.2092)**
	leader-like	**0.3274 (-0.0350)**	**0.5417 (-0.0420)**	**0.5345 (-0.0437)**	**0.4250 (-0.1806)**	**0.3899 (-0.1598)**
comunal	collaborative	**0.2135 (-0.0499)**	**0.3753 (-0.0368)**	**0.3781 (-0.0292)**	**0.4166 (-0.0693)**	**0.3959 (-0.0444)**
	enthusiastic	**0.3392 (-0.0121)**	**0.3941 (-0.1026)**	**0.3897 (-0.1059)**	**0.4908 (-0.0659)**	**0.3979 (-0.2160)**
	friendly	**0.2456 (-0.0373)**	**0.4336 (-0.0662)**	**0.4088 (-0.0902)**	**0.4989 (-0.0661)**	**0.3476 (-0.1999)**
	good-natured	**0.1998 (-0.0367)**	**0.5541 (-0.0191)**	**0.5275 (-0.0360)**	**0.4993 (-0.0985)**	**0.4747 (-0.1025)**
	kind	**0.1489 (-0.0262)**	**0.4400 (-0.0402)**	**0.3808 (-0.0761)**	**0.2849 (-0.2107)**	**0.3273 (-0.1765)**
	likeable	0.2633 (0.0882)	**0.5151 (-0.0463)**	**0.4951 (-0.0596)**	**0.4785 (-0.0874)**	**0.5092 (-0.0417)**
	sincere	**0.2346 (-0.0688)**	**0.4743 (-0.0308)**	**0.4814 (-0.0264)**	**0.4718 (-0.0737)**	**0.5149 (-0.0595)**
	warm	**0.1464 (-0.0035)**	**0.3733 (-0.0918)**	**0.3733 (-0.0918)**	**0.5089 (-0.0370)**	**0.4624 (-0.0229)**
	avg.	**0.3077 (-0.0318)**	**0.4780 (-0.0603)**	**0.4655 (-0.0640)**	**0.4563 (-0.1232)**	**0.4497 (-0.1124)**

7 CONCLUSIONS

In this paper, we investigated automatic social perception recognition using an unimodal model and multimodal fusion. First, we optimized unimodal based on the type of feature. Second, we utilized results from unimodal features to build performance-weighted multimodal fusion. We studied two aspects of multimodal social perception recognition: the effect of prediction normalization before weighting and the number of models in the fusion. Results show that performance-weighted fusion leads to the performance score being over mean and maximum fusions. We found that more models in the performance-weighted fusion tend to improve the prediction score over the top m models (in this case, we evaluated $m = 2, 4, 6$). Results on the development set of the LMU-ELP dataset consistently show that five evaluated fusion models using all six features outperform models from the fusion of top-2 and top-4 models. We also found that prediction normalization before weighting is crucial to improve the prediction score. The average score of the fusion from normalization before prediction weighting is also always better than without normalization. There are discrepancies between development and test results that are possibly caused by the small number and ratio of samples for training, development, and testing, which can be improved in future work.

Table 8: PCC score of late fusions from top-2 and top-4 models; see Tables 2 - 6 for comparison with scores from all (six) models.

	Target	baseline		non-optim		non-optim-2		optim		optim-2	
		Top-2	Top-4	Top-2	Top-4	Top-2	Top-4	Top-2	Top-4	Top-2	Top-4
agentive	aggressive	0.4056	**0.4748**	0.5122	**0.5507**	0.5122	**0.5506**	0.5081	**0.5963**	0.4893	**0.5250**
	arrogant	**0.4854**	0.4577	**0.6577**	0.6304	**0.6547**	0.6279	0.6094	**0.6611**	0.5755	**0.6118**
	assertive	0.4348	**0.4526**	0.4521	**0.5373**	0.4521	**0.5373**	0.4954	**0.5729**	0.5258	**0.5573**
	confident	0.3662	**0.4279**	0.4259	**0.5279**	0.4207	**0.5110**	0.5182	**0.5818**	0.5715	**0.5863**
	dominant	0.4337	**0.4818**	0.5479	**0.5761**	0.5487	**0.5795**	0.5531	**0.6152**	0.5347	**0.5593**
	independent	0.3425	**0.3605**	0.4388	**0.5422**	0.4386	**0.4807**	0.5283	**0.5828**	0.5052	**0.5582**
	risk-taking	0.5012	**0.5465**	0.4925	**0.6049**	0.4890	**0.6021**	0.5685	**0.6034**	0.5987	**0.6244**
	leader-like	0.2953	**0.3438**	0.5090	**0.5613**	0.5014	**0.5553**	0.5041	**0.5913**	**0.5615**	0.5354
communal	collaborative	0.2346	**0.2651**	0.3848	**0.4006**	0.3842	**0.3980**	0.4594	**0.4621**	0.3872	**0.4061**
	enthusiastic	0.3152	**0.3503**	0.4363	**0.4834**	0.4270	**0.4806**	0.4969	**0.5493**	0.4977	**0.5754**
	friendly	0.2687	**0.2851**	0.4735	**0.4823**	0.4607	**0.4839**	0.4757	**0.5406**	0.4809	**0.5403**
	good-natured	0.2093	**0.2311**	0.4874	**0.5710**	0.4185	**0.5526**	0.5523	**0.5922**	0.5169	**0.5551**
	kind	0.1605	**0.1729**	0.4699	**0.4731**	0.4547	**0.4588**	**0.5134**	0.4953	0.4586	**0.4870**
	likeable	0.2847	**0.3012**	0.5236	**0.5642**	0.4203	**0.5501**	0.4837	**0.5366**	0.4874	**0.5332**
	sincere	**0.2795**	0.2753	0.4051	**0.4438**	0.4051	**0.4438**	0.4179	**0.5437**	0.4281	**0.5598**
	warm	**0.1519**	0.1497	0.3843	**0.4674**	0.3843	**0.4674**	0.4247	**0.5058**	0.4320	**0.4859**
	avg.	0.3231	**0.3485**	0.4751	**0.5260**	0.4608	**0.5175**	0.5068	**0.5644**	0.5032	**0.5438**

Table 9: PCC scores of test set; sm: single model, lf: late fusion

Target	sm-optim	lf-non-optim-2	lf-optim	lf-optim-2
aggressive	0.4162	0.3733	0.4215	**0.4425**
arrogant	**0.3809**	0.3130	0.3543	0.3720
assertive	**0.3227**	0.2455	0.3102	0.2502
confident	**0.3283**	0.1011	0.1345	0.0876
dominant	0.2001	**0.3232**	0.2538	0.2393
independent	0.2592	0.3405	**0.3446**	0.3344
risk-taking	0.3682	0.4056	0.3822	**0.4773**
leader-like	0.4098	0.4293	0.4375	**0.4858**
collaborative	0.3280	**0.4217**	0.2178	0.3183
enthusiastic	0.2686	0.2433	**0.3081**	0.2956
friendly	0.3761	0.4146	**0.4519**	0.4435
good-natured	0.0802	0.2234	0.1671	**0.2248**
kind	0.2769	0.2763	**0.4092**	0.3982
likeable	0.1646	0.2491	0.1814	**0.3152**
sincere	0.3660	0.3777	0.2586	**0.4575**
warm	0.3849	0.3361	0.2837	**0.3998**
avg.	0.3082	0.3171	0.3073	**0.3464**

REFERENCES

[1] Andrea E. Abele and Bogdan Wojciszke. 2014. *Communal and agentic content in social cognition: A dual perspective model* (1 ed.). Vol. 50. Elsevier Inc. 195–255 pages. https://doi.org/10.1016/B978-0-12-800284-1.00004-7

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* (2019), 2623–2631. https://doi.org/10.1145/3292500.3330701 arXiv:1907.10902

[3] Shahin Amiriparian, Lukas Christ, Alexander Kathan, Maurice Gerczuk, Niklas Muller, Steffen Klug, Lukas Stappen, Andreas Konig, Erik Cambria, Bj{\"o}rn Schuller, and Simone Eulitz. 2014. The MuSe 2024 Multimodal Sentiment Analysis Challenge: Social Perception and Humor Recognition. In *MM 2024 - Proc. 2024 ACM Multimed. Conf.*

[4] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. 2017. Snore Sound Classification Using Image-Based Deep Spectrum Features. In *Interspeech 2017*. ISCA, ISCA, 3512–3516. https://doi.org/10.21437/Interspeech.2017-434

[5] Shahin Amiriparian, Artem Sokolov, Ilhan Aslan, Lukas Christ, Maurice Gerczuk, Tobias Hübner, Dmitry Lamanov, Manuel Milling, Sandra Ottl, Ilya Poduremennykh, Evgeniy Shuranov, and Björn W. Schuller. 2021. On the Impact of Word Error Rate on Acoustic-Linguistic Speech Emotion Recognition: An Update for the Deep Learning Era. (2021), 2–6. arXiv:2104.10121 http://arxiv.org/abs/2104.10121

[6] B. T. Atmaja and M. Akagi. 2021. Evaluation of error- And correlation-based loss functions for multitask learning dimensional speech emotion recognition. *J. Phys. Conf. Ser.* 1896, 1 (2021), 012004. https://doi.org/10.1088/1742-6596/1896/1/012004 arXiv:2003.10724

[7] Bagus Tris Atmaja and Akira Sasou. 2023. Ensembling Multilingual Pre-Trained Models for Predicting Multi-Label Regression Emotion Share from Speech. In *2023 Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*. IEEE, 1026–1029. https://doi.org/10.1109/APSIPAASC58517.2023.10317109

[8] Bagus Tris Atmaja and Akira Sasou. 2023. Evaluating Variants of wav2vec 2.0 on Affective Vocal Burst Tasks. In *ICASSP 2023 - 2023 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 1–5. https://doi.org/10.1109/icassp49357.2023.10096552

[9] Aayushi Chaudhari, Chintan Bhatt, Achyut Krishna, and Pier Luigi Mazzeo. 2022. ViTFER: Facial Emotion Recognition with Vision Transformers. *Appl. Syst. Innov. 2022* 5, 4 (2022). https://doi.org/10.3390/asi5040080

[10] Robyn Dunlop and Caren Brenda Scheepers. 2023. The influence of female agentic and communal leadership on work engagement: vigour, dedication and absorption. *Manag. Res. Rev.* 46, 3 (feb 2023), 437–466. https://doi.org/10.1108/MRR-11-2021-0796

[11] Florian Eyben, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* 7, 2 (apr 2016), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

[12] Florian Eyben, Felix Weninger, Martin Wollmer, Björn Bjorn Schuller, Martin Wöllmer, and Björn Bjorn Schuller. 2015. *OpenSMILE - The Munich versatile and fast open-source audio feature extractor.* Number December. 1–65 pages. https://doi.org/10.1145/1873951.1874246

[13] Laura Froehlich, Maria I. T. Olsson, Angela R. Dorrough, and Sarah E. Martiny. 2020. Gender at Work Across Nations: Men and Women Working in Male-Dominated and Female-Dominated Occupations are Differentially Associated with Agency and Communion. *J. Soc. Issues* 76, 3 (sep 2020), 484–511. https://doi.org/10.1111/josi.12390

[14] Ammarah Hashmi, Sahibzada Adil Shahzad, Wasim Ahmad, Chia Wen Lin, Yu Tsao, and Hsin-min Wang. 2022. Multimodal Forgery Detection Using Ensemble Learning. 1521–1529.

[15] Reza Lotfian and Carlos Busso. 2019. Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings. *IEEE Trans. Affect. Comput.* 10, 4 (2019), 471–483. https://doi.org/10.1109/TAFFC.2017.2736999

[16] Emad A. Mohammed, Mohammad Keyhani, Amir Sanati-Nezhad, S. Hossein Hejazi, and Behrouz H. Far. 2021. An ensemble learning approach to digital corona virus preliminary screening from cough sounds. *Sci. Rep.* 11, 1 (2021), 1–11. https://doi.org/10.1038/s41598-021-95042-2

[17] Nicolae Cătălin Ristea and Radu Tudor Ionescu. 2021. Self-paced ensemble learning for speech and audio classification. *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH* 2 (2021), 1276–1280. https://doi.org/10.21437/Interspeech.2021-155 arXiv:2103.11988

[18] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 07-12-June (2015), 815–823. https://doi.org/10.1109/CVPR.2015.7298682 arXiv:1503.03832

[19] Ruoqi Shen, Liyao Gao, and Yi-An Ma. 2022. On Optimal Early Stopping: Over-informative versus Under-informative Parametrization. (feb 2022), 1–30. arXiv:2202.09885 http://arxiv.org/abs/2202.09885

[20] Paul D. Trapnell and Delroy L. Paulhus. 2012. Agentic and communal values: Their scope and measurement. *J. Pers. Assess.* 94, 1 (2012), 39–52. https://doi.org/10.1080/00223891.2011.627968

[21] Andrea C. Vial and Jaime L. Napier. 2018. Unnecessary Frills: Communality as a Nice (But Expendable) Trait in Leaders. *Front. Psychol.* 9 (oct 2018). https://doi.org/10.3389/fpsyg.2018.01866

[22] Athena Vongalis-Macrow. 2016. It's About the Leadership: The Importance of Women Leaders Doing Leadership for Women. *NASPA J. About Women High. Educ.* 9, 1 (jan 2016), 90–103. https://doi.org/10.1080/19407882.2015.1114953

[23] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. 2023. Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 9 (sep 2023), 10745–10759. https://doi.org/10.1109/TPAMI.2023.3263585

[24] Bogdan Wojciszke and Olga Bialobrzeska. 2014. Agency versus Communion as Predictors of Self-esteem: Searching for the Role of Culture and Self-construal. *Polish Psychol. Bull.* 45, 4 (dec 2014), 469–479. https://doi.org/10.2478/ppb-2014-0057