frontiers in Computer Science

ORIGINAL RESEARCH
published: 07 January 2022
doi: 10.3389/fcomp.2021.770492

Predicting Activation Liking of People With Dementia

Lars Steinert, Felix Putze, Dennis Küster and Tanja Schultz*

Cognitive Systems Lab, Department of Mathematics and Computer Science, University of Bremen, Bremen, Germany

Physical, social and cognitive activation is an important cornerstone in non-pharmacological therapy for People with Dementia (PwD). To support long-term motivation and well-being, activation contents first need to be perceived positively. Prompting for explicit feedback, however, is intrusive and interrupts the activation flow. Automated analyses of verbal and non-verbal signals could provide an unobtrusive means of recommending suitable contents based on implicit feedback. In this study, we investigate the correlation between engagement responses and self-reported activation ratings. Subsequently, we predict ratings of PwD based on verbal and non-verbal signals in an unconstrained care setting. Applying Long-Short-Term-Memory (LSTM) networks, we can show that our classifier outperforms chance level. We further investigate which features are the most promising indicators for the prediction of activation ratings of PwD.

Keywords: dementia, activation, rating prediction, engagement, LSTM

OPEN ACCESS

Edited by:
Youngjun Cho,
University College London,
United Kingdom

Reviewed by:
Saturnino Luz,
University of Edinburgh,
United Kingdom
Emilie Brotherhood,
University College London,
United Kingdom

****Correspondence:***
Lars Steinert
lars.steinert@uni-bremen.de

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *03 September 2021*
Accepted: *13 December 2021*
Published: *07 January 2022*

Citation:
Steinert L, Putze F, Küster D and Schultz T (2022) Predicting Activation Liking of People With Dementia. Front. Comput. Sci. 3:770492. doi: 10.3389/fcomp.2021.770492

1. INTRODUCTION

Dementia describes a syndrome that is characterized by the loss of cognitive function and behavioral changes. This includes memory, language skills, and the ability to focus and pay attention (WHO, 2017). It has been shown that the physical, social, and cognitive stimulation of People with Dementia (PwD) has significant positive effects on their cognitive functioning (Spector et al., 2003; Woods et al., 2012) and can lead to a higher quality of life (Schreiner et al., 2005; Cohen-Mansfield et al., 2011). It is furthermore often (implicitly) assumed, that activation contents need to be perceived positively to help maintain long-term motivation and well-being. This can be supported by a recommender system that suggests appropriate activation contents. Here, an activation content is defined as a stimulus of a certain type (image gallery, video, audio, quiz, game, phrase or text) on a certain topic, e.g. gardening, sports, or animals to cognitively, socially, or physically activate PwD and which aims for the general maintenance or enhancement of the according functions (Clare and Woods, 2004). However, prompting for explicit user feedback is intrusive as it disturbs the activation flow. Studies have shown that verbal and non-verbal signals can be promising indicators for the internal states of healthy individuals (Masip et al., 2014; Tkalčič et al., 2019). Even PwD who might suffer from blunted affect or aphasia, might remain able to provide verbal and non-verbal signals throughout all stages of the disease (Steinert et al., 2021). For this study, we use the I-CARE dataset (Schultz et al., 2018, 2021) which consists of verbal and non-verbal signals of PwD who used a tablet-based activation system over multiple sessions in an unconstrained care setting. Previous studies have already investigated the recognition of engagement of PwD (Steinert et al., 2020, 2021), which is defined as "the act of being occupied or involved with an external stimulus" (Cohen-Mansfield et al., 2009). Here, we explicitly consider the argument that activation contents should not only be engaging but also need to be perceived positively to maintain long-term motivation and well-being. In this study, we thus first investigate the correlation between engagement responses and self-reported activation ratings.

Second, we analyze if self-reported activation ratings of PwD can be predicted based on verbal and non-verbal signals. Third, we explore the permutation-based feature importance of our classifier to generate hypotheses about possible underlying mechanisms. Last, we discuss the unique challenges involved with predicting activation ratings of elderly PwD. To the best of our knowledge, there are no prior studies that have investigated the prediction of activation ratings of PwD based on verbal and non-verbal signals.

2. RELATED WORKS

Research into the preservation of cognitive resources of PwD has a long history. A number of studies have investigated the effects of activation on perceived well-being, affect, engagement, and other affective states. However, detecting and interpreting the verbal and non-verbal signals of PwD can be particularly challenging due to the broad range of deleterious effects of aphasia or blunted affect on communication (Jones et al., 2015; WHO, 2017). In this section, we will (1) provide an overview of different non-pharmacological interventions that target the activation of PwD and (2) highlight relevant research into the production of (interpretable) verbal and non-verbal signals of PwD.

Over 20 years ago, Olsen et al. (2000) introduced "Media Memory Lane," a system that provides nostalgic music and videos to elicit long term memory stimulation for people with Alzheimer's Disease (AD). An evaluation of this system with 15 day care clients showed positive effects on engagement, affect, activity-related talking, and reduced fidgeting. Astell et al. (2010) evaluated the Computer Interactive Reminiscence and Conversation Aid (CIRCA) system, a touch screen system that presents photographs, music and video clips to enhance the interaction between PwD and caregivers. Their study demonstrated significant differences in verbal and non-verbal behavior when comparing the system with traditional reminiscence therapy sessions. Smith et al. (2009) produced audiovisual biographies based on photographs and personally meaningful music in cooperation with families of PwD. They further used a television set and a DVD player as a familiar interface for their participants. Several studies have also proposed music as a promising factor in non-pharmacological approaches (Spiro, 2010). Accordingly, Riley et al. (2009) introduced a touch screen system that allows PwD to create music regardless of any prior musical knowledge. Evaluating the system in three pilot studies, the authors reported engagement in the activity for all participants. Manera et al. (2015) developed a tablet-based kitchen and cooking simulation for elderly people with mild cognitive impairment. After four weeks of training, most participants rated the experience to be interesting, highly satisfying, and as eliciting more positive than negative emotions. Together, these findings underline the positive effects of non-pharmacological interventions for PwD, as well as for their (in)formal caregivers.

Asplund et al. (1995) investigated affect in the facial expressions of four severe demented participants during activities such as morning care or playing music. The authors compared unstructured judgements of facial expressions with assessments using the Facial Action Coding System [FACS, Ekman et al. (2002)] and showed that while facial cues become sparse and unclear, they are still interpretable to a certain degree. Mograbi et al. (2012) conducted a study with 22 participants with mild to moderate dementia who watched films for emotion elicitation. The authors manually annotated facial expressions, namely happiness, surprise, fear, sadness, disgust, anger, and contempt of the PwD and the controls. While they reported little difference in their production, PwD showed a narrower range of expressions which were less intense. This is in line with other studies that report that PwD may suffer from emotional blunting (Kumfor and Piguet, 2012; Perugia et al., 2020). To examine the quality and the decrease of emotional responses of PwD, Magai et al. (1996) conducted a study with 82 PwD with moderate or severe dementia and their families. Two research assistants were trained to manually code the participants' affective behavior, namely interest, joy, sadness, anger, contempt, fear, disgust, and knit brow expressions. Their results suggest that emotional expressivity, however, may not vary much depending on the stage of the disease.

Another important modality for the recognition of affective states is speech (Schuller, 2018). Nazareth (2019) demonstrated that lexical and acoustic features can be used to predict emotional valence in spontaneous speech of elderly. However, research has shown that speech also undergoes disease-related changes in dementia, e.g. impairments in the production of prosody (Roberts et al., 1996; Horley et al., 2010). This is particularly pertinent in frontotemporal dementia (Budson and Kowall, 2011).

Overall, there seems to be no strong direct link between the ability to produce (interpretable) verbal and non-verbal signals of emotions and the stage of the disease. It rather appears to be a combination of multiple factors such as the dementia type, co-morbidities, medication, and personality. Also, the context seems to play a role. Lee et al. (2017) showed that social and verbal interactions increase positive emotional responses. Notably even the merely implicit presence of a friend has been shown to be sufficient for eliciting this effect in healthy adults (Fridlund, 1991). Thus, emotional expressiveness appears to be extremely sensitive to contextual factors, and PwD might stand to benefit from such factors.

3. DATA COLLECTION

3.1. I-CARE System

The dataset used in this study was collected with the I-CARE system. I-CARE is a tablet-based activation system that is designed to be jointly used by PwD and (in)formal caregivers. The system is mobile and can be used at any location with and internet connection. It provides 346 user-specific activation contents (image galleries, videos, audios, quizzes, games, phrases and texts) on various topics such as gardening, sports, baking, or animals. The system also allows for the uploading of one's own contents to put more emphasis on biographical work (Schultz et al., 2018, 2021). At the same time, it allows for a multimodal

FIGURE 1 | The left figure shows two participants and one instructor from a project partner, who explains the procedure. The right figure shows two participants during an activation session (©AWO Karlsruhe).

data collection using the tablet's camera and microphone to capture video (30 FPS) and audio signals (16 kHz), respectively. The tablet used in the present work was a Google Pixel C (10.2-inch display) or Huawei MediaPad M5 (10.8-inch display). **Figure 1** shows exemplary how an activation session could look like.

3.2. Experimental Setting

The data collection for this study was conducted in different care facilities in Southern Germany as a part of the I-CARE project (Schultz et al., 2018, 2021). Participants of the study were PwD who fulfilled the clinical criteria for dementia according to the ICD-10 system (Alzheimer dementia, vascular dementia, frontotemporal dementia, Korsakoff's syndrome, or Dementia Not Otherwise Specified) ranging from mild to severe, and their (in)formal caregivers. All participants provided written consent and there was no financial compensation. For this study, a setup with minimal supervision and setup requirements was selected with activation sessions taking place in private rooms or in commonly used spaces in the care facilities. The tablet was placed on a stand in front of the participant with dementia so that their face was well-aligned with the field of view of the tablet camera.

At the beginning of each session, the system enquired about the daily well-being ("How are you today?") of the PwD using a smiley rating scale (positive, neutral, negative). Next, the system's recommender system suggested four different activation items, based on interests, personal information of the PwD, and previous ratings. The system also provided the opportunity to search for specific contents and view an activation history. Next, the PwD chose the activation content, e.g. an image gallery on baking, a video on gardening and so on. After each activation, the system asked the PwD for a rating of how well they liked the activation ("Did you enjoy the content?"), again, on a smiley rating scale (positive, neutral, negative). **Figure 2** shows the thumbnail images of four activation recommendations (left) and the rating options after the activation (right). Following the smiley rating, the system went directly back to the overview with recommended activation contents. Here, the PwD could decide whether or not to continue with another activation. Usually, activation sessions consisted of multiple individual activations.

The dataset used in this study consists of 187 activation sessions comprising 804 individual activations and, correspondingly, 804 activation ratings. These sessions cover 25 PwD (gender: 15 f, 10 m; age: 58–95 years, *M*: 82.4 years, *SD*: 9.0 years; dementia stage: 8 mild-moderate, 5 severe, 12 unspecified). Individual participants contributed with different number of sessions ($M = 7.48$, $SD = 2.42$, $Min = 2$, $Max = 12$).

4. METHODS

4.1. Rating Measurement

Self-reported activation ratings of the PwD were collected using an smiley rating scale (positive, neutral, negative) at the end of each activation. **Figure 3** shows the distribution of activation ratings for the participants individually and in total. The colors correspond to the rating (positive = green, neutral = yellow, negative = red). It is evident that activation contents were more frequently perceived as positive than neutral or negative by most participants. A Kruskal-Wallis test shows that these differences are statistically significant ($H = 54.571$, $p < 0.001$). Accordingly, investigating the class distribution across all participants provides a similar picture (positive = 68.23 %, neutral = 25.46 %, negative = 6.3 %). This demonstrates that the activation contents were mostly perceived positively.

4.2. Engagement Analysis

While effective activation contents are typically perceived as positive, not all positive contents are likely to be highly engaging. Furthermore, activation contents will only be effective in the long run if they succeed in engaging PwD. Thus, predicting engagement from verbal and non-verbal signals can be regarded as a separate challenge. As shown by previous work (Steinert et al., 2020, 2021), engagement can indeed be automatically recognized from verbal and non-verbal signals. Engagement in I-CARE was annotated retrospectively based on audio-visual data using the "Video Coding-Incorporating Observed Emotion" (VC-IOE) protocol (Jones et al., 2015) by two independent raters. We computed Cohen's Kappa (κ) between both raters after intensive training on six random test sessions to evaluate inter-rater reliability. The VC-IOE defines different engagement dimensions which were evaluated separately. These are emotional ($\kappa = 0.824$), verbal ($\kappa = 0.783$), visual ($\kappa = 0.887$), behavioral ($\kappa = 0.745$), and agitation ($\kappa = 0.941$) [1]. To obtain the level of engagement for each activation content, we calculated an engagement score by summing up the number of positive engagement outcomes per dimension over all frames of an activation content, divided by the total number of frames covering that activation.

Figure 4 shows the distribution of engagement scores with regards to the self-reported activation ratings of the participants. A Kruskal-Wallis test demonstrated a statistically significant difference ($H = 7.199$, $p < 0.05$) in the group means between the negative ($M = 0.75$, $SD = 0.56$), the neutral ($M = 0.78$, *SD*

[1]The VC-IOE further suggests collective engagement as a dimension which is defined as "Encouraging others to interact with STIMULUS. Introducing STIMULUS to others." (Jones et al., 2015). We interpreted "others" as third persons who did not originally take part in the session. As collective engagement was not apparent in this dataset, we dismissed this dimension.

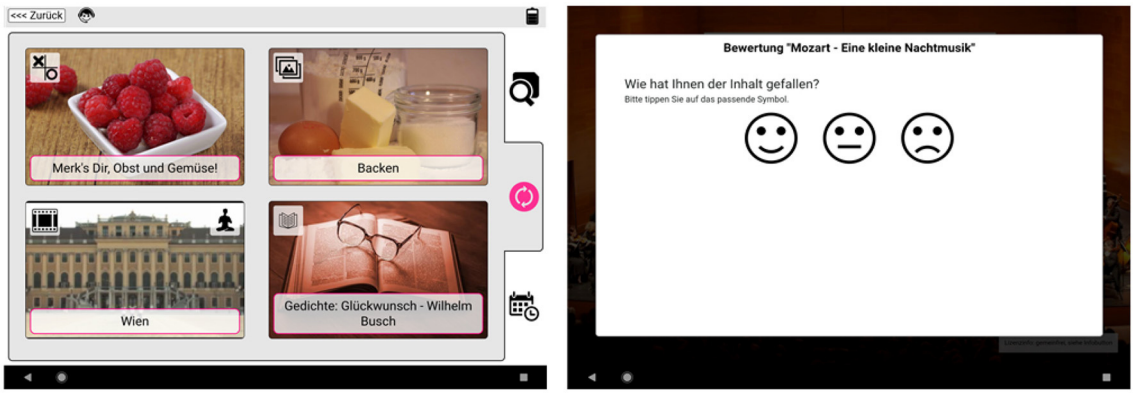

FIGURE 2 | User interface of the I-CARE system. The left figure shows the activation recommendations (top left: memory game, top right: image gallery, bottom left: video, bottom right: phrase). The right figure illustrates the rating options after the activation.

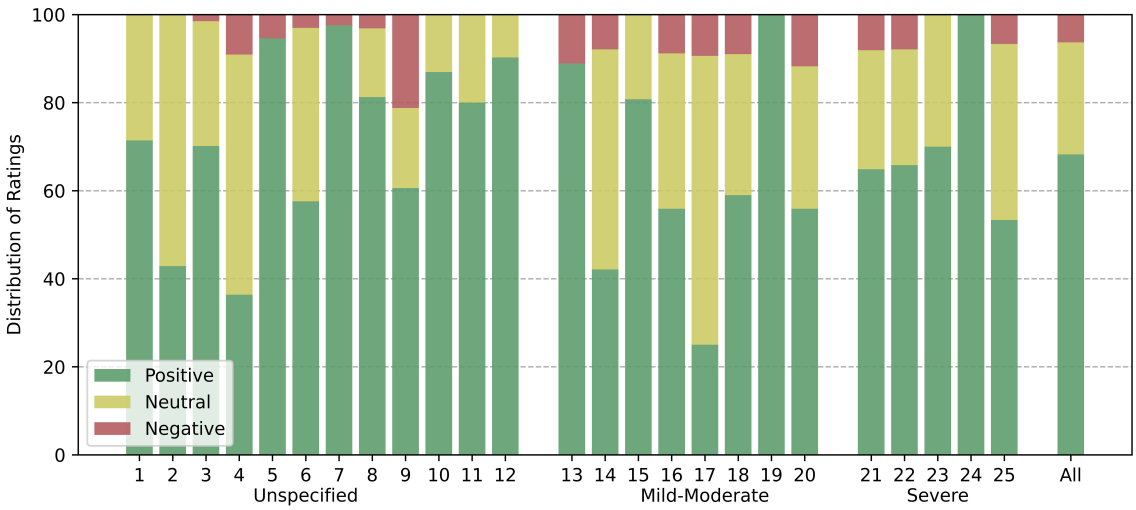

FIGURE 3 | Rating distribution for individual participants **(left)** and in total **(right)** grouped based on their dementia stage. Bar colors correspond to the ratings (positive = green; neutral = yellow; negative = red). Positive ratings significantly outweigh neutral and negative ratings ($H = 54.571$, $p < 0.001$).

$= 0.51$) and the positive class ($M = 0.89$, $SD = 0.47$), indicating a small effect of slightly more evidence for engagement toward positively evaluated activations compared to more negatively perceived contents. Similarly, a Spearman rank correlation analysis ($\rho = 0.094$, $p < 0.001$) showed a significant but small correlation between the engagement score and the rating of individual activation contents.

4.3. Multimodal Features

Human affective behavior and signaling is multimodal by nature. Thus, it can only be fully interpreted by jointly considering information from different modalities (Pantic et al., 2005). We argue that this is especially valid for PwD in an unconstrained care setting because PwD might suffer from aphasia or blunted affect (Kumfor and Piguet, 2012; Perugia et al., 2020). As individual channels begin to degrade, compensation by other channels is well-known to become more important. However, PwD may not only face greater challenges when decoding signals from by their interaction partners (receiver role) - but also with respect to clearly encoding their own socio-emotional signals in any individual channel (sender role). The Signal-to-Noise Ratio (SNR) can also be low for some modalities due to (multiple) background speakers, room reverberation or adverse lighting conditions. Accordingly, we use video-based features (OpenFace, OpenPose, and VGG-FACE) and audio-based (ComParE, DeepSpectrum) features, for the prediction of activation liking of PwD.

4.3.1. Video

The face is arguably the most important non-verbal source for information about another person's affective states (Kappas et al., 2013) and can provide information about affective states throughout all stages of dementia (see section 2). Here, we use the video signal captured with the tablet's camera to detect,

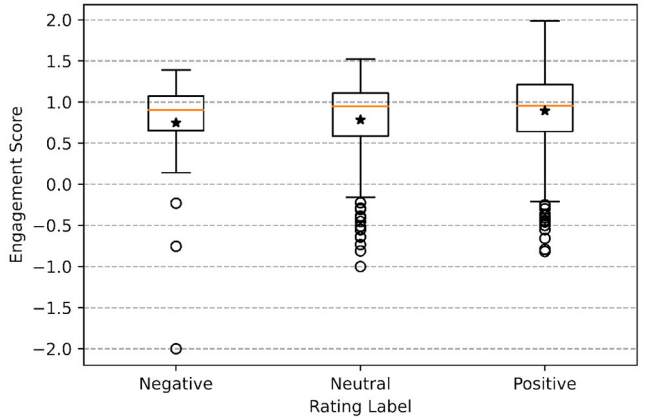

FIGURE 4 | Distribution of engagement scores with regards to the self-reported activation ratings (negative, neutral, positive). There are statistically significant difference ($H = 7.199$, $p < 0.05$) in the group means between the negative ($M = 0.75$, $SD = 0.56$), the neutral ($M = 0.78$, $SD = 0.51$) and the positive class ($M = 0.89$, $SD = 0.47$). Outliers were not removed from further analyses. The * symbol indicates the arithmetic mean.

align, and crop faces from the participants with dementia. From these pre-processed video frames, we extract facial features, namely the (binary scaled) presence of 18 and the (continuously scaled) intensity of 17 Action Units (AUs)[2] ranging from 0 to 5, the location and rotation of the head (head pose), and the direction of eye gaze in world coordinates using OpenFace 2.0 (Baltrusaitis et al., 2018). In the same vein, we extract skeleton features using OpenPose (Cao et al., 2019) to calculate relevant features, namely the distance between shoulders, eyes, ears, hands to nose, and the visibility of the hands. Last, we apply transfer learning using the pre-trained VGG-Face network (Parkhi et al., 2015). We retrained the network for five epochs using the FER2013 dataset with stochastic gradient descent, a learning rate of 0.0001, and a momentum of 0.9. Next, all video frames are rescaled to 224x224 pixels to match the input size of the Convolutional Neural Network (CNN), and normalized by subtracting the mean. The feature vectors for each video frame is the extracted from the *fc6* layer of the network. Overall, concatenating the feature vectors from all feature extractors leads to a 4138-dimensional feature vector for each video frame.

4.3.2. Audio

The recognition of affective states from speech is also a highly active research area (Akçay and Oğuz, 2020). While previous research has shown that speech undergoes disease-related changes in dementia, e.g. impairments in the production of prosody (Roberts et al., 1996; Horley et al., 2010), recent studies suggest that speech of PwD may still help to improve the automatic recognition of engagement (Steinert et al., 2021). We first apply denoising on all raw audio files recorded with the tablet's microphone to remove stationary and non-stationary background sounds, and to enhance participant's speech (Defossez et al., 2020). From the denoised audios, we extract the 2013 Interspeech Computational Paralinguistics Challenge features set (ComParE) using OpenSMILE (Eyben et al., 2010, 2013). We extract audio frame-wise (60 ms frame size; 10 ms steps) frequency, energy, and spectral related Low-Level Descriptors (LLD) which leads to a 130-dimensional feature vector (65 LLDs + deltas) for each step of 10 ms. Next, we create mel spectrograms using Hanning windows (512 samples size, 256 samples steps). We forward spectrograms (227x227 pixels, viridis colormap) to the pre-trained CNN AlexNet to receive bottleneck features from the *fc7* layer which results in a 4096-dimensional feature vector (Amiriparian et al., 2017).

4.4. Data Pre-processing

To take interpersonal and intrapersonal variations into account, we scale each feature to a range between zero and one. We assume that the verbal and non-verbal signals from the time interval shortly before the rating are likely to be most diagnostic for the subsequent activation rating. Correspondingly, we consider the 30 s of verbal and non-verbal signals before the rating was provided. Next, we slice features into 1 s segments with 25 % overlap and assign each segment to the corresponding rating label. Due to the class imbalance (see **Figure 3**), we combine the neutral and negative classes to formulate a two-class prediction problem. This seems reasonable as especially the prediction of positively perceived activation contents is relevant for an individual's well-being and motivation (Cohen-Mansfield, 2018). These pre-processed and labeled feature sequences are then forwarded to the classifier.

4.5. Prediction and Evaluation

The applied prediction approach is based on Long-Short-Term-Memory (LSTM) networks which allow for the preservation of temporal dependencies. This is especially important as verbal and non-verbal signals such as speech or facial expressions are subject to continuous change, especially in interactive activation sessions. Due to the different sampling rates of the feature sets of video and audio features (ComParE and DeepSpectrum), the classifier consists of three different input branches. Each input branch consists of a CNN layer (filter size = 256, 64, 256) followed by a MaxPooling layer (pool size = 3, 5, 3). Next, outputs are forwarded to an LSTM layer (units = 512, 64, 512). The three resulting context vectors are concatenated and passed to a Dense layer (units = 256) followed by the output layer (units = 2) with a Softmax activation function which outputs the class prediction. **Figure 5** shows the proposed system architecture. For regularization, we use a dropout rate of 0.3 in the LSTM layers and after the concatenation layer. We train the model for 50 epochs with a batch size of 16. We use a cross-entropy loss function and Adam optimizer with a learning rate of 0.001. To retrieve the overall rating prediction from individual segments, we apply majority voting. We apply a session-independent model evaluation through 10-fold cross-validation on session level where individual folds contain multiple sessions (18–19) and, thus, multiple activation ratings (67–87) ranging from negative

[2]AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU45. For AU28, OpenFace only provides information about whether the AU is present.

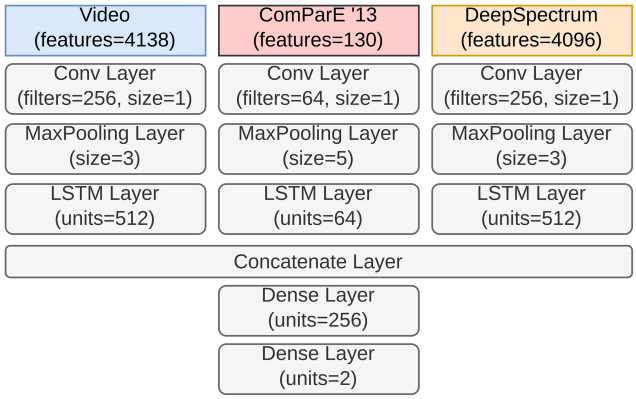

FIGURE 5 | Overview of our proposed prediction model.

to positive. Based on this approach, the proposed system learns behavioral characteristics elicited through subjective activation likings of multiple participants for inference on unseen sessions. The performance of our approach is compare to chance level. We select Unweighted Average Precision, Recall and F1-Score as the evaluation metrics as they are particularly suitable for unevenly distributed classes. To test for statistical significance between our model and the baseline, i.e. chance level, we apply a McNemar Test.

4.6. Permutation-Based Feature Importance

Explainable artificial intelligence has become an important research field in recent years (Linardatos et al., 2021). Knowing about the underlying mechanisms behind the predictions of black-box classifiers such as neural networks helps to understand and interpret their output. Accordingly, we compute permutation-based feature importances to investigate the importance of individual features for the prediction results (Molnar, 2020). For this, we break the association between individual features and labels by shuffling each feature sequence and adding random noise. For particularly relevant features, this should increase the model's prediction error, i.e. the cross-entropy loss (Kuhn and Johnson, 2013; Molnar, 2020). This is especially useful because it (1) provides insights into which verbal and non-verbal signals are relevant for the prediction of activation rating/ liking of PwD and allows for comparison with healthy individuals, and (2) it can help reveal irrelevant features, which can then be removed to decrease model complexity and computational costs.

5. RESULTS AND DISCUSSION

Table 1 shows the prediction results as the *M* and *SD*, Precision, Recall and F1-Score for each class individually and as an unweighted average over all folds. It is apparent that the model is especially capable of correctly predicting the positive class. A possible explanation for this may be the imbalance toward this

TABLE 1 | Prediction results based on the session-independent 10-fold cross-validation on session level.

Class	Precision	Recall	F1-Score
Pos.	0.726 (0.096)	0.754 (0.209)	0.729 (0.127)
Neu./ Neg.	0.308 (0.224)	0.364 (0.277)	0.328 (0.238)
Unweighted avg.	0.517 (0.272)	0.559 (0.312)	0.528 (0.277)
Chance	0.342 (0.354)	0.500 (0.513)	0.405 (0.417)

Results are reported as the M and SD Precision, Recall and F1 Score for each class individually and as the unweighted average over all folds.

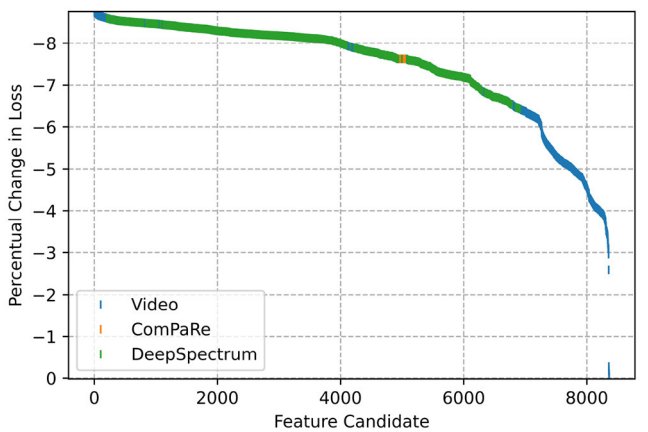

FIGURE 6 | Permutation-based feature importance averaged over all folds. The y-axis represents the perceptual change when comparing the cross-entropy loss before and after permutation, the x-axis shows the feature candidates. The colors indicates the set the feature belongs to (Video = blue, ComParE = Orange, DeepSpectrum = green).

class (see **Figure 3**). The model might not have seen a sufficient variation of data to accurately predict neutral and negative activation ratings. We also assume that participants showed only rather subtle negative expressions due to the highly supportive social context (Lee et al., 2017).

What stands out is that overall the prediction model significantly ($\chi^2 = 4.91$, $p < 0.05$) outperforms the baseline. Accordingly, verbal and non-verbal signals of PwD in different stages of the disease contain sufficient information for the prediction of activation ratings - despite the challenging recording conditions. The standard deviation indicates performance fluctuations throughout the folds. There are several possible explanations for this result. Participants in our study contributed substantially different numbers of sessions and, thus, different numbers of training samples (see section 3.2). As individual folds do not necessarily represent the overall data distribution, predictions can be based on a variable number of training samples of the same participant. The unstable recording conditions (background speakers, room reverberation, or lighting) throughout individual sessions might further increase the heterogeneity within folds. At the same time, this seems inevitable as the I-CARE system is designed for mobile usage.

Thus, these results are not comparable to clean and unambiguous data obtained in laboratory studies with healthy individuals.

Figure 6 provides an overview of the permutation-based feature importance averaged over all folds. The y-axis indicates the percentage change when comparing the cross-entropy loss before and after permutation. The bigger the negative change, the more important we consider the feature to be. This x-axis represents all 8364 feature candidates (see section 4.3). It is apparent that video-based and DeepSpectrum features seem to be important for the prediction. Especially video-based have been found as an import predictor in other tasks, namely the investigation of music (Tkalčič et al., 2019) or image (Masip et al., 2014) preferences. The curve progression further suggests that there are no individual features that stand out. Instead, it is rather the combination of different features on which the model relies. This finding could also be due to colinearity in the features, i.e. if one feature is permuted, the model relies on a highly correlated neighbor.

6. CONCLUSION

The main goal of the current study was to determine if activation ratings of PwD can be predicted in a real-life environment. We investigated a dataset collected with the I-CARE system of 25 PwD throughout all stages of the disease, and showed that contents provided by the system are mainly perceived positively, which can lead to more engagement and positive mood (Cohen-Mansfield, 2018). Moreover, participants' verbal and non-verbal signals contain sufficient information to successfully predict their activation ratings. Also, we could show that, in line with studies on healthy individuals (Masip et al., 2014; Tkalčič et al., 2019), the face remains an important source of information for inferring preferences. Interestingly, in our sample, there seems to be only a weak link between observed engagement and subjective activation liking. In general, this finding is indeed more consistent with prior reviews and meta-analyses focused on healthy adults, which have demonstrated only weak to moderate associations between subjective experience and different types of physiological or behavioral responses to emotion-eliciting stimuli in healthy adults (Mauss and Robinson, 2009; Hollenstein and Lanteigne, 2014). However, it is remarkable that (1) this relationship appears to be even further degraded among PwD and (2) that machine learning approaches based on multimodal data may still succeed in successfully predicting subjective ratings of PwD. At the same time, our approach still faces a number of limitations. A session-independent model evaluation implies the existence of annotated samples of the participants. While user-independent modeling would be preferable for the real-world application, this seems too ambitious with a small and heterogeneous dataset. As the presented results are not easily comparable to other studies, future work could also consider the assessments of the present caregivers. This could provide further information about the validity of our results. Despite these limitations, the present results make an important contribution to a, thus far, sparsely populated part of the field with regards to predicting activation liking of PwD.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available as the used dataset consists of data of People with Dementia. Requests to access the datasets should be directed to lars.steinert@uni-bremen.de.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University Of Bremen. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

LS conceived and designed the analyses, performed the analyses, and wrote the paper. FP conceived and designed the analyses, collected the data, and wrote the paper. DK conceived and designed the analyses and wrote the paper. TS conceived and designed the analyses, collected the data, and supervision of project. All authors contributed to the article and approved the submitted version.

FUNDING

This work was partially funded by the Klaus-Tschira-Stiftung. Data collection and development of the I-CARE system was funded by the BMBF under reference BMBF-number V4PIDO62. We also gratefully acknowledge the support of the Leibniz ScienceCampus Bremen Digital Public Health (lsc-diph.de), which is jointly funded by the Leibniz Association (W4/2018), the Federal State of Bremen and the Leibniz Institute for Prevention Research and Epidemiology—BIPS.

REFERENCES

Akçay, M. B., and Oğuz, K. (2020). Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* 116, 56–76. doi: 10.1016/j.specom.2019.12.001

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., et al. (2017). "Snore sound classification using image-based deep spectrum features," in *Interspeech 2017* (Stockholm), 3512–3516. doi: 10.21437/Interspeech.2017-434

Asplund, K., Jansson, L., and Norberg, A. (1995). Facial expressions of patients with dementia: A comparison of two methods of interpretation. *Int. Psychogeriatr.* 7, 527–534. doi: 10.1017/S1041610295002262

Astell, A. J., Ellis, M. P., Bernardi, L., Alm, N., Dye, R., Gowans, G., et al. (2010). Using a touch screen computer to support relationships between people with dementia and caregivers. *Interact. Comput.* 22, 267–275. doi: 10.1016/j.intcom.2010.03.003

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). "Openface 2.0: facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (Xi'an: IEEE), 59–66. doi: 10.1109/FG.2018.00019

Budson, A. E., and Kowall, N. W. (2011). *The Handbook of Alzheimer's Disease and Other Dementias*, Vol. 7. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9781444344110

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. (2019). Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 172–186. doi: 10.1109/TPAMI.2019.2929257

Clare, L., and Woods, R. T. (2004). Cognitive training and cognitive rehabilitation for people with early-stage Alzheimer's disease: a review. *Neuropsychol. Rehabil.* 14, 385–401. doi: 10.1080/09602010443000074

Cohen-Mansfield, J. (2018). Do reports on personal preferences of persons with dementia predict their responses to group activities? *Dement. Geriatr. Cogn. Disord.* 46, 100–108. doi: 10.1159/000491746

Cohen-Mansfield, J., Dakheel-Ali, M., and Marx, M. S. (2009). Engagement in persons with dementia: the concept and its measurement. *Am. J. Geriatr. Psychiatry* 17, 299–307. doi: 10.1097/JGP.0b013e31818f3a52

Cohen-Mansfield, J., Marx, M. S., Thein, K., and Dakheel-Ali, M. (2011). The impact of stimuli on affect in persons with dementia. *J. Clin. Psychiatry* 72:480. doi: 10.4088/JCP.09m05694oli

Defossez, A., Synnaeve, G., and Adi, Y. (2020). "Real time speech enhancement in the waveform domain," in *Interspeech* (Shanghai). doi: 10.21437/Interspeech.2020-2409

Ekman, P., Friesen, W. V., and Hager, J. C. (2002). *Facial Action Coding System (FACS), 2nd Edn*. Salt Lake City, UT: Research Nexus Division of Network Information Research Corporation.

Eyben, F., Weninger, F., Groß, F., and Schuller, B. (2013). "Recent developments in opensmile, the Munich open-source multimedia feature extractor," in *MM '13: Proceedings of the 21st ACM International Conference on Multimedia* (Barcelona). doi: 10.1145/2502081.2502224

Eyben, F., Wöllmer, M., and Schuller, B. (2010). "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia, MM '10* (New York, NY: Association for Computing Machinery), 1459–1462. doi: 10.1145/1873951.1874246

Fridlund, A. J. (1991). Sociality of solitary smiling: potentiation by an implicit audience. *J. Pers. Soc. Psychol.* 60, 229–240. doi: 10.1037/0022-3514.60.2.229

Hollenstein, T., and Lanteigne, D. (2014). Models and methods of emotional concordance. *Biol. Psychol.* 98, 1–5. doi: 10.1016/j.biopsycho.2013.12.012

Horley, K., Reid, A., and Burnham, D. (2010). Emotional prosody perception and production in dementia of the Alzheimer's type. *J. Speech Lang. Hear. Res.* 53, 1132–1146. doi: 10.1044/1092-4388(2010/09-0030)

Jones, C., Sung, B., and Moyle, W. (2015). Assessing engagement in people with dementia: a new approach to assessment using video analysis. *Arch. Psychiatr. Nurs.* 29, 377–382. doi: 10.1016/j.apnu.2015.06.019

Kappas, A., Krumhuber, E., and Küster, D. (2013). "Facial behavior," in *Nonverbal Communication*, eds J. A. Hall and M. L. Knapp (Berlin: Mouton de Gruyter), 131–166. doi: 10.1515/9783110238150.131

Kuhn, M., and Johnson, K. (2013). *Applied Predictive Modeling*, Vol. 26. New York, NY: Springer. doi: 10.1007/978-1-4614-6849-3

Kumfor, F., and Piguet, O. (2012). Disturbance of emotion processing in frontotemporal dementia: a synthesis of cognitive and neuroimaging findings. *Neuropsychol. Rev.* 22, 280–297. doi: 10.1007/s11065-012-9201-6

Lee, K. H., Boltz, M., Lee, H., and Algase, D. L. (2017). Does social interaction matter psychological well-being in persons with dementia? *Am. J. Alzheimers Dis. Other Dement.* 32, 207–212. doi: 10.1177/1533317517704301

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable AI: a review of machine learning interpretability methods. *Entropy* 23:18. doi: 10.3390/e23010018

Magai, C., Cohen, C., Gomberg, D., Malatesta, C., and Culver, C. (1996). Emotional expression during mid- to late-stage dementia. *Int. Psychogeriatr.* 8, 383–395. doi: 10.1017/S104161029600275X

Manera, V., Petit, P.-D., Derreumaux, A., Orvieto, I., Romagnoli, M., Lyttle, G., et al. (2015). "Kitchen and cooking," a serious game for mild cognitive impairment and Alzheimer's disease: a pilot study. *Front. Aging Neurosci.* 7:24. doi: 10.3389/fnagi.2015.00024

Masip, D., North, M. S., Todorov, A., and Osherson, D. N. (2014). Automated prediction of preferences using facial expressions. *PLoS ONE* 9:e87434. doi: 10.1371/journal.pone.0087434

Mauss, I. B., and Robinson, M. D. (2009). Measures of emotion: a review. *Cogn. Emot.* 23, 209–237. doi: 10.1080/02699930802204677

Mograbi, D. C., Brown, R. G., and Morris, R. G. (2012). Emotional reactivity to film material in Alzheimer's disease. *Dement. Geriatr. Cogn. Disord.* 34, 351–359. doi: 10.1159/000343930

Molnar, C. (2020). *Interpretable Machine Learning*. Morrisville: lulu.com.

Nazareth, D. S. (2019). "Emotion recognition in dementia: advancing technology for multimodal analysis of emotion expression in everyday life," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (Cambridge), 45–49. doi: 10.1109/ACIIW.2019.8925059

Olsen, R. V., Hutchings, B. L., and Ehrenkrantz, E. (2000). "Media memory lane" interventions in an Alzheimer's day care center. *Am. J. Alzheimers Dis.* 15, 163–175. doi: 10.1177/153331750001500307

Pantic, M., Sebe, N., Cohn, J. F., and Huang, T. (2005). "Affective multimodal human-computer interaction," in *Proceedings of the 13th Annual ACM International Conference on Multimedia, MULTIMEDIA '05* (Singapore: Association for Computing Machinery), 669–676. doi: 10.1145/1101149.1101299

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). "Deep face recognition," in *British Machine Vision Conference* (Swansea). doi: 10.5244/C.29.41

Perugia, G., Diaz-Boladeras, M., Catala, A., Barakova, E. I., and Rauterberg, M. (2020). ENGAGE-DEM: a model of engagement of people with dementia. *IEEE Trans. Affect. Comput.* 1. doi: 10.1109/TAFFC.2020.2980275

Riley, P., Alm, N., and Newell, A. (2009). An interactive tool to promote musical creativity in people with dementia. *Comput. Hum. Behav.* 25, 599–608. doi: 10.1016/j.chb.2008.08.014

Roberts, V. J., Ingram, S. M., Lamar, M., and Green, R. C. (1996). Prosody impairment and associated affective and behavioral disturbances in Alzheimer's disease. *Neurology* 47, 1482–1488. doi: 10.1212/WNL.47.6.1482

Schreiner, A. S., Yamamoto, E., and Shiotani, H. (2005). Positive affect among nursing home residents with Alzheimer's dementia: the effect of recreational activity. *Aging Mental Health* 9, 129–134. doi: 10.1080/13607860412331336841

Schuller, B. W. (2018). Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* 61, 90–99. doi: 10.1145/3129340

Schultz, T., Putze, F., Schulze, T., Steinert, L., Mikut, R., Doneit, W., et al. (2018). *I-CARE - Ein Mensch-Technik Interaktionssystem zur Individuellen Aktivierung von Menschen mit Demenz* (Oldenburg).

Schultz, T., Putze, F., Steinert, L., Mikut, R., Depner, A., Kruse, A., et al. (2021). I-CARE-an interaction system for the individual activation of people with dementia. *Geriatrics* 6:51. doi: 10.3390/geriatrics6020051

Smith, K. L., Crete-Nishihata, M., Damianakis, T., Baecker, R. M., and Marziali, E. (2009). Multimedia biographies: a reminiscence and social stimulus tool for persons with cognitive impairment. *J. Technol. Hum. Serv.* 27, 287–306. doi: 10.1080/15228830903329831

Spector, A., Thorgrimsen, L., Woods, B., Royan, L., Davies, S., Butterworth, M., et al. (2003). Efficacy of an evidence-based cognitive stimulation therapy programme for people with dementia: randomised controlled trial. *Brit. J. Psychiatry* 183, 248–254. doi: 10.1192/bjp.183.3.248

Spiro, N. (2010). Music and dementia: observing effects and searching for underlying theories. *Aging Ment. Health* 14, :891–899. doi: 10.1080/13607863.2010.519328

Steinert, L., Putze, F., Küster, D., and Schultz, T. (2020). "Towards engagement recognition of people with dementia in care settings," in *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event), 558–565. doi: 10.1145/3382507.3418856

Steinert, L., Putze, F., Kuster, D., and Schultz, T. (2021). "Audio-visual recognition of emotional engagement of people with dementia," in *Proc. Interspeech 2021* (Brno), 1024–1028. doi: 10.21437/Interspeech.2021-567

Tkalčič, M., Maleki, N., Pesek, M., Elahi, M., Ricci, F., and Marolt, M. (2019). "Prediction of music pairwise preferences from facial expressions," in *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19* (New York, NY: Association for Computing Machinery), 150–159. doi: 10.1145/3301275.3302266

WHO (2017). *Dementia*. Available online at: https://www.who.int/news-room/fact-sheets/detail/dementia (accessed August 5, 2021).

Woods, B., Aguirre, E., Spector, A. E., and Orrell, M. (2012). Cognitive stimulation to improve cognitive functioning in people with dementia. *Cochrane Database Syst. Rev*. 2:CD005562. doi: 10.1002/14651858.CD005562.pub2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.