Can mood primitives predict apparent personality?

Gizem Sogancioglu
Information and Computing Sciences
Utrecht University
Utrecht, Netherlands
g.sogancioglu@uu.nl

Heysem Kaya
Information and Computing Sciences
Utrecht University
Utrecht, Netherlands
h.kaya@uu.nl

Albert Ali Salah
Information and Computing Sciences
Utrecht University
Utrecht, Netherlands
Dept. Computer Engineering
Boğaziçi University
Istanbul, Turkey
a.a.salah@uu.nl

2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII) | 978-1-6654-0019-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/ACII52823.2021.9597444

***Abstract*—First impressions play a critical role in shaping social interactions and consequently have a high impact on people's lives. This study presents an explainable system that models apparent personality traits that influence first impressions as a function of automatically predicted arousal, valence and likeability (AVL) scores. To this end, we enrich the ChaLearn Looking at People - First Impressions (LAP-FI) dataset by annotating a portion of it for the AVL dimensions and carry out extensive uni-modal and multimodal experiments by using state-of-the-art acoustic, visual and linguistic features. We propose to use a glass-box model, namely, Explainable Boosting Machine, to model the Big Five personality traits. Our results demonstrate that personality trait impressions can be effectively predicted through the mood and likeability scores of a given video. We show that the proposed model, which is trained on only a few features, not only provides more meaningful explanations but also yields competitive performance (with a 0.09 Mean Absolute Error) compared to the state-of-the-art methods. The annotated benchmark dataset and the scripts to reproduce the results are available at: https://github.com/gizemsogancioglu/mood-project.**

***Index Terms*—Big Five personality traits, valence, arousal, mood recognition, affective computing, multimodal fusion**

I. Introduction and Related Work

The relationship between affect and personality is an important research question in affective computing and has applications including analysis of mental health [1], personality disorders [2], and personal assessment [3]. Several related databases were previously released for research purposes [3], [4], as well as approaches to analyse affect dimensions such as mood, emotions, and personality, including both unimodal and multimodal [5] approaches.

Apparent personality, i.e. the trait impressions perceived by an observer regarding other people, is one of the key elements in personality computing [6]. Personality is typically summarized and assessed along the "Big Five" personality traits [7]. The five factors are Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (commonly abbreviated as OCEAN). While there are other personality models, such as HEXACO [8], adding Honesty-Humility as a sixth dimension, the Big Five is used more in the literature.

A single modality may not provide sufficient information to predict real or apparent personality [3]. Although most of the earlier works have analysed a single modality, such as textual cues [9], phonetic information [10], or facial expressions [11], availability of recent multimodal affect datasets helped to investigate this problem in a richer way.

The relation of personality traits and emotions and likeability is well studied in psychology. An earlier study [12] reported substantial correlations between primary emotions, mood, and the Big Five traits, such that all traits except Neuroticism were shown to be positively correlated with a positive mood. Another study [13] showed that Extraversion is highly associated with likeability. Based on the available literature, we hypothesized that apparent personality traits can be accurately modeled with the mood states and likeability of the people.

Taking a similar approach with [4], which presents a system for the use of valence and arousal as a meta-feature for longer mood state monitoring such as depression, we used the mood and likeability as intermediary features to predict apparent personality traits. In this paper, we define the mood in terms of valence (positive vs negative) and activation (calm vs. excited), both of which are observable from expressed behaviors such as speech, facial expressions, and language.

The ChaLearn Looking at People First Impressions (LAP-FI) challenge series, which were conducted in 2016 [14] and 2017 [15], have also boosted research in the multimodal personality computing field. These challenges asked for an audiovisual prediction of (apparent) Big Five personality impressions, and whether a person would be invited to a job interview, using explainable models [3]. The BU-NKU system that won the CVPR 2017 edition of the challenge proposed audio, video, and scene-based two-staged system [16]. Two extreme learning machine (ELM) models were trained from an early fusion of face, scene, and audio features, followed by stacking the predictions of sub-systems to an ensemble of decision trees. The most recent studies on this task (published after the challenge) employ a trimodal (audio, visual and linguistic) approach, showing positive contribution via linguistic modeling [17], [18]. In all state-of-the-art systems, the visual modality is found to be the most predictive and extensively benefits from transfer learning. Aslan et al. [18] proposed more complex deep multimodal architectures with modality-specific deep sub-networks, which are subsequently combined and complemented by feature attention and regression layers.

978-1-6654-0019-0/21/$31.00 ©2021 IEEE

TABLE I
AVERAGE INTER RATER AGREEMENT (COHEN'S KAPPA) SCORES USING THREE WEIGHTING SCHEMES.

Task	Unweighted	Linear	Quadratic
Arousal	0.37	0.41	0.48
Valence	0.36	0.39	0.44
Likeability	0.23	0.30	0.41

TABLE II
CLASS DISTRIBUTION OF THE NEWLY ANNOTATED DIMENSIONS.

Task	Low	Medium	High
Arousal	57	588	315
Valence	36	709	215
Likeability	119	655	186

Li et al. approached the problem differently from the previous studies and proposed a deep Classification-Regression network to overcome the regression-to-mean problem [17]. Most of the successful studies used deep architectures with high-dimensional feature spaces, which makes interpretability difficult.

We used the ChaLearn LAP-FI dataset (detailed in the next section) to evaluate our approach to predict the personality impressions. First, we annotated a portion of the dataset for the new AVL dimensions. By using the three different modalities (transcript, audio, and video) that are already available, an extensive set of uni-modal and multimodal experiments with an SVM classifier were conducted for the classification of arousal, valence, and likeability. The predictions of the AVL models were used as mid-level features for the second-level learner. Another motivation to use mood and likeability as features in this study was to increase the explainability of the systems by using a few and easily understandable features unlike the previous studies on this dataset. For this reason, Explainable Boosting Machine (EBM) [19], which is an explainable model by its nature, was used as a second-level learner.

The contributions of this paper are twofold:

- We demonstrate that using mood and likeability as mid-level features can effectively help recognize apparent personality traits and obtain comparable performance to the literature in addition to providing more transparent predictions.
- We present a set of manual annotations and benchmark experiments for AVL dimensions that can be used for training and evaluation in future studies. Experiments were performed on this set using the state-of-the-art linguistic, acoustic, and visual feature sets.

II. DATASET

We used the publicly available ChaLearn LAP-FI dataset [3], [14], which comprises 10,000 video clips with an average duration of 15 seconds. The clips were collected from over 3,000 videos available on YouTube. The videos were annotated for apparent personality traits, using the Big Five model, and later normalized into [0, 1].

Besides sensory data, the dataset also contains manual text transcriptions of the videos. In total, 435.984 words were transcribed (183.861 non-stopwords), which corresponds to 43 words per clip on average. The ethnicity, age, and sex of people in the videos were annotated later to investigate annotator and algorithm biases [3].

In order to answer our research question of whether the overall apparent mood in the video is significant to predict apparent personality, we enriched a portion of the ChaLearn dataset (a total of 960 video clips) with three new dimensions: valence, arousal, and likeability. These dimensions were annotated by three different annotators (Gender: all male, Age: 21-22, Native language: Dutch) for three categories (1:low, 2:medium, 3:high). Most voted class among three different annotations was assumed as a ground-truth value for the corresponding clip. In case of a tie, the "low" class was used, i.e. we broke the tie in favor of the minority class (see Table. II).

As shown in Table I, Kappa scores for unweighted (original), linear, and quadratic weighting (also reported, since categories are ordinal) schemes were computed to measure the inter-rater agreement of the annotations and show fair to a moderate agreement. The arousal dimension has the highest agreement among annotators, the agreement level for likeability is lower than both valence and arousal annotations. While Kappa scores are admittedly low due to the subjective nature of the assessments, we illustrate next that a system built using these annotations is still capable of achieving state-of-the-art performance.

Table II shows a histogram of the AVL annotations. For all three dimensions, annotations of the dataset were imbalanced, as more than 60 % of the examples were annotated in the 'medium' class. We also observed that there are very few examples belonging to the 'low' class, especially for valence and arousal dimensions, which indicates a possible bias in the YouTube videos.

The annotated dataset was partitioned into development (training/validation) and test sets to set a standard for future studies. The development partition consists of 660 examples, while the remaining 300 instances served as the test set.

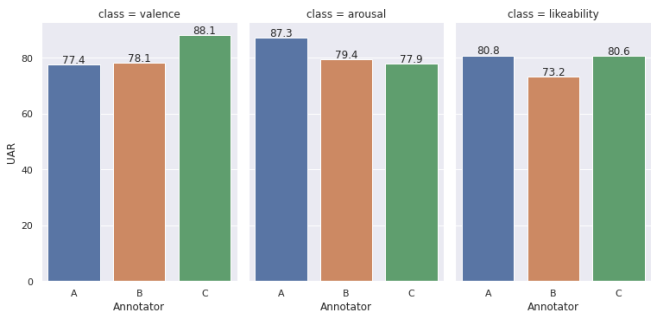

Fig. 1. The test set UAR (%) performance of each human annotator with respect to the ground truth.

We computed each annotator's performance to assess task difficulty. Figure 1 shows the test set unweighted average recall

TABLE III
MEAN ANNOTATIONS PER SENSITIVE GROUP. AROUSAL MEAN: 2.27, VALENCE MEAN: 2.19, LIKEABILITY MEAN: 2.07

	Arousal	Valence	Likeability
Female	2.29	2.24	2.18
Male	2.24	2.13	1.94
African American	2.27	2.13	2.06
Caucasian	2.27	2.19	2.07
Asian	2.13	2.33	1.97
Female (Younger)	2.28	2.24	2.26
Female (Older)	2.31	2.22	1.92
Male (Younger)	2.27	2.15	1.93
Male (Older)	2.19	2.05	1.99

(UAR) (%) scores of each annotator's labels per dimension with respect to the ground truth. The lowest accuracy scores among annotators are 77.4, 77.9, 73.2 for valence, arousal, likeability, respectively. These results are rather optimistic since annotators voted for the ground-truth label. However, scores still give a good idea about human-level performances on a given task and set upper bound algorithmic approaches. Both upper-bound scores and inter-rater agreement scores given in Table I show that likeability is a more difficult dimension to predict accurately.

A. Limitations

We observe two important limitations about the ChaLearn dataset and the annotations. Firstly, since the main task was recognition of first impressions about personality from short videos, the original video clips were randomly cut off to 15 seconds. Most of the transcripts thus end with incomplete sentences. Unfortunately, this causes losing an important linguistic context and weaker representation compared to other modalities.

Secondly, the joint study of the ChaLearn organizers and the competitors [3] showed that personality trait annotations are biased towards sensitive groups such as gender and ethnicity. We performed a similar analysis and reported the mean value of each sensitive group for the new AVL annotations in Table III. We classified the numerical age annotations as 'younger' and 'older' by setting a threshold ($t = 33$) based on the distribution of the examples in our dataset. For the likeability dimension, a larger gap was observed between sensitive groups of gender and ethnicity. Interestingly, while younger female groups had a higher likeability score than older females, older males were considered to be more likeable compared to younger males. A similar bias was also reported in [3]. As such annotator biases are learned by machine learning models, we caution the readers in the usage of these methods: They are suitable for assessing the assessors and for training purposes, but they should not be used directly for screening job candidates.

B. Correlation Analysis

Figure 2 illustrates the correlation matrix of personality traits and mood variables for the annotated portion of the ChaLearn First Impression dataset (960 clips). We first observe

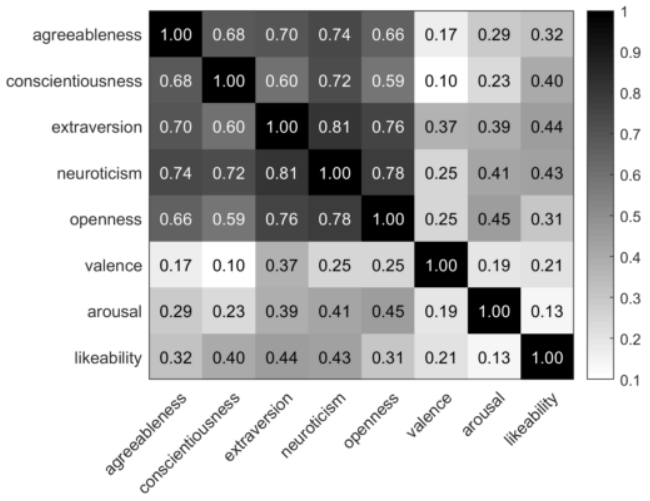

Fig. 2. Pearson correlation matrix of target variables (all correlations are significant at $p < .0001$ level).

that all correlations are positive, noting that Neuroticism scores in this dataset refer to Non-Neuroticism. The second general observation is that the majority of the inter-trait correlations are substantial (higher than 0.6), while inter-state (mood/likability) correlations are fair at best. The high inter-trait correlations imply that there may be a common factor (e.g. an overall impression left on the annotator) that affects all personality trait impressions. We observe moderate trait-state correlations, particularly between arousal and OEN (openness, extroversion, and neuroticism) trait impressions as well as likeability and CEN (conscientiousness, extroversion, and neuroticism) trait impressions. Although lower than inter-trait correlations, these moderate state-trait correlations motivate an investigation to use them in a two-level framework as interpretable mid-level predictive features for trait impressions.

III. METHODOLOGY

In this section, we present the proposed feature sets for apparent personality analysis. As depicted in Fig.III-C, the proposed system consists of two main components; 1. mood and likeability classification through a support vector machine (SVM) classifier, 2. personality trait impression prediction using the predicted mood and likeability scores with an explainable boosting machine (EBM) regressor.

A. Feature Extraction

We experimented with a rich set of state-of-the-art and hand-crafted linguistic features that can be correlated with personality traits and affect. On the other hand, for acoustic and visual modalities, we followed the work of Kaya et al. [16]. We selected the visual feature set used in their system that won the ChaLearn LAP challenge.

1) Linguistic Features: As a common linguistic baseline, Term Frequency-Inverse Document Frequency (TF-IDF) was used. Two affective lexicon-based approaches, a state-of-the-art (SoA) embedding method, and two statistical feature sets were experimented with. We describe each of these in turn.

a) Text Entity Per Second (TEPS): The number of the word- and sentence entities per second were used as one of the linguistic features in this study. Since the speech length is fixed with 15 seconds for all the videos, the total number of words and sentences were divided by 15. The resulting two-dimensional features are not purely linguistic, because speech duration was also exploited to construct these features.

b) Sentence-BERT: BERT [20] is a contextualized word embedding method, which is the SoA for various Natural Language Processing (NLP) tasks including sentiment analysis and semantic textual similarity. We used the pre-trained Sentence-BERT [21] model, which is a modification of BERT word embeddings to sentence-level space. Each transcript was processed by the Sentence-BERT encoder to construct 768-dimensional vectors. We used the BERT-Base-NLI, which was pre-trained on the NLI dataset [21].

c) Valence, Arousal, Dominance (VAD): As an affect lexicon-based feature, the NRC VAD lexicon [22], consisting of more than 20,000 English words and their VAD scores, was used. After extracting the scores for each word, functional statistics, namely mean, standard deviation, minimum, maximum, range, and sum, were computed per transcript. Scores of the words that do not exist in the VAD lexicon were considered as 0 (disregarded for the computation of the minimum statistic).

d) TF-IDF: TF-IDF is commonly used as a text representation technique in NLP studies [23]. It was used as the baseline linguistic feature in this study. In the preprocessing step, stop words that are available in the NLTK library [24] were removed and stemming was applied by using the Porter Stemmer algorithm [25]. TF-IDF weights were computed over the set of uni-grams and bi-grams. Only the most frequent 500 entities were used for feature representation to reduce the dimension of the highly sparse vector.

e) Linguistic Inquiry and Word Count (LIWC): Earlier studies [9] evidenced that there are small but significant correlations between linguistic dimensions and personality traits. As words and the ways people use them can provide rich information about their relationships, personalities, emotions, and many more dimensions, Pennebaker et al. [26] developed the LIWC tool, which allows doing text analysis by means of rich dictionaries and pre-defined categories. We used the LIWC 2015 tool to extract information from the given text about 93 LIWC categories, which can be grouped into main categories such as affect information, language metrics, informal speech, etc. All the features that were extracted by the tool were used, without any pre-selection.

f) Polarity Statistics: As sentiment analysis is a hot research topic in NLP, various pre-trained models and tools have been made available for research purposes. In this study, three SoA sentiment analysis libraries, namely NLTK Vader [27], TextBlob [28], and Flair [29], were used to extract polarity features from the transcripts. Since each of these libraries have some strengths and drawbacks in different dimensions for assessing the sentiment of the sentences, we combined the features from these three methods to benefit from the strengths of each approach. However, those libraries are designed to work at the sentence level, while transcripts in our dataset may consist of more than a sentence. To compute the polarity scores over a transcript, sentence-level scores are summarized with the same set of functional statistics that were used for calculating the VAD features.

2) Acoustic Features: The open-source openSMILE[1] tool [30] is popularly used to extract acoustic features in a number of international paralinguistic and multimodal challenges [31], [32]. The idea is to obtain a large pool of potentially relevant features by passing an extensive set of summarizing functionals on the low level descriptor contours (e.g. Mel Frequency Cepstral Coefficients, pitch, energy, and their first/second order temporal derivatives). We use the toolbox with a standard feature configuration (called IS13 hereafter) that served as a baseline for challenges since the INTERSPEECH 2013 Computational Paralinguistics Challenge [31], [32]. This configuration was found to be the most effective acoustic feature set for personality impression prediction [33].

3) Visual Features: Following the winner systems in ICPR 2016 and CVPR 2017 ChaLearn LAP-FI challenges [16], [33], we use embeddings from a fine-tuned CNN for face representation. Facial features are extracted over an entire video segment and summarized by statistical functionals. Faces are detected on all frames of the video input. The Supervised Descent Method (SDM) is used for face registration, which gives 49 landmarks on each detected face [34]. The roll angle is estimated from the eye corners to rotate the image accordingly. Then a margin of 20% of the interocular distance around the outer landmarks is added to crop the facial image. Faces are detected, aligned, and resized to 64×64 pixels. After aligning the faces, image-level deep features are extracted from a convolutional neural network (CNN) trained for facial emotion recognition. A deep neural network pre-trained with VGG-Face [35] and fine-tuned with FER-2013 database [36] is used from [37]. The final trained network has a 37-layer architecture (involving 16 convolution layers and 5 pooling layers). The response of the 33^{rd} layer, which is the lowest-level 4 096-dimensional embedding, is used.

After extracting the frame-level features from each aligned face using the fine-tuned CNN, videos are summarized by computing functional statistics of each dimension over time, including mean, standard deviation, offset, slope, and curvature. Offset and slope are calculated from the first order polynomial fit to each feature contour, while curvature is the leading coefficient of the second order polynomial. An empirical comparison of these individual functionals is given in [38].

Unlike the winning systems in the ChaLearn LAP-FI challenges, we do not use Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) video descriptor [39], due to its very high-dimensionality (~100K). However, we should note that LGBP-TOP representation of the face sequence was

[1]Available from https://www.audeering.com/opensmile/

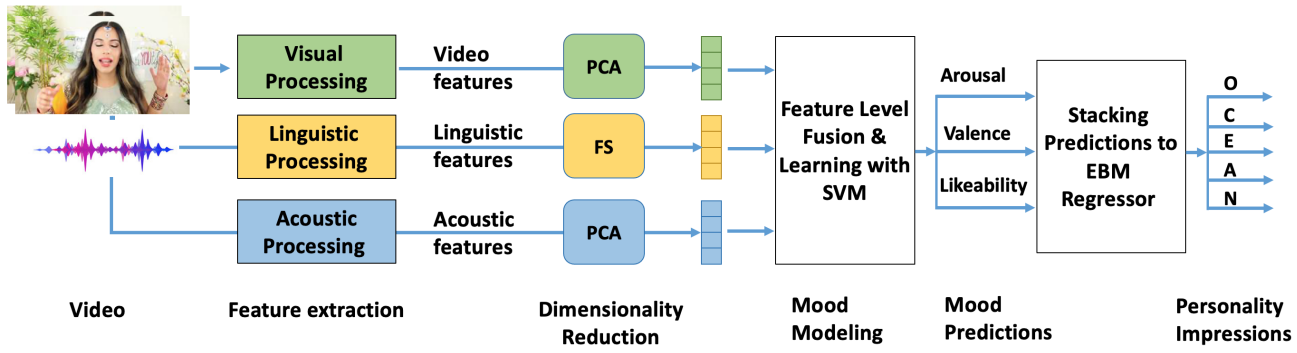

Fig. 3. Pipeline of the proposed two-staged personality traits prediction system. FS: Feature Selection, PCA: Principal Component Analysis, SVM: Support Vector Machine, EBM: Explainable Boosting Machine.

TABLE IV
3-FOLD CROSS-VALIDATION UAR (%) PERFORMANCES OF UNI-MODAL (PART I AND II) AND MULTIMODAL (PART III) MODELS. FOR UNI-MODAL MODELS, THE DIMENSION OF THE FEATURE VECTOR IS GIVEN IN PARENTHESES. FF: FEATURE FUSION.

Features	Arousal	Valence	Likeability
LIWC (93)	47.58	42.96	34.43
Polarity Stats (35)	48.46	47.49	36.47
TEPS (2)	55.61	43.83	42.58
VAD (18)	53.26	42.67	40.03
TF-IDF (500)	34.96	37.40	34.52
Sentence-BERT (768)	43.30	48.90	37.14
IS13 (1309)	65.73	37.62	48.15
VGG-FER (2138)	48.82	51.75	52.73
FF(VAD, IS13, VGG-FER)	**70.02**	54.45	55.47
FF(Polarity, TEPS, VGG-FER)	52.79	**61.76**	53.63
FF(TEPS, VAD, IS13, VGG-FER)	69.08	55.60	**55.99**

found to yield the most successful uni-modal descriptor for predicting the personality impressions [16], [33].

B. Mood/Likeability Prediction

The first stage of the proposed system is the prediction of mood and likability variables from the given video. As mentioned in Section II, 660 examples were used for training and validation, while the remaining 300 examples were used as the test set. Hyperparameter tuning was performed with a 3-fold Cross Validation method. Separate SVM classifiers were trained for valence, arousal, and likeability. To overcome the data imbalance problem stated above, class weights inversely proportional to respective prior probabilities were used in training the SVMs.

Using the features explained in Section III-A, extensive uni-modal and multimodal experiments were performed. Because of the high dimensionality of acoustic (6374) and visual features (20480), not only the end-results are un-interpretable, but also the training procedure is costly. With the purpose of improving these issues, principal components analysis (PCA) was applied to both acoustic and visual features, resulting in reduced dimensions of 1309 and 2138, respectively. We have selected the projection dimensionality that retains % 99 of the variance.

Along with uni-modal experiments, early fusion experiments on all the combinations of different modalities were conducted. Before using feature fusion with acoustic and visual modalities, a further stage of feature reduction using the L1 feature selection method available in Scikit-learn [40] was applied to avoid dominance over lower dimensional linguistic features. As a linear model penalizing L1 error, the linear SVM model was used and only the features having non-zero coefficients were kept by the model. Then, the retained features were concatenated with the other modality features, and the same training procedure was applied with the SVM classifier.

C. Personality Impression Prediction

As illustrated in Figure 3, apparent personality traits were modeled as a function of mood and likeability prediction probabilities that are provided by the trained models described in Section III-B. Since one of the goals of this study is to promote an intelligible model with a small set of features that help us to understand what is learned and the reason for the decisions made by the model, a glass-box model, explainable boosting machine (EBM) [19], was used as a regression model.

EBM is a type of Generalized Additive Model (GAM) [41], whose formula is given in Eq. 1.

$$g(y) = f_0 + \sum f_j(x_j) \tag{1}$$

GAM is interpretable because the impact of each feature x_j and the learned function f_j can be known and visualized. As two key improvements over traditional GAMs, EBM uses modern machine learning methods such as bagging and gradient boosting [42] and includes the pairwise interaction terms as given in Eq. 2.

$$g(y) = f_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j) \tag{2}$$

This extended version has accuracy comparable to the SoA techniques such as SVM and Random Forests, and additionally, is highly intelligible and explainable due to the GAM-based additive structure of the model.

We used only low and high classes' probabilities of mood and likeability models to prevent co-linearity among features. This results in six continuous features. The original training

TABLE V
TEST SET UAR (%) PERFORMANCES OF THE MODELS THAT PERFORMED BEST ON THE VALIDATION SET.

Features	Arousal	Valence	Likeability
Polarity Stats	43.89	49.63	31.74
TEPS	53.71	38.07	37.22
VAD	56.97	33.44	37.08
IS13	60.29	34.19	41.4
VGG-FER	46.12	52.64	50.36
FF(VAD, IS13, VGG-FER)	**71.15**	52.98	48.50
FF(Polarity, TEPS, VGG-FER)	46.29	**54.29**	**55.21**
FF(TEPS, VAD, IS13, VGG-FER)	70.53	53.72	50.27

(6000), development (2000), and test set (2000) partitions given in the ChaLearn LAP-FI challenge were used for personality impression prediction experiments.

IV. EXPERIMENTAL RESULTS

In this section, first, we present the preliminary results for the mood and likeability models, which comprise the first stage of the proposed framework. Next, we show the results of the apparent personality recognition system and compare it with the previous studies that are described in Section I.

A. Mood/Likeability Preliminary Results

As explained in Section II, mood and likeability annotations were done in three categories, high, medium, and low. To evaluate the performance of the ternary SVM classifiers, an unweighted average recall (UAR) measure was chosen due to its common use as a performance measure for imbalanced datasets [31], [32].

Table IV gives the validation set performances of uni-modal and top-performed multimodal models, which were obtained with feature-level fusion (FF). Results show that visual features (VGG-FER) perform best for both valence and likeability dimensions, while acoustic features obtain the highest score for arousal prediction, which is in line with the literature [43].

A chance-level baseline returning the majority ("medium") class for all the examples gives a 33.3% UAR score. The TF-IDF model that was used as a strong linguistic baseline performs just slightly higher than this simple approach. We observed that different sets of linguistic features perform well for different dimensions. For example; on the arousal task, TEPS and VAD features gave 55.61% and 53.26% validation set UAR scores, respectively, which are quite higher than the baseline and outperform the visual model. On the other hand, Polarity Stats and Sentence-BERT outperform other linguistic approaches and the acoustic model for the valence task. Although the linguistic attributes alone did not rank best in any of the three dimensions, they contributed to all of the most successful multimodal models. Fusing visual features with Polarity Stats and TEPS yielded a ~10% absolute UAR improvement over the highest uni-modal score for valence model while combining VAD features with the acoustic and visual features improved the performance of the arousal model by ~5%. Due to the page limit, we were not able to report all feature fusion experiments. These are made available as extra material on the Github repository of the paper.

Test set performances of the multimodal models that performed best on the validation set are shown in Table V along with the performance of their uni-modal components. Results are consistent with the upper-bound scores that were determined for each AVL dimension. While the multimodal models performing best on the validation set also obtained the highest score for arousal and valence, we did not see this pattern for the likeability dimension. The top likeability fusion scheme on the validation set does not generalize as well as the model that fuses Polarity, TEPS, and VGG-FER features.

B. Personality Impression Prediction over Mood and Likeability

We employ the Mean Absolute Error (MAE) as the evaluation measure for a fair comparison with the LAP-FI challenge results. The performance of each trait is evaluated in terms of this measure, which is formulated as:

$$E = \frac{1}{N} \sum |y_i - \hat{y_i}|, \tag{3}$$

where N indicates the number of predicted samples, while y_i and $\hat{y_i}$ denote the ground truth and predicted value of sample i, respectively.

As explained in detail in Section III-C, the EBM model was trained to predict the personality traits over mood and likeability predictions. For each dimension, predictions of the models that were most successful on the validation set (regardless of the test set performance, obviously) were used as features at this stage. The test set performances of the proposed system and some of the previous studies on this dataset are given in Table VI. The baseline is a simple, but effective approach that returns the per-trait training set average for all test examples. The models using only valence, arousal, and likeability features obtain 0.113, 0.108, 0.107 MAE, respectively, each of which outperforming the baseline. Moreover, the combination of AVL features yielded an average MAE of 0.098, outperforming some of the previous studies. However, the proposed system does not outperform the state-of-the-art (SoA) models that use much more complex features and architectures. We should also note that those systems are trained with a large set of annotated data (6000 train + 2000 validation set video clips annotated for OCEAN), while we have used only 660 video clips to train and validate our mood and likeability recognition models, as our goal was to investigate the effectiveness of mood based explainable and cost-efficient modeling.

The relative importance of EBM features for each personality trait is visualized in Fig. 4. On the overall, negative weights for low and positive weights for high personality impressions is inline with the former works in social psychology [12]. Although as a single factor, likeability has the strongest impact (according to Table VI, with 0.107 MAE), results show that different dimensions have the highest impact on the recognition of personality traits in the trained model. For

TABLE VI
TEST SET PERFORMANCE OF THE PROPOSED MODELS IN TERMS OF MEAN ABSOLUTE ERROR.

Method	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness	MEAN
Ours	0.097	0.107	0.094	0.099	0.095	0.098
Valence	0.105	0.124	0.108	0.116	0.109	0.113
Arousal	0.101	0.120	0.107	0.107	0.103	0.108
Likeability	0.104	0.115	0.103	0.111	0.103	0.107
Baseline [3]	0.107	0.126	0.122	0.123	0.117	0.119
Vo et al. [5]	0.104	0.123	0.118	0.123	0.114	0.116
ROCHCI [3]	0.097	0.105	0.097	0.099	0.095	0.099
Gurpınar et al. [38]	0.093	0.087	0.084	0.098	0.092	0.091
Kaya et al. [16]	0.086	0.080	0.079	0.085	0.083	0.083
Aslan et al. [18]	0.084	0.078	0.080	0.085	0.084	0.082
Li et al. [17]	0.082	0.078	0.080	0.085	0.081	0.081

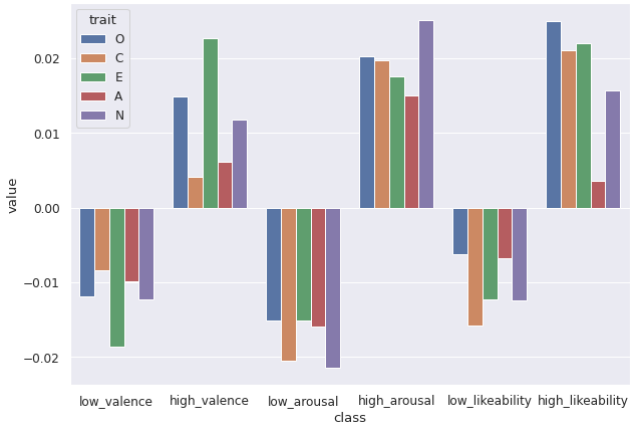

Fig. 4. Relative importance of features for the indirectly modeled personality trait impressions.

instance, valence was found to be the most important factor to predict extraversion, closely followed by likeability; while the predicted arousal score of the person has the highest impact on the remaining four personality impressions. On the other hand, the probability of high likeability has the biggest importance for openness to experience and conscientiousness. These findings are corroborated by the test set prediction performances reported in Table VI, where we observe that likeability, which is less studied in affective computing compared to the other two dimensions, exhibits the highest performance in three impressions including extraversion, which is inline with [13].

V. DISCUSSION

Inspired by the significant correlation between mood, likeability, and personality traits available in the literature, in this study, we investigated whether we can implement a successful personality prediction model by using mood predictions on short videos of people. To the best of our knowledge, this is the first study that uses predicted mood dimensions as features to model apparent personality traits directly. We demonstrate that mood and likeability features can effectively predict apparent personality traits. Although not accurate as the state-of-the-art methods, we have shown that the proposed glass-box model with a small set of intelligible features can produce insightful explanations. When we analysed the explanations that were provided by the EBM model, we observed that trait predictions were linearly correlated with the valence, arousal, and likeability values. We observed that the highest total importance is attributed to arousal, which may be explained by the relatively higher predictive accuracy of this dimension. The proposed approach yields comparable performance to the current state-of-the-art methods, although it learns the personality traits from only six features of mood and likeability and provides explanations for the user to understand the underlying decisions made.

Another important contribution of this study is that we provide a strong baseline, as well as enriched annotations for valence, arousal, and likeability dimensions on a portion of the ChaLearn dataset. The manually annotated dataset is made publicly available for research purposes in this domain. Preliminary unimodal and multimodal experiments were conducted using a clear experimental protocol to provide a comparable baseline for future studies. Experimental results demonstrate that although unimodal features were more successful in the prediction of individual traits (audio features for arousal and visual features for valence and likeability), early fusion of the three modalities yields the highest performance for each mood and likeability dimension.

It should be noted that the proposed model's accuracy is dependent on the performance of the mood and likeability prediction systems. There is a multitude of ways to improve the accuracy of the first stage, such as increasing the size of the training dataset, and upsampling the minority classes (low and high) to balance the training data distribution. On the other hand, achieving good performance by using only six meta-level mood features is very promising. Moreover, recent studies showed that both the dataset and the top algorithms trained on it carry some gender and ethnicity bias [3], [44], [45]. In terms of biases towards sensitive groups, the proposed approach serves to highlight them and to provide opportunities for systematic analysis. Bias mitigation can be done via pre-processing/balancing the data, but this will result in losing representativeness. Since we recommend using these systems for training and to gain a better understanding of the biases instead of directly for job screening, we suggest using bias mitigation strategies outside the system (e.g. as post-processing).

REFERENCES

[1] D. N. Klein, R. Kotov, and S. J. Bufferd, "Personality and depression: explanatory models and review of the evidence," *Annual review of clinical psychology*, vol. 7, p. 269, 2011.

[2] L. J. Simms, T. Yufik, and D. F. Gros, "Incremental validity of positive and negative valence in predicting personality disorder." *Personality Disorders: Theory, Research, and Treatment*, vol. 1, no. 2, p. 77, 2010.

[3] H. J. Escalante, H. Kaya, A. A. Salah, S. Escalera, Y. Güç, U. Güçlü, X. Baró, I. Guyon, J. C. Jacques, M. Madadi *et al.*, "Modeling, recognizing, and explaining apparent personality from videos," *IEEE Transactions on Affective Computing*, 2020.

[4] S. Khorram, M. Jaiswal, J. Gideon, M. McInnis, and E. M. Provost, "The priori emotion dataset: Linking mood to emotion detected in-the-wild," *arXiv preprint arXiv:1806.10658*, 2018.

[5] N. N. Vo, S. Liu, X. He, and G. Xu, "Multimodal mixture density boosting network for personality mining," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2018, pp. 644–655.

[6] X.-S. Wei, C.-L. Zhang, H. Zhang, and J. Wu, "Deep bimodal regression of apparent personality traits from short video sequences," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 303–315, 2017.

[7] W. T. Norman, "Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings." *The Journal of Abnormal and Social Psychology*, vol. 66, no. 6, p. 574, 1963.

[8] K. Lee and M. C. Ashton, "Psychometric properties of the HEXACO personality inventory," *Multivariate behavioral research*, vol. 39, no. 2, pp. 329–358, 2004.

[9] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of artificial intelligence research*, vol. 30, pp. 457–500, 2007.

[10] G. Mohammadi and A. Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 273–284, 2012.

[11] A. Basu, A. Dasgupta, A. Thyagharajan, A. Routray, R. Guha, and P. Mitra, "A portable personality recognizer based on affective state classification using spectral fusion of features," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 330–342, 2018.

[12] K. L. Davis and J. Panksepp, "The brain's emotional foundations of human personality and the affective neuroscience personality scales," *Neuroscience & Biobehavioral Reviews*, vol. 35, no. 9, pp. 1946–1958, 2011.

[13] D. Van der Linden, R. H. Scholte, A. H. Cillessen, J. te Nijenhuis, and E. Segers, "Classroom ratings of likeability and popularity are related to the big five and the general factor of personality," *Journal of Research in Personality*, vol. 44, no. 5, pp. 669–672, 2010.

[14] 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 *Proc. ECCV*. Springer, 2016, pp. 400–418.

[15] S. Escalera, X. Baró, H. J. Escalante, and I. Guyon, "Chalearn looking at people: A review of events and resources," in *Proc. IJCNN*, 2017, pp. 1594–1601.

[16] H. Kaya, F. Gürpınar, and A. Ali Salah, "Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video CVs," in *Proc. CVPRW*, 2017.

[17] 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*, pp. 1–18, 2020.

[18] S. Aslan, U. Güdükbay, and H. Dibeklioğlu, "Multimodal assessment of apparent personality using feature attention and error consistency constraint," *Image and Vision Computing*, p. 104163, 2021.

[19] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "InterpretML: A unified framework for machine learning interpretability," *arXiv preprint arXiv:1909.09223*, 2019.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[21] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proc. EMNLP*, 2019.

[22] S. M. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proc. ACL*, 2018.

[23] D. Hiemstra, "A probabilistic justification for using TF $\times$ IDF term weighting in information retrieval," *International Journal on Digital Libraries*, vol. 3, no. 2, pp. 131–139, 2000.

[24] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[25] M. F. Porter *et al.*, "An algorithm for suffix stripping." *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[26] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," Tech. Rep., 2015.

[27] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th International AAAI Conference on Weblogs and Social Media*, 2014.

[28] S. Loria, "Textblob documentation," *Release 0.15*, vol. 2, 2018. [Online]. Available: https://textblob.readthedocs.io/en/dev/

[29] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "Flair: An easy-to-use framework for state-of-the-art nlp," in *Proc. NAACL-HTL (Demonstrations)*, 2019, pp. 54–59.

[30] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich Versatile and Fast open-source Audio Feature Extractor," in *Proc. ACM MM*, 2010, pp. 1459–1462.

[31] B. Schuller *et al.*, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *INTERSPEECH*, Lyon, France, Proceedings, 2013, pp. 148–152.

[32] ——, "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates," in *Proceedings of Interspeech*, Brno, Czechia, September 2021, p. 5 pages, to appear.

[33] F. Gürpınar, H. Kaya, and A. A. Salah, "Multimodal Fusion of Audio, Scene, and Face Features for First Impression Estimation," in *Proc. ICPR*, 2016, pp. 43–48.

[34] X. Xiong and F. de la Torre, "Supervised Descent Method and Its Application to Face Alignment," in *Proc. CVPR*, 2013, pp. 532–539.

[35] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015.

[36] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Proc. NeurIPS*, 2013, pp. 117–124.

[37] H. Kaya, F. Gürpınar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image and Vision Computing*, vol. 65, pp. 66–75, 2017.

[38] F. Gürpınar, H. Kaya, and A. A. Salah, "Combining deep facial and ambient features for first impression estimation," in *Proc. ECCVW*, 2016, pp. 372–385.

[39] T. R. Almaev and M. F. Valstar, "Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *Proc. ACII*, 2013, pp. 356–361.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[41] T. Hastie and R. Tibshirani, "Generalized additive models: some applications," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 371–386, 1987.

[42] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[43] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[44] A. Köchling, S. Riazy, M. Wehner, and K. Simbeck, "Highly accurate, but still discriminatory: A fairness evaluation of algorithmic video analysis," in *Academy of Management Proceedings*, vol. 2020, no. 1. Academy of Management Briarcliff Manor, NY 10510, 2020, p. 13339.

[45] S. Yan, D. Huang, and M. Soleymani, "Mitigating biases in multimodal personality assessment," in *Proc. ICMI*, 2020, pp. 361–369.