IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 July 2024, accepted 30 July 2024, date of publication 12 August 2024, date of current version 23 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3442377

RESEARCH ARTICLE

Intermediate Layer Attention Mechanism for Multimodal Fusion in Personality and Affect Computing

P. SREEVIDYA[1], J. ARAVINTH[1], (Member, IEEE), AND SATHISHKUMAR SAMIAPPAN[2]

[1]Department of Electronics and Communication Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore 641112, India
[2]Geosystems Research Institute, Mississippi State University, Starkville, MS 39762, USA

Corresponding author: J. Aravinth (j_aravinth@cb.amrita.edu)

ABSTRACT This article introduces a versatile multimodal architecture designed for personality-aware systems, encompassing tasks such as personality trait prediction, sentiment analysis, and emotion recognition. This is a unique attempt to develop a general pipeline that is applicable to the personality affect computing applications within the context of multimodal data. The proposed model employs task-specific feature extraction models that are appropriately trained for each application. An intermediate layer, employing both inter- and intra-attention mechanisms for fusion, is presented. This dual attention mechanism is further improved with a binary search algorithm, which is notably the key contribution of the work. This fusion models discerns distinctive features crucial for classification and regression tasks. To evaluate the system's efficacy, short-duration video clips and corresponding transcriptions from databases were utilized. Low-level acoustic features were derived from audio signals, while high-level and mid-level audio features were extracted through a transformer-based sentence-RoBERTa model applied to audio transcripts. Visual features were obtained from context and facial images through deep face networks, followed by the use of CNN and LSTM models. Dimensionality reduction and multimodal fusion techniques were implemented prior to applying machine learning-based classification and prediction tasks. Performance metrics such as mean accuracy and squared correlation coefficients (R^2) were chosen for prediction tasks, while accuracy and F1-score were employed for classification tasks. The study explored various fusion techniques and dimension-reduction approaches to establish an efficient pipeline, ultimately aiming to reduce uncertainties and enhance robustness. The results indicate that the proposed architecture performs comparably with state-of-the-art systems across all evaluated domains.

INDEX TERMS Big-five personality traits, emotion recognition, fusion techniques, attention.

I. INTRODUCTION

Personality awareness refers to the understanding, recognition, and comprehension of an individual's personality traits, emotional states, behavioural patterns, preferences and sentimental connections [1]. It involves the ability to perceive and comprehend the unique psychological, cognitive and affective aspects that define an individual's characteristics, reactions and responses along with disposition in various situations [2], [3]. Personality awareness has growing applications in education &learning, job interviews, healthcare &medicine, recommender systems, content creation for social medial platforms, adaptive & emotion aware human-machine interaction systems and visual question answering [4] to name a few. Human personality is graded based on a detectable and distinctive collection of emotions and adaptive patterns, and personality traits are the means by which this scale is explained. The Myers-Briggs Type Indicator (MBTI) model, the Big-Five Personality (BFI) Model [5] and shorter scales such as the NEO-FFI or BFI

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

or Ten Item Personality Inventory (TIPI) [6] are some of the tools for computing personality traits. The OCEAN model or BFI model is a discrete form of traits classification which divides the personality traits into 5 categories with certain specialities:-

1) Openness: very creative & adventurous people, capable of tackle changes
2) Conscientiousness: Paying attention to details, schedule things and finishes task duty bound
3) Agreeableness: Cooperative empathetic, helping and caring
4) Extroversion: Enjoying being the center of attraction and out going and
5) Neuroticism: Mood swings, gets upset easily and anxious.

To understand and evaluate the emotion or sentiment, psychologists put forth two typical theories to model human emotion: discrete emotion model or Ekman model and dimensional emotion model which takes emotions as bipolar entities. The emotion recognition aims to detect the emotional state of human beings and mostly focuses on visual emotion recognition (VER), audio/speech emotion recognition (AER/SER), and physiological emotion recognition (PER). The sentiment analysis is typically classifying the human into positive, negative, or neutral categories [7]. The conventional approaches like psychometric charts and clinical diagnosis have paved ways to non-invasive techniques based on artificial intelligence for computing the various attributes relating to personality awareness. Normally, data from text, image and video signals contain huge information about personality awareness. Deep learning and machine learning techniques have revolutionized the field of speech processing, computer vision, and text processing. When it comes to the applications such as deception detection, stress detection or advertisement recommendation, information extracted from a single modality will not be sufficient. The effective utilization of disparate datastream mainly taken from social media platforms and physiological signals such as EEG, ECG etc involving different modalities can be materialized through multimodal fusion strategies [8], [9]. Affective and personality computing being an integral part of Human computer interaction systems have benefited, greatly from combining the information from different modalities, rather than taking the features from a single modality [10].

The feature selection from this expanded data horizon is a great challenge. The conventional handcrafted feature extraction methods are getting replaced by deep feature extraction strategies. It is proved that there pre-trained deep neural networks such as VGGNet, ResNet, can be used for extracting embeddings from different modalities, and identifying suitable features from these embeddings play a pivotal role in predicting the personality awareness [11]. The next hurdle is to identify appropriate fusion strategies for combining the features from different modalities so as to improve the performance of the existing systems. The major factors regarding multimodal approaches are:-

1) The modalities can be heterogeneous in terms of structures, qualities and representations.
2) Modalities are connected, since they share commonalities.
3) While the modalities are interacting with each other, they may give no information.

Studies in psychology shows that emotional dispositions and personality are closely related to each other. From analyzing the advancements in the fields of affect and personality computing, it can be observed that social media based information retrieval has high impact on the classification and recognition based tasks. Apart from that, the systems requirements for improved accuracy has teamed up with multi modal fusion in order to extract more information.

This inspiration led us to conceptualize a comprehensive pipeline for the mentioned tasks. Additionally, we recognize the necessity for customization based on specific applications. Consequently, our work puts forth the following contributions.:-

- To propose a multi modal fusion technique employing inter and intra-modal attention mechanisms which takes embeddings from intermediate layers.
- To compare the performance of different fusion strategies such as early fusion, intermittent fusion, weighted intermittent fusion and attention mechanisms for combining the features from different modalities.
- To study the impact and importance of feature reduction before undertaking the classification/regression tasks.
- Finally, to propose a comprehensive pipeline for personality and affect computation tasks.

The feature selection plays a pivotal role in the classification/regression tasks. The proposed model was using features from visual and speech modalities. The task was to select the best features from the selected modalities which was done through the proposed cross modal approach that included a combination of intra and inter modality fusion. This along with dimensional reduction techniques was identified as the simplified architecture for the prediction purposes.

The subsequent parts of the document is organized as follows: Section II covers the research on multimodal architectures and the feature extraction techniques put out in the most recent past. The architecture of the suggested deep multimodal attention-based system and the modules that go with it are provided in Section III. The dataset and the implementation details of the tests conducted with different configurations are provided in Section IV. In Section IV, we also go over the outcomes acquired under five different setups. The ablation study in Section V aims to determine how dimension reduction techniques affect various classification techniques. Lastly, we address the work's potential scope and constraints in Section VI.

II. RELATED WORKS

This section is a brief survey of existing works on estimating personality. In the beginning, personality prediction was mainly based on written scripts. The Linguistic factors were

analyzed to recognize the underlying personality. Linguistic Inquiry Word Count (LIWC) is an analytic tool that can analyze quantitative text based on a psychological dictionary [12]. The works on personality have got a new dimension with the advent of new AI techniques. Image, Audio, video, and text modalities were explored to develop models to predict personality. Term Frequency-Inverse Document Frequency (TF-IDF), Glove, Word2Vec are some of the popular methods for text feature extraction [13]. Of late, Bidirectional Encoder Representations from Transformers (BERT), and RoBerta XLXLNet are increasingly used for getting the feature embeddings [14]. The work by Sreevidya et al. [15] on the portrait personality dataset proposed that the FaceNet and its derivatives, such ArcNet, can be used to categorize personality traits because of their discriminating abilities. Transfer learning methods can also be used to tailor face recognition algorithms. An ontology based multimodal fusion of text and images was tried by Biswas et al. [16]. In the work of Suman et al. [17], a multimodal personality trait prediction system was developed. The deep features were developed for different modalities. They achieved an average accuracy of 90.43% on ChaLearn First Impressions dataset. ResNet-based networks extract facial and ambient features from the visual modality. The audio features are extracted using the VGGish Convolutional Neural Networks. Textual features include BERT and Glove embeddings. They experimented with traditional early and late fusion strategies and attentive mechanisms. The experiments showed the advantages of early fusion, which we adopt in the proposed framework. Güçlütürk et al. [18] proposed a deep video visual framework for multimodal personality recognition with a Long-Short-Term-Memory (LSTM) network and volumetric Convolutional Neural Network. Personality traits are associated with human behavior, response to contexts, the pattern of thoughts, and approaches to handling emotions. An end-to-end AI-based automatic personality recognition system was developed based on video features by Suen et al. [19]. The work focuses on Asynchronous Video Interviews embedded with a Tensorflow-based semi-supervised deep learning network, applied to accurately auto-recognize the interviewer's true personality. Explainability and interpretability were brought into the context of apparent personality recognition by Escalante et al. [20]. Here baseline models were developed on the text and sensory modalities. The predictions were explained using visualization or using an audio/visual occlusion method, which marks the decision-sensitive region or feature identification. Uncertainty modeling is an extension to the supervised learning techniques for personality prediction [21]. The authors give a holistic perspective of uncertainties in video-based works. They were trying to quantify epistemic and aleatoric segments of emotion estimation from emotion-based datasets. Further, these emotion predictions are associated with personality prediction pipelined as a downstream task. CNN-GRU network was used for the estimation task. Monte-Carlo dropouts [22] and predictive modeling techniques [23] were used to address the uncertainties. Predicting the Big five personality traits based on emotional features such as Arousal, valence, and likability was the objective of the work by Sogancioglu et al. [24]. An explainable boosting machine (EBM) regressor was employed for trait prediction using the mood and likability scores. [25].

In work by Aslan and his team [26], the personality traits were estimated by integrating a consistency constraint along with the trait-specific errors in the cost function. For each modality, pre-trained networks viz, ResNet, VGGish, and ELMo were used to extract the features. A multilayered Long Short-Term Memory network was finally used to capture the temporal dynamics that eventually improved the performance of the system. They proposed two-stage modeling, where a modality-specific sub-networks is first trained individually, and then fine-tuned multimodal data to accomplish the target jointly. It is pointed out by Giritlioğlu [27] that induced behavior includes personality traits by applying multimodal and deep learning-based algorithms. Long- and Short-term Time-series Networks (LSTNet) and Recurrent Convolutional Neural Networks (RCNN) were the main networks chosen, and the experiments were conducted on Self-presentation and Induced Behavior Archive for Personality Analysis (SIAP) and First Impressions datasets. The LSTNet and RCNN combine the benefits of Long Short-Term Memory (LSTM) with CNN. The plethora of features from different modalities is a salient feature in video-based datasets. Tharini et al. [28] applied a hierarchical fusion for sentiment analysis with autoencoder based model. Meanwhile, the work on emotion recognition by combining four different modalities was performed in the work by Priyadarshini et al. [29]. In the work the authors are extracting the physiological features from EEG, ECG, temperature and respiration from a GRU-LSTM network which is then combined using simple fusion techniques. Zhu et al. worked on emotion recognition using the state of the art datasets and they introduced the input as low ranked tensor with a multimodal attention level fusion [30]. Also, a multitasking frame work was recently tried by Akhtar et al. [31] for classification sentiment and emotion datasets employing contextual attention mechanisms.

The challenge is to extract the relevant features from it. As the dimensions of the dataset increases, more complex networks, which are memory hungry and computationally rich, would become essential for predicting personality. With this in mind, the authors decided to develop a simple and efficient framework for predicting the personality by applying multimodal fusion along with dimensionality reduction techniques and state-of-the-art regression algorithms such as SVR, Radial Basis Regressor and GPRs. While numerous studies focus on the classification and regression of emotions, sentiments, and personality traits, earlier research did not attempt to develop a general framework for personality awareness tasks. Limited attention is also given to minimizing

the computational bottleneck that results from the high dimensional data from multimodal fusion.

A. THE MOTIVATION

The multimedia inputs itself can support in identifying the affect computing aspects without relaying on physiological signals. Figure 1 represents the different images from a video of Chalearn V2 dataset which is annotated for personality traits. Here the actor is uttering the sentence "[''Whatever you say, but I did not … I know you can still get like three years afterwards, so knock on wood. I still don't, but they feel really natural. They're very squishy.'']."'. There are some pause and expressions in between and the way in which the actor communicates could act as the input features which can very well coined out by MFCC coefficients and NLP tools. The paraphernalia that consists of emphasis on different words, the face expressions, the shape of the face, the text formations all can be well thought-out as the identifiers of the personality. The said features are surfaced from different modalities. The comprehend fusion of these modalities are critical and therefore analyzing the performance of the proposed model. Additionally, it can be observed that all the personality awareness systems have commonalities and they differentiate in frequencies. The emotion system is a central subsystem of personality, and that inter-individual differences traceable to this system are important for describing individuals [32]. This key ins from psychology have paved ways for us to propose a general pipeline for personality awareness problems. The precise selection of the features and discriminate multimodal fusion techniques with reduced dimensions are giving the results at par with the state-of-the-art systems.

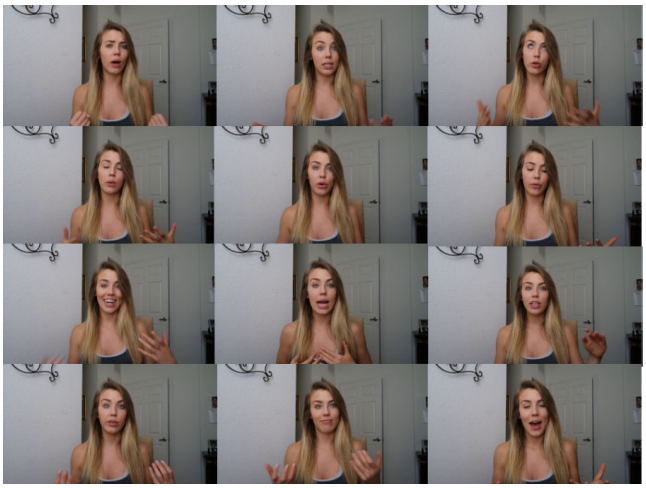


FIGURE 1. Video images indicating neuroticism.

III. METHODOLOGY

A general block diagram for the proposed system which can be applied to personality and affect computations is given in Figure 2. Here, Deep features are extracted through transfer learning based CNN and LSTM networks. The features are taken from multiple modalities including audio, video, and text. Suitable fusion strategies are chalked out for feature fusion, and dimensioanlity reduction techniques were applied for reducing the feature complexity. These embeddings are applied to machine learning based classification/regression tasks.

We have considered Chalearn First Impression V2 Dataset predicting the personality traits, while ElderReact dataset was used for classification of discrete emotions. The MOUD dataset was selected for binary classification of positive and negative sentiments. We applied the proposed general approach on each task.

A. TECHNIQUES FOR EXTRACTING DEEP FEATURES

1) FOR PERSONALITY COMPUTATION

a: AUDIO FEATURES

We used both Low-Level Descriptors (LLD) and High-Level Descriptors (HLD) for extracting the audio features. The LLD consists of energy information, auditory spectra characteristics, Mel Frequency Cepstral Coefficients (MFCC), pitch, auto-correlation coefficients, Perceptual Linear Predictive (PLP) Coefficients, Linear Predictive Coefficients (LPC), etc., which were extracted by using the openSMILE tool. The openSMILE tool was used with a standard feature configuration that served as a challenge baseline since the INTERSPEECH 2013 Computational Paralinguistics Challenge [33], [34]. The above-said configuration is identified as the most suitable one for automatic personality recognition tasks. The feature set extracted has a dimension of 6373D.

The speech in the video files were converted to transcriptions by the transcription service, the professional transcribers. In total, 435984 words were transcribed (183861 non-stop words), corresponding to 43 words per video on average (18 non-stopwords). Among these words, 14535 were unique (14386 non-stopwords). We used the sentence transformer model, specifically the RoBERTa model. RoBERTa (short for ''Robustly Optimized BERT Approach'') is a variant of the BERT (Bidirectional Encoder Representations from Transformers) model, which can generate contextualized representations of words in a sentence. Sentence-RoBERTa is a transformer-based language model which is operating in sentence level-space. It maps sentences and paragraphs to a 1024D dense vector space and can be used for tasks such as clustering [34]. The model selected here for feature embeddings were the pre-trained model from the HuggingFace Model hub. The Sentence-RoBERTa use Siamese and triplet networks with cosine similarity function as loss function [35].

b: FACE FEATURES

The frames containing face images are identified from the video through image registration and the faces are extracted using Multi-Task Cascaded Convolutional Neural Networks (MTCNN) [36]. The MTCNN initially put forth a series of bounding boxes on the faces through Proposal Network

FIGURE 2. Generalized workflow of the proposed architecture.

(P-Net), which is fine tuned with a Refine Network (R-Net). The Output Network (O-Net) finally gives the face cropped image. We are using OpenFace architecture, which is a library installed in torch by DeepFace for extracting face embedding. The face images are resized to 96 × 96 pixels. The 2-D Affine transformations are used to normalize the faces in OpenFace architecture. OpenFace is trained with 500k images from CASIA-WebFace and FaceScrub datasets. The backbone of the network is Inception-Resnet V1 model, where the loss function is a triplet loss function [37]. From each video file 10 frames with faces where sampled the features are extracted by retraining the network. We created a CNN network for classification of personality traits by transfer learning through the said OpenFace model and our model had an additional layer for classifying each personality traits as shown in Figure 3. Before selecting the said model we compared the performance with FaceNet and ArcNet architectures which are under the same category. The comparison of performance in terms of accuracy is shown in Figure 4 The 'sigmoid' activation function was used. The optimizer used was 'adam', with a learning rate of 0.001, with a minibatch size of 64. The deep fearures are extracted from the intermediate layer with a dimension of 7360D which ensures a time series analysis.

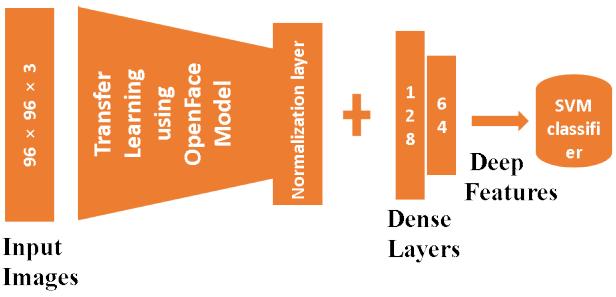


FIGURE 3. Personality classification with OpenFace Model.

c: SCENE/CONTEXT FEATURES

The video sequences of 15 seconds duration are presented with a single background; only a single frame was enough for extracting the context information. The VGG-19 network, which was trained for object recognition tasks was selected. The 4096 D feature was extracted from the 39^{th} layer of the 43-layer architecture into a 1-D vector space. [38]

The overall pipeline for predicting personality trait is shown in Figure 2. The deep features are combined together using fusion techniques and the features are effectively

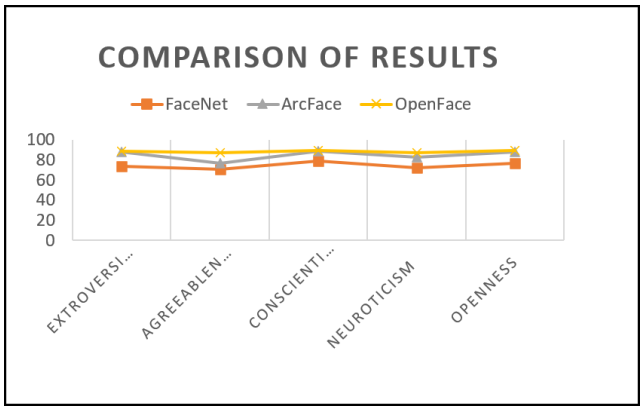


FIGURE 4. Comparison of face feature extraction networks for classification of personality traits.

selected based on the Eigen vectors. The prediction of personality traits are done by SVR and similar machine learning algorithms.

2) FOR RECOGNITION OF EMOTION

a: AUDIO FEATURES

The audio features are extracted using the open-source tool, COVARAP [39] at a frame length 10 ms. The audio features including prosody, MFCC (Mel-Frequency Cepstral Coefficients), and voice quality features like tenseness, creakiness, etc were given to a 1-D CNN network, implementing two dense layers. The number of filters are 64 and 32, respectively. The dense layers which had 256 and 128 neurons carried 128 filters in each layer. The Mean square error is monitored for convergence. The network was optimized with Adam optimizer at a learning rate of 0.001. The 'relu' activation function was used on each layer. The network is trained for the classification of six discrete emotions and the embeddings of 256D were extracted from the intermediate layer.

b: IMAGE FEATURES

From the 30 seconds videos. 90 images of size were sampled and face-only images were cropped through MTCNN which are resized to the dimension of 96 × 96. these images are passed to a transfer learning network to extract the embeddings. We used OpenFace model from deep face for this purpose. Finally, there are 1213 forlders with 90 × 128 embeddings. An LSTM layer of size 128 is added to the network followed by a dense layer of 256 neurons and 'relu' activation function, before the final classification layer. The

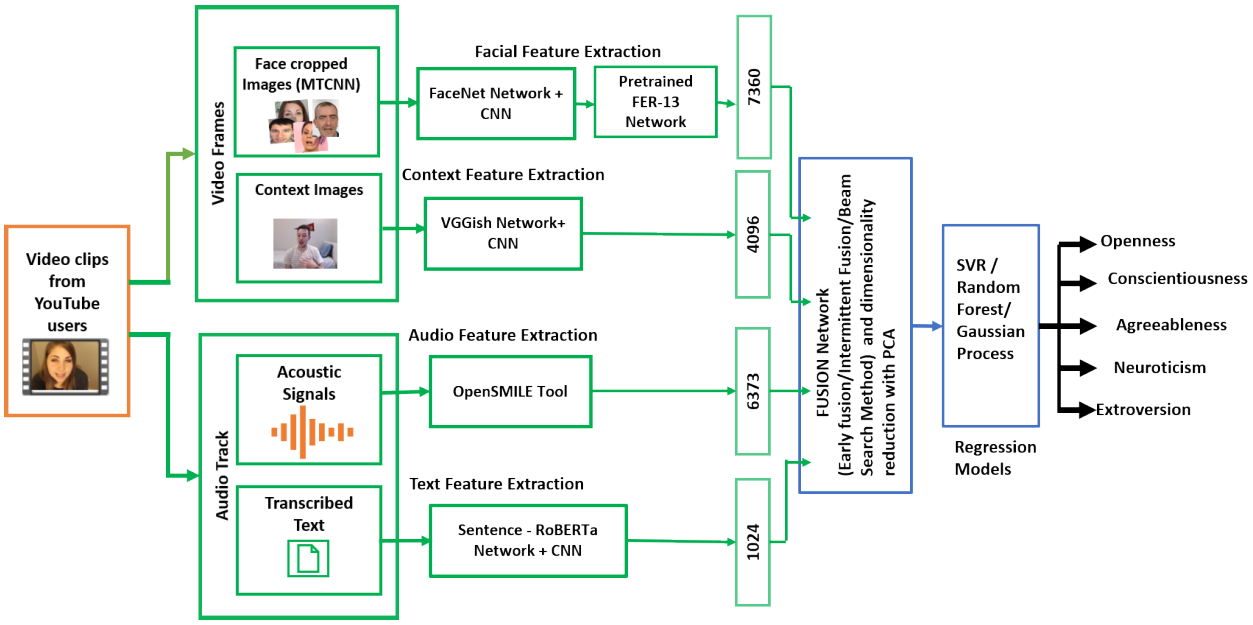


FIGURE 5. Framework to predict the personality traits.

256D extracted from dense layer forms the final embeddings for fusion models.

3) FOR SENTIMENT ANALYSIS

a: AUDIO FEATURES

The extracted features include prosody, energy, voice probabilities, spectral, and cepstral estimates. Each file contains 28 fields per line, where each line represent an audio frame sampled at 25ms. Each field is separated by a semicolon. Each file contains the following features: frameIndex, frameTime, intensity, loudness, 12 MFCC coefficients, 8 lspfreq components, zero crossing detector,voiceProb, F0, F0env and RMSenergy. The features are passed to neural network model with three fully connected layers before adding the classification layer. Each layer has 64, 128 and 32 layers respectively in it. The 'relu' activation function is used and the network is trained with 'binary cross entropy' as loss function. The training is done with 16 epochs with 'adam' optimizer at a learning rate of 0.01.

b: IMAGE FEATURES

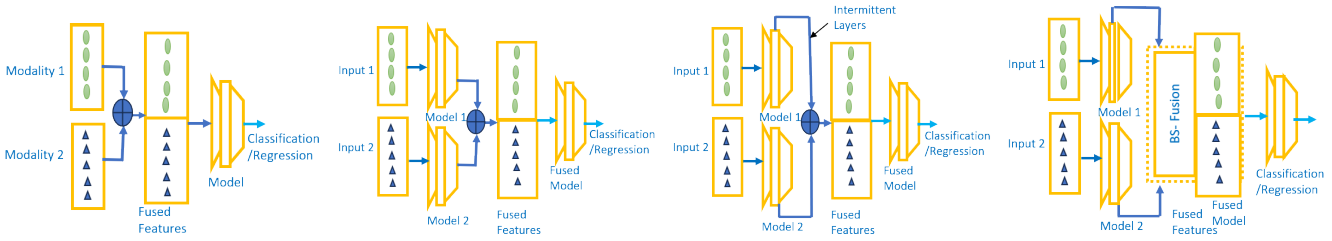
Visual features are acquired at frame level. Each file contains visual features including smile and head pose, action units (AUs), and six basic emotion estimates. The features were extracted using The Computer Expression Recognition Toolbox (CERT). Each video is sampled at 30 frames per second and the features corresponding to each utterance can be extracted using the transcription time stamp. The samples are stacked together and applied as input to deep neural network model. The model has two Bidirectional LSTM (B-LSTM) layers with 128 filters each which will be helpful in collecting the context information from these stacked image details [40]. The BLSTM is effective in handling variable length sequences by reducing the noise. The network is complete with two dense layers of 256 and 64 filters each with a 'dropout' layer before applying the final classification layer with 'sigmoid' activation function. The model is compiled with adam optimizer at a learning rate of 0.01. The training is done with 500 epochs by applying 'binary cross entropy' as the cost function.

c: TEXT FEATURES

Each video of 30 seconds was conditioned into an average of six utterances, resulting in a final dataset of 498 1-D utterances. The spoken sentence is linked to the corresponding audio and video streams, as well as its manual transcription. The utterances have an average duration of 5 seconds, with a standard deviation of 1.2 seconds and is annotated using ELAN tool [41]. The Sentence-Roberta model from Hugging face was used for extracting the embeddings. 1024 embeddings are collected and vertically stacked.

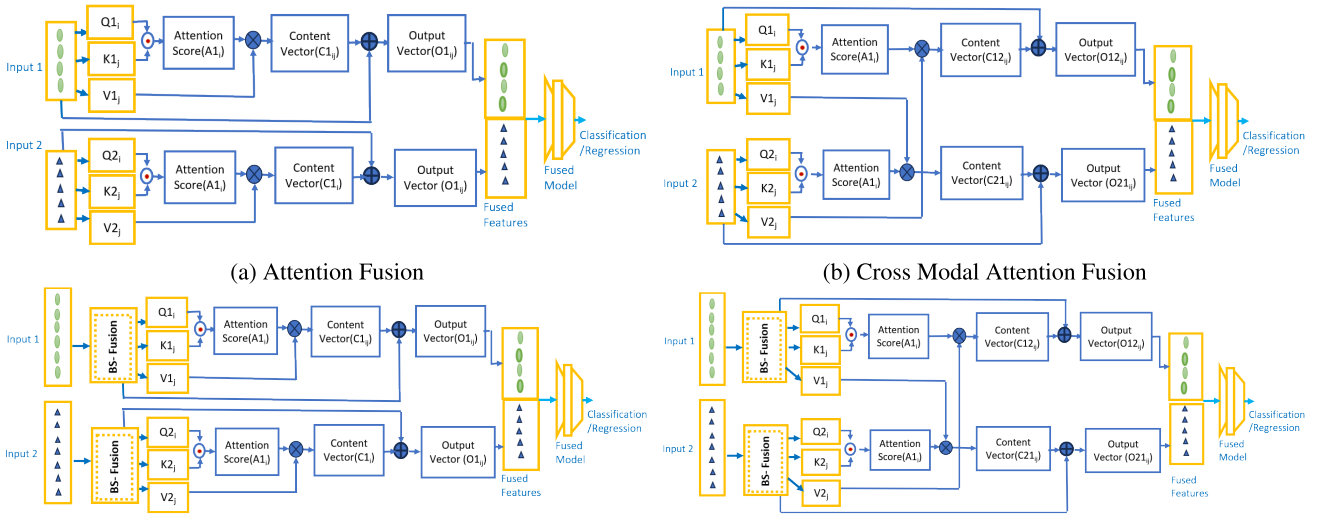
B. MULTIMODAL FUSION TECHNIQUES

The methods adopted for feature fusion are discussed in the following section. We applied fusion of different modalities to complete the pipeline with reduced dimensions. The attention based mechanisms are incorporated to select the relevant features. We formulated new cross modal attention (inter model) and simple attention model (inter model) strategies on intermediate layers. We also tested the system with various fusion strategies. The system underwent rigorous testing with various fusion strategies, including early fusion, intermittent layer fusion, weighted score level fusion and Beam Search fusion(BS-Fusion) as shown in Figure 6.



(a) Early Fusion (b) Late Fusion (c) Intermittent Level Fusion (d) BS-Fusion

FIGURE 6. State-of-the-art multimodal fusion techniques.



(a) Attention Fusion (b) Cross Modal Attention Fusion

(c) BS-Fusion based Cross Self Attention Fusion (d) BS-Fusion based Cross Modal Attention Fusion

FIGURE 7. Proposed attention based fusion techniques.

Before applying feature fusion with acoustic and visual modalities, an intermediate stage with feature standardization was performed. We used the standard scalar function available in the sci-kit learn package for this purpose. Standardization is a scaling technique wherein the data is converted to a new range that maps the statistical distribution of the data into a new selection. The new distribution of the data z_j can be represented as:-

$$z_j = \frac{x_j - \mu}{\sigma} \tag{1}$$

where μ is the mean and σ is the standard deviation of a modality features x_j is the feature under consideration.

In the early fusion technique, the feature-level information from different modalities is concatenated to achieve the final feature set as in Figure 5(a). Consider that there are two modalities M1 and M2 with features X1i and X2j. The features in this individual modalities can be concatenated together to get the combined feature X, which is applied on the prediction model.

$$X = [X1i, X2j] \tag{2}$$

It has less complexity and computational burden.

Intermittent layer fusion is a technique used in multi-modal data fusion, where two or more modalities are fused at specific layers of a neural network architecture. The mathematical approach for intermittent layer fusion with 2 modalities can be summarized as follows: Let $H1^i$ represents the i^{th} layer of modality 1 and $H2^j$ represents the j^{th} layer of modality 2, the embeddings from the respective layers are taken out before applying the dimensional reduction techniques.

$$H^l = [H1^i, H2^j] \tag{3}$$

The layers from which the embeddings are selected is fixed through a grid based search method.

In the case of weighted fusion, the above technique is improved as:-

$$H1^{(i)} = \sum_{l=1}^{M1} W_l H_l^{(i)} \tag{4}$$

where W_l represents the learned weight of modality 1 with M1 features. The weighting factor is calculated based on the

score secured for individual modality.

$$W_l = \frac{Score_{test_l}}{\sum_{j=1}^{N} Score_{test_j}} \tag{5}$$

where W_l is the weight of the l^{th} model.

In the case of BS-fusion [42], the algorithm works in two steps. In the first step a beam search is carried out to find the best features, and in the second step the fusion of the features is done. Beam search is a method that uses a heuristic approach to explore a graph. It focuses on extending the most promising node within a limited set. It is an optimization of best-first search that reduces its memory requirements. It is run to find the best combination of modalities for applying feature-level fusion by implementing the scoring mechanism. The layer which is giving the best score is selected in each modality. The search space in our case includes the different layers of the OpenFace architecture for image modality, BERT based model for text modality and selected layers in the CNN/ LSTM architecture for other modalities. The score can be calculated as

$$score(H1_i) = \sum_{m=1}^{M} W_m \cdot score_m(Hm_i) \tag{6}$$

M is the number of modalities, and W_m is the weight for modality m.

$score_m \cdot (H_{mi})$ is the score of the i^{th} layer for modality m.

Turning to the attention mechanisms, we tried both simple attention (intra-modality attention) and cross modal attention (inter-intra modality attention) schemes. Attention mechanisms can suitably select the relevant features from a large number of available features. The attention mechanism computes the degree of attention (or weight) to be assigned on each value vector based on the similarity to the query and key vectors, allowing it to capture and emphasize relevant information from the input data [43]. The attention weights are normalized between 0 and 1. The weights determine the strength of attention. The novelty of our approach is that we applied the attention mechanisms on the previously selected intermediate layers and attention scores were calculated on audio and video features. Through this proposition more general features were collected. That is the focus was on mid level features as well as low level features.

In the cross model attention mechanism, which is acting between audio and video embedding. The approach is a two step process, where in the first step, a simple attention mechanism is applied on individual modalities. In the next step, the cross modal attention mechanism is applied where the video based attention features are applied on audio features and vice versa. Thus the impact of audio modality could be calculated on the video modality and vice versa. Here the modalities are run with complimentary information, rather than the taking the attention features from the same modality. The schematic of the proposed system is given in Figures 7(a) and 7(b).

The attention weight on attention mechanism using inter-modality and intra-modality approaches are as under:-

$$W_{mi} = \frac{e^{(Q_{mi} \cdot K_{mi})}}{\sum_{i=1}^{M} e^{(Q_{mi} \cdot K_{mi})}} \tag{7}$$

where mi represents the i^{th} feature of modality m. Thus the context vectors are obtained as

$$C_{mi} = W_{mi} \cdot V_{mi} \tag{8}$$

for intra-modality attention fusion and

$$C_{M1i} = W_{M1i} \cdot V_{M2j} \tag{9}$$

for inter modality attention fusion where Q_m, K_m and V_m are the parameters Query, Key and Value respectively and is obtained by:-

$$[Q_{mi}, K_{mi}, V_{mi}] = \sigma(w_m[Q_{mi}, K_{mi}, V_{mi}] + q_m \tag{10}$$

With w_m and q_m as weights and bias values.

Further, we applied Bilinear-Search on the feature set so as to select the most relevant data for feature fusion as in Figure 7(c) and 7(d). This approach could reduce the features without PCA and we could observe further improvement in results, which can be attributed to the suppression of noise and redundancy, while reducing the features.

C. REGRESSORS

1) SUPPORT VECTOR CLASSIFIER/REGRESSOR

The Support Vector Regression (SVC/R) is a non-parametric technique, with a symmetric loss function, with an objective to fit the error within a margin ϵ [44]. The SVC/R is characterized by linear and non-linear kernels with the classic kernel trick to mitigate the curse of dimensionality, sparse solution and Vapnik-Chervonenki (VC) control of marginal supports, and the presence of support vectors. The uniqueness of SVR is its excellent prediction accuracy and generalization capability. We are applying SVR with a non-linear kernel, the Radial Basis Function (RBF). The hyper-parameters can be trained to optimize prediction accuracy. We are controlling the regularization parameter 'C' and gamma, which shows how far the influence of a single parameter reaches [45]. The tuning of hyper-parameters is always a trade-off between 'C' and 'gamma'.

2) RANDOM FOREST CLASSIFIER/REGRESSOR

Random Forest Classifier/Regressor is a multiple-tree structure based on ensemble learning. It uses the bagging technique or bootstrap aggregating to avoid overfitting issues. It predicts by taking the average of the output from various trees. Increasing the number of trees increases the precision of the outcome [46].

3) GAUSSIAN PROCESS CLASSIFIER/REGRESSOR

The third type of algorithm we applied here is the Gaussian Process Regressor (GPC/R). This is a non-parametric supervised learning model which learns mainly from data with

minimal hyper-parameters. GPC/R calculates a probability distribution as an overall admissible function that fits the data. During training, posterior is calculated using training and predictive posterior distribution over test data [47].

The regressors and classifiers can offer the best results by applying the features skimmed out by reducing the correlated and irrelevant data. Dimensionality reduction is a proven strategy for the same. In this study, Principal Component Analysis (PCA) and its non-linear variant Kernel PCA are used for dimensionality reduction. The input of PCA is original extensive dimensional data, and this high dimensional data finds new coordinates, which are its principal components. These components are selected based on the variance and are orthogonal to each other. The features were pre-processed with a standard scalar before applying dimensionality reduction.

IV. EXPERIMENTS AND RESULTS

In this section, we evaluate the effectiveness of the proposed framework on the three selected datasets using different machine learning algorithms viz, Support Vector Machines, Random Forest and Gaussian Process Classifier.After being retrieved from CNN-based networks, the deep features underwent additional processing before being applied to the chosen regressors. We scaled the data using the statistical mean and the standard deviation. We examined the effect of each individual features as well as combined modality features on the perception of the personality, emotion and sentiments both in terms of the robustness of the model and the model performance. The multimodal scenario is carefully constructed using attention based fusion methods described in the earlier section.

The attention mechanism is implemented has a hidden dense layer with 'relu' activation function. We used 'adam' optimizer with learning rate as 0.001. L-2 regularization was applied to improve the performance of the model. The SVC/R with the RBF kernel is applied with hyperparameters 'gamma' as 0.01 and penalty parameter 'C' as 10. It was selected by running a grid based search algorithm. For Random Forest Classifier/Regressor, the number of trees built by the algorithm before averaging the products is kept at 6. Finally, for GPC/R, the kernel is RBF with white noise, the number of parameters is 350 and the loss function chosen is 'mean squared error'.

A. DATASET DESCRIPTION

1) THE CHALEARN FIRST IMPRESSIONS DATASET

The ChaLearn First Impression dataset was released in 2016 and enhanced in 2017 as ChaLearn V2. It contains 10000 talking-to-the-camera audio and visual clippings from around 2764 YouTube users. Each clipping typically lasts around 15 seconds with 30 frames per second. The participants in the videos are from different gender, ages, nationality, and ethnicity. These are single-person videos, the scene is clear, and speakers keep their faces aligned with the camera for at least 80 percent of the allotted time. Each video has a label attributed to the Big-Five personality traits (OCEAN model) by using Amazon Mechanical Turk annotators. The value allocated for each trait was then normalized to the range 0 - 1. The dataset is subdivided into training (6000 videos), validation (2000 videos), and testing (2000 videos), which comes as a 3:1:1 split.

2) THE ELDERREACT DATASET

ElderReact is a comprehensive dataset containing 46 elderly individuals. It includes 1323 video clips, each lasting between 3 to 8 seconds. These films were sourced from the YouTube Reach channel. The dataset is annotated with 6 basic emotions, viz, Anger, Disgust, Fear, Happiness, Sadness and Surprise, through Amazon Mechanical Turk, based on Discrete Emotions Questionnaire. Here 615 video clips are reserved for training, 355 clips are kept for validation and the rest 353 clips are taken only for testing purposes [10].

3) THE MOUD DATASET

The Multimodal Opinion Utterances Dataset (MOUD) product review videos in Spanish with multiple segments labeled to display positive, negative or neutral sentiment [41]. We used the dataset as a binary class problem with positive and negative sentiments only. There were 80 videos of 30 seconds duration, and it is manually ensured that the videos covered recommendations single topic only. The videos are split in utterance level and out of the 498 splits 438 segments were selected after the pre-processing steps.

B. EVALUATION METRICS

The performance metrics used in this study are Mean Accuracy(1 - MAE (Mean Absolute Error)), and Square of correlation coefficient (R^2).

1) MEAN ACCURACY

The metric used here is Mean Accuracy, which is calculated by taking 1-MAE, which is the score proposed by the organizers of the First Impressions challenge.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_{tar_i} - y_{pred_i}| \tag{11}$$

2) SQUARE OF CORRELATION COEFFICIENT (R^2)

R^2 is defined as

$$R^2 = 1 - (SS_{res}/SS_{tot}) \tag{12}$$

where SS_{res} –> residual sum of squares and SS_{tot} –> total sum of squares.

The relation between the predicted and actual values can be analyzed based on the R-square metric for all the regression models. The higher the value of R^2 nearer to one, the better the model will be. The value of R^2 can be positive or negative based on whether the model is fitted better than the average or worse when compared to the average value of the model. It shows how well the data fit in the regression model(goodness of fit) [48].

3) F1-SCORE

The performance matrices based on **Usage Prediction** [49]

- Recall: What proportion of items that a user likes were actually recommended
  It is given by:

$$Recall = \frac{tp}{tp + fn} \tag{13}$$

  Here tp represents the number of items recommended to a user that he/she likes and tp + fn represents the total items that a user likes
- Precision: Out of all the recommended items, how many did the user actually like? It is given by:

$$Precision = \frac{tp}{tp + fp} \tag{14}$$

  Here tp represents the number of items recommended to a user that he/she likes and tp + fp represents the total items recommended to a user
- F1-score: Harmonic mean of accuracy and precision

$$F1 - Score = \frac{Precision * Recall}{Precision + Recall} \tag{15}$$

C. PREDICTION OF PERSONALITY TRAITS

In this section, the performance analysis of unimodal systems and multimodal systems applied on Chalearn dataset is presented. The selected individual modalities are video features such as context and face-cropped images and acoustic features such as audio and text transcriptions. Transfer learning using OpenFace and VGG-Face-based networks were used for extracting the image features. The context and face images give 7360 and 4096 embeddings respectively. There are 6373 features in audio modality, and BERT based 1024 features from transcriptions. The performance outcome of the experiments by applying the mean accuracy metric is presented in Tables 1 and 2. The results of all the three regressors are given as three segments in the table. It gives the notion of the performance of the individual modalities and all the fused ones in estimating the personality traits using the selected regression algorithms.

On a detailed perusal of individual modalities, it can be observed that the face image embeddings yielded the best outcomes, while embeddings from transcriptions yielded the lowest. Based on individual traits, it can be noted that the estimation rate of 'openness' and 'conscientiousness' is higher than 'agreeableness'. The second and third column segments of Table1 give scores with dimensionality reduction using PCA and Kernel-PCA, respectively. The number of features selected for each modality is listed in Table 4. We have selected the number of components based on the hyper-parameter tuning. It can be observed that both PCA and Kernel-PCA are improving the prediction scores. This can be attributed to the improved generalization capabilities of the regression algorithms. The number of selected features depends on the 'explained variance ratio' which is based on the Eigen vectors [50]. Explained variance is a measure of

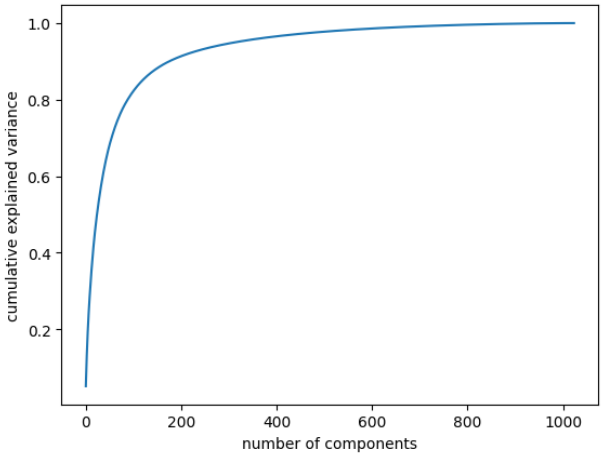


FIGURE 8. Plot for explained variance for PCA.

the amount of information that is accounted for by distinct principal components (eigenvectors). It is derived by taking the ratio of the eigenvalue of a specific principal component (eigenvector) to the sum of all eigenvalues. The explained variance is applied on the normalized data only. The sample plot of explained variance is shown in 8 for text data in chalearn dataset. It can be observed that a maximum of 700 features are most relevant ones The major limitation of PCA is that it can be affected by outliers. In our algorithm, the influence of outliers is reduced by the normalization applied in the preprocessing stages. In the case of BS-Fusion based attention mechanism, without PCA, better results are achieved. In this case, the total number of features are restricted to 1900.

Next, the discussion is on the effect of fusion strategies on the system. First, we will discuss mean accuracy itself. All the fusion strategies discussed here are showing better results than the performance of the individual modalities. We are taking the features from intermediate layers, which has mid level information along with the low level features for the proposed attention mechanisms. The simple attention mechanism takes up features from intermediate layer in the same modality, while in the cross modal approach, inter-intra model features are considered for the proposed attention mechanisms. The attention mechanism is further enhanced through a bilinear search algorithm. They are compared with four other fusion mechanisms. Among the fusion strategies the proposed system is found to be giving the best results consistently with dimension reduction. Next, while investigating the performance of the regressors, and the best results are given by SVR.

Next, we look over the influence of multimodal fusion in the dataset. It indicates the positive correlation with feature selection in the proposed models. We are showing the results of the experiments on the fusion architectures given in Figure 6. The results on all the selected regressors are tabulated in Table 1. Here, GPR is yielding the results better than the other two with independent modalities, while coming

TABLE 1. Performance Analysis of Personality Traits based on 1-MAE values.

Modality	Without PCA					PCA					K-PCA				
	Extr	Neu	Agr	Conn	Open	Extr	Neu	Agr	Conn	Open	Extr	Neu	Agr	Conn	Open
Support Vector Regression															
Audio	0.906	0.897	0.896	0.903	0.901	0.902	0.903	0.904	0.911	0.909	0.909	0.909	0.911	0.914	0.915
Context	0.909	0.902	0.900	0.907	0.908	0.913	0.915	0.912	0.914	0.913	0.913	0.912	0.914	0.914	0.917
Image	0.912	0.909	0.911	0.914	0.915	0.912	0.911	0.91	0.917	0.916	0.915	0.912	0.910	0.916	0.917
Text	0.893	0.894	0.897	0.901	0.897	0.904	0.904	0.902	0.906	0.905	0.904	0.904	0.902	0.906	0.905
Early	0.913	0.912	0.909	0.911	0.913	0.916	0.917	0.915	0.919	0.918	0.916	0.916	0.915	0.918	0.918
Intermittent	0.914	0.913	0.911	0.913	0.914	0.917	0.917	0.921	0.921	0.922	0.920	0.919	0.919	0.920	0.920
Weighted	0.917	0.918	0.915	0.915	0.917	0.916	0.916	0.917	0.919	0.919	0.916	0.916	0.916	0.919	0.918
BS-Fusion	0.920	0.919	0.919	0.920	0.920	0.921	0.920	0.921	0.921	0.920	0.921	0.921	0.922	0.923	0.917
Attention	0.919	0.921	0.920	**0.921**	0.920	0.922	0.921	0.923	0.924	**0.926**	0.922	0.923	0.923	0.925	**0.927**
Cross modal	0.918	0.919	0.917	0.918	0.920	0.922	0.921	0.924	0.923	0.923	0.921	0.922	0.921	0.924	0.923
Random Forest Regression															
Audio	0.887	0.889	0.895	0.901	0.899	0.902	0.903	0.904	0.909	0.904	0.909	0.909	0.911	0.914	0.905
Context	0.902	0.901	0.900	0.902	0.901	0.913	0.912	0.912	0.911	0.912	0.913	0.913	0.912	0.915	0.917
Image	0.904	0.907	0.903	0.911	0.910	0.912	0.9112	0.91	0.917	0.916	0.915	0.915	0.91	0.916	0.916
Text	0.893	0.894	0.897	0.900	0.901	0.895	0.899	0.903	0.904	0.904	0.903	0.904	0.893	0.903	0.905
Early	0.909	0.912	0.911	0.908	0.911	0.911	0.917	0.915	0.918	0.918	0.916	0.917	0.915	0.918	0.918
Intermittent	0.915	0.912	0.911	0.910	0.912	0.916	0.914	0.918	0.917	0.916	0.919	0.917	0.918	0.920	0.920
Weighted	0.900	0.901	0.905	0.908	0.915	0.909	0.905	0.905	0.914	0.917	0.916	0.909	0.911	0.914	0.915
BS-Fusion	0.916	0.91	0.912	0.915	0.917	0.913	0.913	0.912	0.915	0.917	0.913	0.913	0.912	0.915	0.917
Attention	0.916	0.915	0.910	0.916	0.917	0.915	0.915	0.916	0.918	0.920	0.915	0.917	0.916	0.919	0.920
Cross Modal	0.914	0.914	0.911	0.917	0.915	0.911	0.915	0.914	0.915	0.919	0.916	0.911	0.913	0.916	0.915
Gaussian Progress Regression															
Audio	0.902	0.903	0.904	0.911	0.909	0.906	0.906	0.908	0.914	0.914	0.906	0.906	0.906	0.915	0.914
Context	0.913	0.912	0.912	0.911	0.912	0.916	0.918	0.917	0.916	0.917	0.917	0.919	0.916	0.915	0.915
Image	0.915	0.916	0.912	0.915	0.916	0.915	0.916	0.916	0.917	0.916	0.918	0.912	0.913	0.921	0.920
Text	0.895	0.899	0.903	0.904	0.904	0.897	0.904	0.906	0.906	0.907	0.900	0.904	0.908	0.904	0.905
Early	0.916	0.917	0.915	0.918	0.918	0.916	0.921	0.919	0.920	0.921	0.915	0.919	0.920	0.922	0.922
Intermittent	0.921	0.919	0.920	0.921	0.921	0.922	0.919	0.922	0.923	0.923	0.920	0.919	0.921	0.923	0.922
Weighted	0.910	0.911	0.914	0.916	0.917	0.916	0.915	0.919	0.920	0.917	0.916	0.916	0.917	0.920	0.918
BS-Fusion	0.913	0.913	0.912	0.915	0.921	0.917	0.917	0.917	0.918	0.922	0.916	0.917	0.918	0.919	0.923
Attention	0.917	0.918	0.916	0.915	0.919	0.920	0.921	0.919	0.924	0.921	0.924	0.922	0.923	0.928	0.926
Cross modal	0.915	0.915	0.916	0.916	0.917	0.917	0.916	0.918	0.921	0.919	0.918	0.917	0.919	0.922	0.923

TABLE 2. 1-MAE values for Bilinear Search based Attention Fusion.

Support Vector Regression					
Fusion Method	Extr	Neu	Agr	Conn	Open
BS-Attention	0.929	0.927	0.924	0.923	0.923
BS-Cross Modal	0.927	0.928	0.923	0.921	0.926
Random Forest					
BS-Attention	0.920	0.916	0.917	0.917	0.921
BS-Cross Modal	0.920	0.918	0.916	0.917	0.920
Gaussian Process Regression					
BS-Attention	0.921	0.924	0.920	0.920	0.925
BS-Cross Modal	0.925	0.919	0.920	0.923	0.924

TABLE 3. The R^2 measure of regressors.

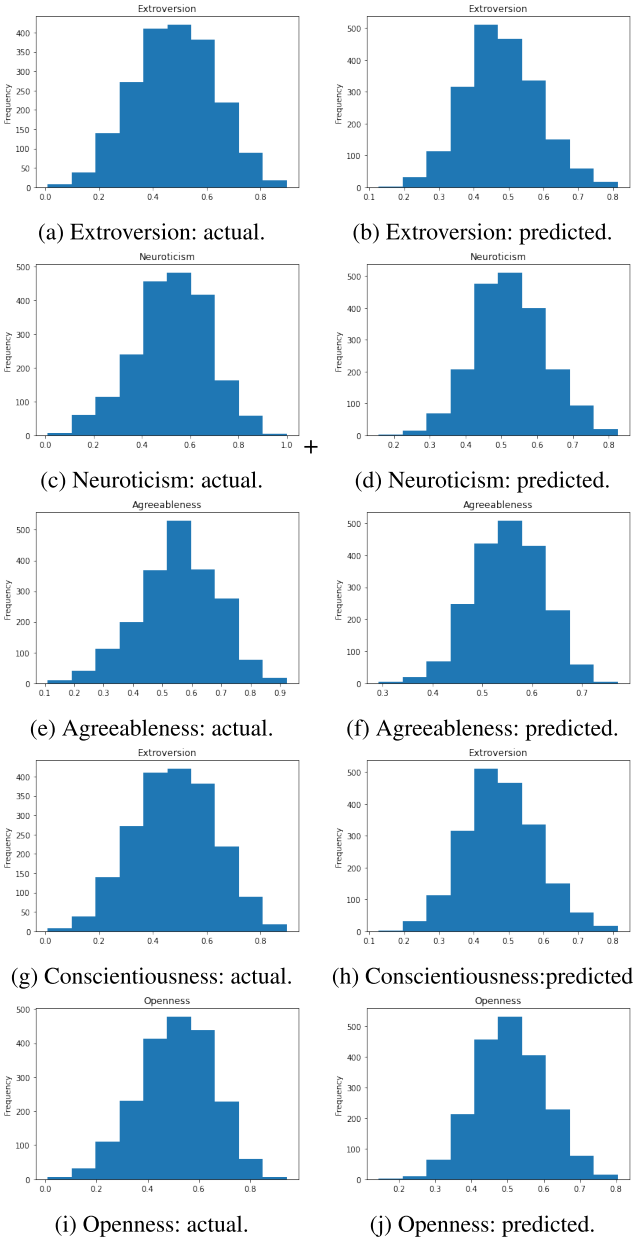
s.no	Modality	SVR	RF-R	GPR
1	Audio	0.551	0.517	0.558
2	Context	0.529	0.511	0.519
3	Face Image	0.560	0.521	0.504
4	Text	0.518	0.363	0.454
5	early fusion	0.569	0.523	0.552
6	intermittent	0.589	0.525	0.571
7	weighted	0.571	0.533	0.541
8	BS-Fusion	0.609	0.592	0.627
9	attention level	**0.689**	0.625	**0.664**
10	cross modal	0.673	0.623	0.641

to the proposed attention based fusion model, SVR gives the best outcomes. Further, the R^2 values for different regressors are presented in Table 3.

The attention level fusions were giving better than the other methods. In both the said methods all the accuracy traits are predicted equally well, which indicates the robustness of the system. The distribution of predicted and actual scores of the 5 traits in the test data using histograms are shown in Figure 9. The comparison is done on the SVR model with BS-Fusion, where the best of audio, video, and text modalities are selected. It shows the consistency of the algorithm.

TABLE 4. Performance analysis with the state-of-the-art for Chalearn dataset.

S. No	Features	Dim	MAE
1	OpenFace	7360	0.912
2	VGGFace-FER13	4096	0.906
3	VGG-VD-19	6373	0.908
4	Transcript	1024	0.891
5	Attention fusion	188516	0.922
6	Fusion with PCA	2000	0.927
7	BS-Attentiom	1900	**0.929**
8	[18] Yagmur 2016	-	0.911
9	[27] Dersu 2020	-	0.915
10	[24] Soganciogl 2021	-	0.902
11	[26] Aslan2021	-	0.917
12	[51] Madan 2021	-	0.906
13	[52] Yan 2021	-	0.913
14	[21] Tellamekala 2022	-	**0.926**
15	[17] Suman2022	-	0.914



(a) Extroversion: actual. (b) Extroversion: predicted.

(c) Neuroticism: actual. (d) Neuroticism: predicted.

(e) Agreeableness: actual. (f) Agreeableness: predicted.

(g) Conscientiousness: actual. (h) Conscientiousness:predicted

(i) Openness: actual. (j) Openness: predicted.

FIGURE 9. Histogram of Actual and Predicted value of all the traits.

D. CLASSIFICATION OF EMOTIONAL STATES

The next objective is to analyze the performance of the suggested framework on the classification of six emotional states. The performance analysis takes place based on the accuracy and F1-Score. In this case we are taking the deep features from audio and image inputs only. Since the emotions are more dynamic and profound, facial expressions and micro-expressions can provide clues about emotional states which is suitably extracted from the openFace architecture which is enhanced by the LSTM network incorporating the time dependency among the frames. There are only two modalities, audio and image, considered for fusion model. There are 256 features taken from both modalities and fusion was done using the same set of algorithms that we applied on the chalearn dataset. The ML algorithms used for classification of emotional states are SVM, Random Forest and Gaussian Process Classifier. We applied dimensionality reduction with PCA and K-PCA techniques before applying it to the final classifier.The features where normalized through standard scalar operations. The results are consolidated in Tables 5-7. It can be observed that fused models are outperforming the individual models consistently and here also visual modalities are yielding with more information. Among the various fusion techniques that is applied here, attention based mechanisms are performing better than their counterparts. Apart from that cross modal attention mechanisms are offering better performance than that on personality traits. This can be accounted for the increased dynamics of the features selected for emotion classification. Except for the emotion 'surprise', SVM is performing better than the other two classifiers. The best results are given by 'fear' and 'disgust'.

E. SENTIMENT ANALYSIS

Our third experiment is to evaluate the proposed general architecture and fusion algorithm for sentiment analysis, which is a binary classification problem. We selected the MOUD dataset for binary classification. The selected features are from audio, image and text modalities. There were 256 features taken from audio and image modalities which are extracted from the intermediate stages. The BERT model had 1024 features which was fused together using the suggested fusion mechanisms. The temporal and deep information are extracted by the networks The feature fusion was performed on the dataset and the binary class problem gave very good results. The results indicating the accuracy and F1-score are shown in Table 8. All the fusion methods are outperforming the results with individual modalities. The attention based methods are found to be outperforming the other fusion methods. The cross modal approach could bring out the complementary information in a better manner here. The performance of the classification model with multimodal fusion can be measure by a ROC curve (receiver operating characteristic curve), which is a graph between True Positive rate and False Positive Rate. The Figure 10 shows the ROC curve for SVM classification fused with attention mechanism. The area under the curve is 0.94. All the algorithms are performing equally well here, but the best results are obtained by SVM itself.

F. COMPARISON WITH OTHER WORKS

Now, we contrast our best findings with those of other researchers in terms of network architecture complexity. Accuracy reported on the ChaLearn V2 dataset are shown in Table 4. In the first kind, deep architectures were directly deployed.

In [18] audio modality-based 17-layer deep residual network and a video modality-based 17-layer deep residual network were merged to a fully-connected layer. This whole audiovisual network is a total 18-layer deep residual network

TABLE 5. Performance Analysis of Emotion dataset based on Accuracy and F1-Score using SVM.

Modality	Anger		Disgust		Fear		Happy		Sad		Surprise	
	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)
Audio	55.4	57.14	57.0	51.0	65.3	67.7	65.9	70.0	57.9	57.7	55.4	61.3
Image	57.4	61.3	63.0	67.0	58.0	65.1	59.0	67.0	53.9	58.1	61.3	66.0
Early fusion	61.4	61.7	65.0	61.5	65.4	73.0	66.9	76.8	56.0	65.0	65.3	70.1
Intermittent	78.0	77.67	84.0	84.67	89.9	90.33	49.1	54.67	79.9	80.67	64.0	65
Weighted	64.1	63.2	66.2	67.4	75.4	76.4	77.9	79.1	68.2	65.6	64.2	67.6
BS- Fusion	77.8	78.1	82.1	79.9	83.7	81.6	67.8	69.2	80.1	81.2	65.1	68.5
Attention	79.3	79.5	87.2	84.3	91.2	90.4	79.8	80.1	82.3	81.4	74.2	73.4
Cross Modal	78.1	78.7	88.1	82.4	88.4	87.9	78.3	79.2	80.1	80.2	73.2	72.4
BS-Attention	81.1	80.9	90.1	83.1	90.7	89.8	80.4	81.4	81.2	80.7	74.3	83.4
BS-Cross Modal	79.1	78.7	90.9	82.4	91.4	87.9	79.5	79.2	83.6	80.2	76.2	72.4

TABLE 6. Performance Analysis of Emotion database based on Accuracy and F1-Score using Random Forest Classification.

Modality	Anger		Disgust		Fear		Happy		Sad		Surprise	
	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)
Audio	49.8	50.1	50.3	56.2	58.6	56.3	52.7	53.4	56.7	59.1	50.4	51.4
Image	54.6	55.6	59.7	56.8	60.5	61.2	62.1	61.2	55.6	56.7	57.3	54.9
Early fusion	60.2	57.8	61.2	60.9	63.4	65.5	67.8	65.8	60.1	59.8	62.3	64.5
Intermittent	70.9	71.2	64.4	64.8	65.9	67.6	60.2	64.8	62.8	66.7	65.7	64.5
Weighted	59.8	60.5	61.8	62.3	60.1	63.4	66.2	67.5	59.7	64.2	62.1	63.2
BS- Fusion	65.4	64.5	65.3	65.3	69.0	69.5	67.8	68.3	66.8	64.5	62.8	65.8
Attention	74.9	72.5	76.7	69.6	78.8	78.7	69.3	70.9	66.3	69.1	72.4	73.4
Cross Modal	70.5	69.8	64.6	65.9	75.6	79.7	71.2	70.2	68.3	69.5	70.9	73.4
BS-Attention	73.4	72.3	74.5	69.8	76.5	77.6	67.8	68.9	65.7	67.9	70.1	71.2
BS-Cross Modal	70.3	68.5	64.6	65.7	74.5	78.3	69.1	68.7	67.2	67.4	70.2	69.9

TABLE 7. Performance Analysis of Emotion database based on Accuracy and F1-Score using Gaussian Process Classification.

Modality	Anger		Disgust		Fear		Happy		Sad		Surprise	
	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)
Audio	54.3	56.2	58.9	62.3	59.6	58.7	64.5	63.5	53.8	54.6	54.3	58.9
Image	55.7	56.8	60.9	63.0	57.3	64.8	61.9	55.1	52.5	55.7	57.6	59.4
Early fusion	60.7	59.4	63.4	61.2	64.3	69.7	67.1	73.8	55.6	63.4	65.2	70.1
Intermittent	70.2	71.4	78.9	80.3	81.3	83.6	59.8	70.8	72.3	77.4	78.9	80.5
Weighted	67.8	67.9	62.3	62.4	67.8	67.9	65.7	71.2	56.7	64.5	64.5	69.8
BS- Fusion	71.0	73.4	77.8	80.1	80.5	78.9	66.0	76.9	76.3	78.1	77.9	80.4
Attention	73.6	78.9	77.8	85.6	80.9	83.6	76.8	79.6	80.2	78.9	81.3	79.9
Cross Modal	70.4	78.1	68.1	76.1	79.8	82.3	73.9	79.4	79.2	78.5	80.2	81.1
BS-Attention	74.6	79.2	79.9	87.3	82.2	83.7	77.8	80.5	81.2	80.9	81.7	81.9
BS-Cross Modal	74.4	79.4	70.2	78.2	80.2	82.9	74.5	79.9	80.1	79.7	80.5	82.0

and reported 0.9109 on the test set. In [21] there were two parallel streams. The first stream termed the baseline model composed of the CNN followed by a two-layer bidirectional GRU-RNN with 256 hidden units and finally one fully connected output layer to predict the estimates of the Big-Five personality traits. The second stream consisted of Conditional Latent Variable Models (CLVM) capturing the epistemic (model) and aleatoric (data) emotion uncertainties using the distributions of predicted valence and arousal as inputs. They reported the highest performance of 0.926 in the literature so far. In the second approach, pre-trained networks were used to extract different features, and one-stage or two-stage machine-learning models were used to learn the final outcome. Dersu et al. [27] extracted Facial appearance features using ResNext + CNN-GRU model; the facial action unit features Voice features and Transcribed speech features using separate LSTM Nets respectively. With a late fusion of these features, they reported a performance of 0.915. In [24], audio, video, and linguistic features were extracted separately for learning mood (Arousal, valence) and likeability predictions. These mood and likeability predictions were then used to predict the apparent personality traits using Explainable Boosting Machine (EBM) regressors. They also contributed Arousal, valence, and likeability annotations for the ChaLearn dataset. In [26] two-stage training was proposed. In the first stage,

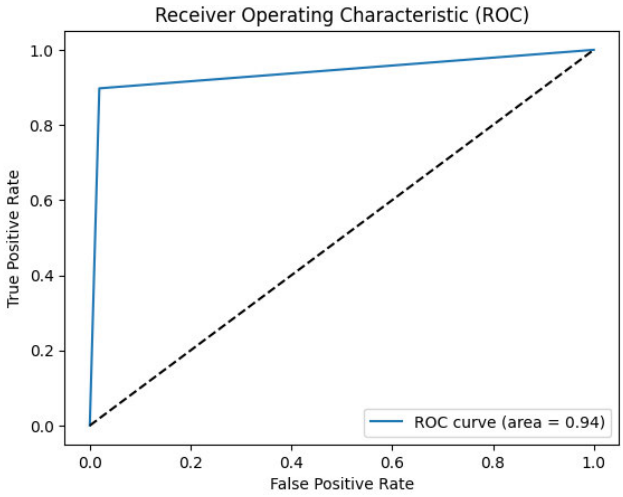


FIGURE 10. The ROC curve showing sentiment analysis.

TABLE 8. The Performance Analysis based on Accuarcy and F1-Score for sentiment analysis using SVM, Random Forest and Gaussian Process Classifiers.

Modality	SVM		Random Forest		Gaussian Process	
	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)	Acc (%)	F1-Score (%)
Audio	72.6	65.2	70.1	62.1	73.4	70.1
Image	73.1	63.1	71.12	69.2	72.7	71.12
Text	70.5	63.2	67.9	65.7	68.3	59.9
Early	81.3	80.6	81.7	79.9	81.5	80.7
Intermittent	82.6	75.7	83.7	78.7	85.8	79.7
Weighted	83.5	72.5	80.9	80.9	81.1	80.9
BS- Fusion	92.4	90.1	91.3	91.3	91.2	90.4
Attention	95.7	94.9	91.5	91.5	91.0	91.5
Cross Modal	92.4	89.0	90.4	90.4	89.8	88.4
BS-Attention	96.2	95.6	92.2	92.1	92.1	91.9
BS-Cross Modal	95.4	93.2	91.6	93.4	91.8	89.8

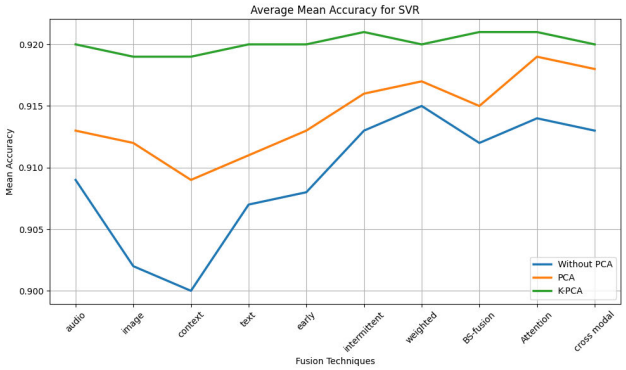
separate sub-networks were trained to estimate the personality scores to learn modality-specific representations. In the latter stage, the trained modality-specific representations obtained earlier were concatenated and processed through an attention module and finally used for training regressors to estimate the personality trait. Surbhi et al. [51] proposed kinemes feature representation based on elementary head-motion units. An LSTM was trained using Kinemes and Facial Action features. They reported a performance of 0.907. Shen et al. [52] investigated biases in multimodal systems and proposed two fairness metrics to quantify the biases of the model outcomes – Statistic Parity (SP) and Equal Accuracy (EA). They extracted facial features, scene features, audio features and performed feature-level fusion. Kernel extreme learning machines (ELM) were trained to get intermediate predictions. These intermediate predictions were used to train the Random Forest regressor for final Big-Five predictions. Suman et al. [17] extracted audio, video, and textual features using different available pre-trained CNNs: ResNet, MTCNN, and VGGish CNN. The extracted deep features are used for training a separate neural network. They reported 0.914 on the test set.

The comparison of results on personality traits is given in Table 4 Keeping the simplicity of the framework as the objective, we have adopted the second strategy for our solution. After extracting deep features using the pre-trained networks, we investigated dimensional reduction techniques for two-purpose: reduction in feature dimensions and exploring sublime feature embedding in the lower-dimensional space. Our best performance is 0.928, which is better than all works, and this with very few dimensions (2000D). It can also be observed that multimodal approaches can bring out the salient features in all the modalities effectively. Especially CNN-based deep features are more specific and free from biases that are in hand-crafted features. SVR was the best option when comparing the performance of all the regressors because of its exceptional qualities, including strong generalization performance, the lack of local minima, and the sparse representation of the solution. It is due to the reason that support vectors are the sparse representations of the original training dataset. There are no local minima in SVR as it is a convex quadratic optimization problem that has linear constraints. The SVR performs better than Gaussian Process Regressor, which has multiple Gaussian curves, and the decision tree-based Random Forest Regressor. The comparison table for emotion recognition using ElderReact dataset is also prepared. Jannat et al. [53]considered a model which can be used to identify emotional expressions across the age and developed a Siamese network considering EmoReact and ElderReact datasets. A spatial Transformer Network viz, Face-STN was introduced by Barros et al. [54] which gave a result of 60.8, and this is because the model was trapped in the bias of age factor, which is not in the proposed system as it is able to overcome the local minima problem. Similar to our method CNN and LSTM-BLSTM networks are proposed by Hotterfield et al. [55]. But they applied only feature level fusion and the fusion results are tabulated in the Table 9. Lately, a self supervised Multi label Peer Collaborative knowledge Distillation based on multimodal transformer network was postulated, [56] which gave excellent results in spite of its complex structure compared to our network.

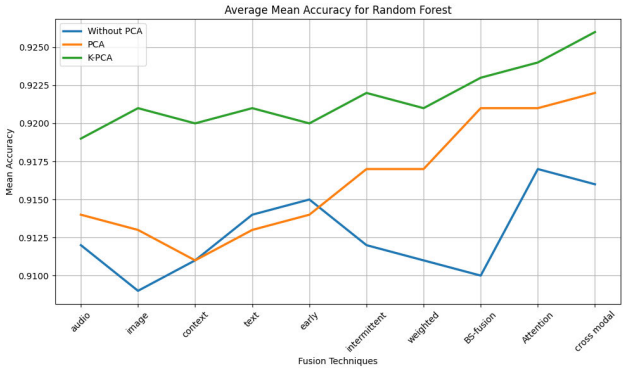
TABLE 9. Performance analysis with the state-of-the-art for ElderReact dataset.

Author	Technique	Acc(%)
1 Barros 2022 [54]	Facial image Transformers	60.8
2 Jannat 2021 [53]	Siamese Network	86
3 Hetterscheid 2020 [55]	Feature level fusion	71.1
4 SVP 2022 [10]	Intermediate level fusion	68
5 **proposed**	attention level fusion	**91.4**

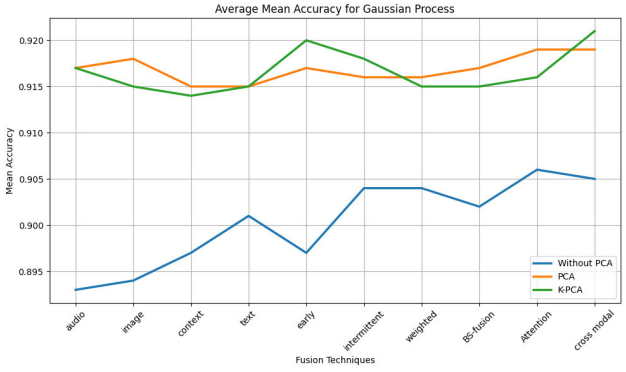
Next, objective is to compare the performance of our proposed pipeline on the state of the art networks for sentiment analysis. Table 8 summarizes the results on MOUD dataset and it outperforms the other approaches. Using multiple linear transformations, a multi head attention mechanism was introduced by Chen Li et al. and they performed a multimodal sentimental analysis which gave an accuracy of



(a) Support Vector Regressor

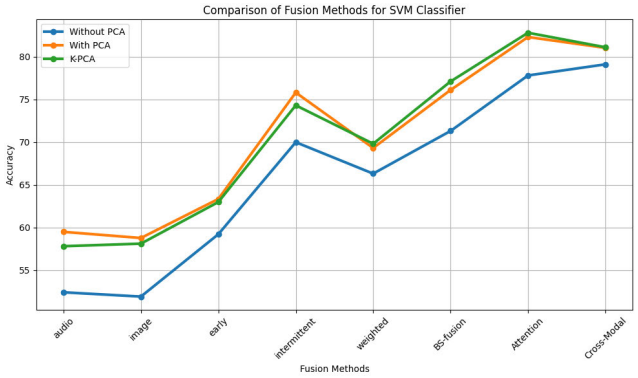


(b) Random Forest Regressor

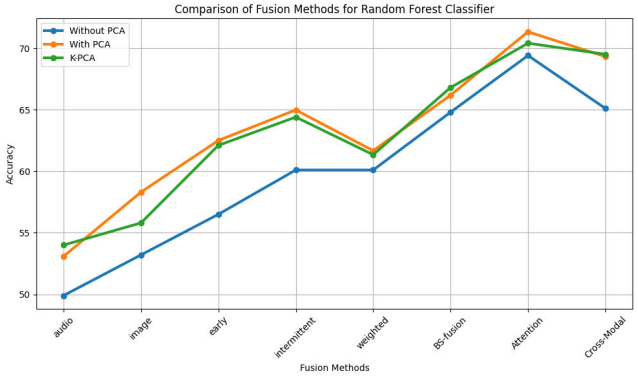


(c) Gaussian Progress Regressor

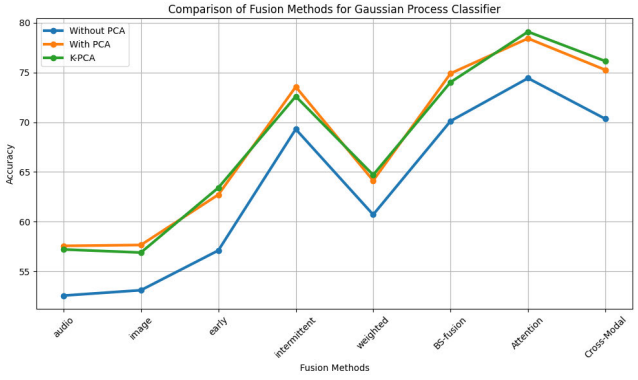
FIGURE 11. The comparison graph for dimensional reduction for personality trait analysis.



(a) Support Vector Regressor



(b) Random Forest Regressor



(c) Gaussian Progress Regressor

FIGURE 12. The comparison graph for dimensional reduction for emotion recognition.

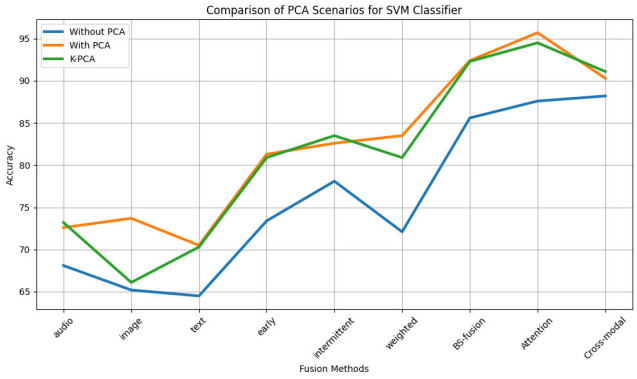
90.43% which is better than the rest of the results so far. Sun et al. introduced a gated approach with an inter-modality attention mechanism to perform modality interactions and filters inconsistencies from multiple modalities in an adaptive manner. The generated features where passed through various encoders with attentive mechanism, and they arrived at an accuracy and F1-score of 84.9%. We paid attention to the procedure for deep feature extraction also and the best results are achieved for Attention mechanisms.

V. ABLATION STUDY

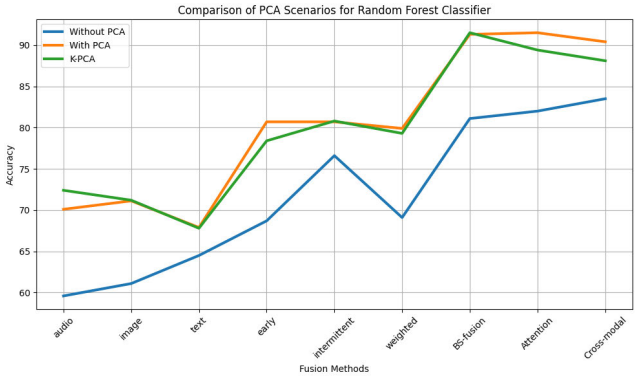
The change in the performance indices by the inclusion of dimension reduction is discussed in this section. Here, we did the comparison between the scores while applying PCA or K-PCA. It can be observed From Figure 11, 12 and Figure13 that the performance has increased by filtering the unnecessary features through dimensional reduction. Apart from that PCA and K-PCA are giving closer outcomes. It can be observed that in the case of personality trait prediction, SVR with K-PCA is giving the best results. In the case of GPR, PCA and K-PCA are giving similar results. While taking the case of emotion recognition, the improvement in the results can be observed from graphs. As it is already observed, except for 'surprise' SVM is performing better. It is maintained through out the emotions. Similarly in the case of sentiment analysis, SVM and Gaussian Process Classifier

TABLE 10. Performance analysis with the state-of-the-art for MOUD dataset.

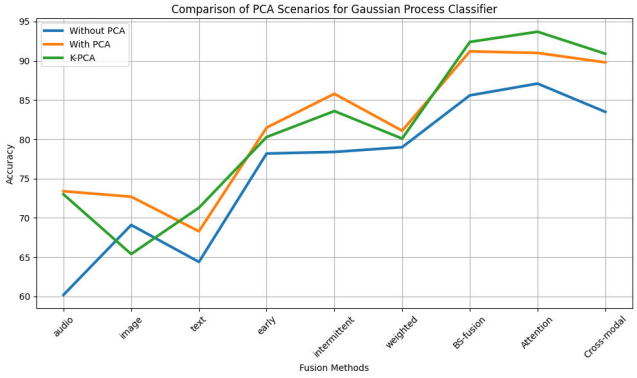
Author	Technique	Accuracy (%)	F1-score (%)
1 Xi 2020 [57]	Multimodal Multilevel attention	90.43	
2 poria 2015 [58]	Multi kernel fusion	88.6	
3 Li 2020 [59]	Feature level fusion	71.1	62.8
4 Svp 2020 [60]	Intermediate level fusion	84.8	84.8
5 Zadeh 2018 [61]	Memory Fusion	81.4	80.8
6 Chou 2019 [62]	Context aware fusion	82.07	82.04
7 Sun 2023 [62]	late fusion	84.9	84.9
8 **proposed**	attention multi level fusion	**96**	**95**



(a) Support Vector Regressor



(b) Random Forest Regressor



(c) Gaussian Progress Regressor

FIGURE 13. The comparison graph for dimensional reduction for sentiment analysis.

are offering similar performance, and for all the classifiers, the results improved with the reduced dimensions. Also, it is evident that there is no significant difference in performance between PCA and K-PCA.

VI. CONCLUSION

Through this work we tried to introduce fusion architectures with inter and intra attention mechanisms which operated on the intermediate features. The key characteristic of the finalized architecture is the consistency in the score while estimating all the personality awareness based tasks regardless of the biases like age, ethnicity or gender details. The proposed system could bring out the discriminate features by employing the attention based mechanisms and the outliers are filtered out by applying the dimensionality reduction techniques. Thus the proposed general architecture is suitable for applying on personality related tasks. The claim is proved through the results on various dataset. While considering the individual modalities, facial features were giving the maximum scores. The reason behind it is that the facial cues always contribute to the prediction of perceived personality and deep learning architectures are rich enough to capture these facial expressions and messages. When the generalized features are fused together, the personality traits and affect states could be analysed more closely than the individual modalities. Our technique consistently generates competitive performance across many multimodal fusion challenges in diverse datasets.Our method simplifies the parameters and also minimizes the complexity of the measurements. This is a novel approach to implement the attention mechanism in multimodal fusion, demonstrating improved efficiency and superior performance across many subsequent tasks.Incorporating the attention mechanism along with BS-Fusion enhances the classification capacities of our model, while maintaining a low parameter count and good efficiency.

In the future, the deep features may be extracted with standard transformer based techniques. Also, the context information can be extended to all the problem statement other than the personality trait prediction.

REFERENCES

[1] V. Kaushal and M. Patwardhan, "Emerging trends in personality identification using online social networks—A literature survey," *ACM Trans. Knowl. Discovery Data*, vol. 12, no. 2, pp. 1–30, Apr. 2018.

[2] Z. Lv, F. Poiesi, Q. Dong, J. Lloret, and H. Song, "Deep learning for intelligent human–computer interaction," *Appl. Sci.*, vol. 12, no. 22, p. 11457, 2022.

[3] S. Dhelim, N. Aung, M. A. Bouras, H. Ning, and E. Cambria, "A survey on personality-aware recommendation systems," *Artif. Intell. Rev.*, vol. 55, no. 3, pp. 2409–2454, Mar. 2022.

[4] F. Milella, C. Natali, T. Scantamburlo, A. Campagner, and F. Cabitza, "The impact of gender and personality in human–AI teaming: The case of collaborative question answering," in *Proc. IFIP Conf. Human-Computer Interact.* Cham, Switzerland: Springer, 2023, pp. 329–349.

[5] P. T. Costa Jr. and R. R. McCrae, *The Revised Neo Personality Inventory (NEO-PI-R)*. Newbury Park, CA, USA: Sage, 2008.

[6] E. Romero, P. Villar, J. A. Gómez-Fraguela, and L. López-Romero, "Measuring personality traits with ultra-short scales: A study of the ten item personality inventory (TIPI) in a Spanish sample," *Personality Individual Differences*, vol. 53, no. 3, pp. 289–293, Aug. 2012.

[7] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, and W. Zhang, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Inf. Fusion*, vols. 83–84, pp. 19–52, Jul. 2022.

[8] M. Umair, N. Rashid, U. S. Khan, A. Hamza, and J. Iqbal, "Emotion fusion-sense (Emo Fu-Sense)—A novel multimodal emotion classification technique," *Biomed. Signal Process. Control*, vol. 94, Aug. 2024, Art. no. 106224.

[9] D. Mamieva, A. B. Abdusalomov, A. Kutlimuratov, B. Muminov, and T. K. Whangbo, "Multimodal emotion detection via attention-based fusion of extracted facial and speech features," *Sensors*, vol. 23, no. 12, p. 5475, Jun. 2023.

[10] P. Sreevidya, S. Veni, and O. V. Ramana Murthy, "Elder emotion classification through multimodal fusion of intermediate layers and cross-modal transfer learning," *Signal, Image Video Process.*, vol. 16, no. 5, pp. 1281–1288, Jul. 2022.

[11] X. Zhao, Y. Liao, Z. Tang, Y. Xu, X. Tao, D. Wang, G. Wang, and H. Lu, "Integrating audio and visual modalities for multimodal personality trait recognition via hybrid deep learning," *Frontiers Neurosci.*, vol. 16, Jan. 2023, Art. no. 1107284.

[12] G. D. Salsabila and E. B. Setiawan, "Semantic approach for big five personality prediction on Twitter," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, pp. 680–687, Aug. 2021.

[13] S. Garg and A. Garg, "Comparison of machine learning algorithms for content based personality resolution of tweets," *Social Sci. Humanities Open*, vol. 4, no. 1, 2021, Art. no. 100178.

[14] H. Christian, D. Suhartono, A. Chowanda, and K. Z. Zamli, "Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging," *J. Big Data*, vol. 8, no. 1, pp. 1–20, Dec. 2021.

[15] P. Sreevidya, S. Veni, and V. R. M. Oruganti, "The role of face embeddings in classification of personality traits from portrait images," in *Proc. IEEE Region 10 Conf. (TENCON)*, Dec. 2021, pp. 845–850.

[16] K. Biswas, P. Shivakumara, U. Pal, and T. Lu, "A new ontology-based multimodal classification system for social media images of personality traits," *Signal, Image Video Process.*, vol. 17, no. 2, pp. 543–551, Mar. 2023.

[17] C. Suman, S. Saha, A. Gupta, S. K. Pandey, and P. Bhattacharyya, "A multi-modal personality prediction system," *Knowl.-Based Syst.*, vol. 236, Jan. 2022, Art. no. 107715.

[18] Y. Güçlütürk, U. Güçlü, M. A. J. van Gerven, and R. van Lier, "Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 349–358.

[19] H.-Y. Suen, K.-E. Hung, and C.-L. Lin, "TensorFlow-based automatic personality recognition used in asynchronous video interviews," *IEEE Access*, vol. 7, pp. 61018–61023, 2019.

[20] H. J. Escalante, H. Kaya, A. A. Salah, S. Escalera, Y. Güçlütürk, U. Güçlü, X. Baró, I. Guyon, J. C. S. J. Junior, M. Madadi, S. Ayache, E. Viegas, F. Gürpinar, A. S. Wicaksana, C. C. S. Liem, M. A. J. van Gerven, and R. van Lier, "Modeling, recognizing, and explaining apparent personality from videos," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 894–911, Apr. 2022.

[21] M. K. Tellamekala, T. Giesbrecht, and M. Valstar, "Dimensional affect uncertainty modelling for apparent personality recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2144–2155, Oct. 2022.

[22] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 48, Jun. 2016, pp. 1050–1059.

[23] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[24] G. Sogancioglu, H. Kaya, and A. A. Salah, "Can mood primitives predict apparent personality?" in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Los Alamitos, CA, USA, Sep. 2021, pp. 1–8.

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

[26] S. Aslan, U. Güdükbay, and H. Dibeklioğlu, "Multimodal assessment of apparent personality using feature attention and error consistency constraint," *Image Vis. Comput.*, vol. 110, Jun. 2021, Art. no. 104163.

[27] D. Giritlioğlu, B. Mandira, S. F. Yilmaz, C. U. Ertenli, B. F. Akgür, M. Kınıklıoğlu, A. G. Kurt, E. Mutlu, Ş. C. Gürel, and H. Dibeklioğlu, "Multimodal analysis of personality traits on videos of self-presentation and induced behavior," *J. Multimodal User Interfaces*, vol. 15, no. 4, pp. 337–358, Dec. 2021.

[28] T. A D and A. J, "Multimodal sentimental analysis using hierarchical fusion technique," in *Proc. IEEE 4th Annu. Flagship India Council Int. Subsections Conf. (INDISCON)*, Aug. 2023, pp. 1–8.

[29] N. Priyadarshini and J. Aravinth, "Emotion recognition based on fusion of multimodal physiological signals using LSTM and GRU," in *Proc. 3rd Int. Conf. Secure Cyber Comput. Commun. (ICSCCC)*, May 2023, pp. 1–6.

[30] H. Zhu, Z. Wang, Y. Shi, Y. Hua, G. Xu, and L. Deng, "Multimodal fusion method based on self-attention mechanism," *Wireless Commun. Mobile Comput.*, vol. 2020, pp. 1–8, Sep. 2020.

[31] M. S. Akhtar, D. S. Chauhan, and A. Ekbal, "A deep multi-task contextual attention framework for multi-modal affect analysis," *ACM Trans. Knowl. Discovery from Data*, vol. 14, no. 3, pp. 1–27, Jun. 2020.

[32] P. J. Corr and G. Matthews, *The Cambridge Handbook of Personality Psychology*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[33] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*, Lyon, France, Aug. 2013, pp. 148–152.

[34] B. W. Schuller et al., "The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates," 2021, *arXiv:2102.13468*.

[35] H. Choi, J. Kim, S. Joe, and Y. Gwon, "Evaluation of BERT and Albert sentence embedding performance on downstream NLP tasks," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 5482–5487.

[36] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[37] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "OpenFace: A general-purpose face recognition library with mobile applications," School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-16-118, 2016.

[38] F. Gürpinar, H. Kaya, and A. A. Salah, "Multimodal fusion of audio, scene, and face features for first impression estimation," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 43–48.

[39] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 960–964.

[40] M. Sushmitha, K. Suresh, and K. Vandana, "To predict customer sentimental behavior by using enhanced bi-LSTM technique," in *Proc. 7th Int. Conf. Commun. Electron. Syst. (ICCES)*, Jun. 2022, pp. 969–975.

[41] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, Sofia, Bulgaria, Aug. 2013, pp. 973–982.

[42] Z. Lian, Y. Li, J. Tao, and J. Huang, "Investigation of multimodal features, classifiers and fusion methods for emotion recognition," 2018, *arXiv:1809.06225*.

[43] S. Ji, Y. Xie, and H. Gao, "A mathematical view of attention models in deep learning," Texas A&M Univ., College Station, TX, USA, Tech. Rep., 2019.

[44] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.

[45] K. S. Naveenkumar, R. Vinayakumar, and K. P. Soman, "Amrita-CEN-SentiDB 1: Improved Twitter dataset for sentimental analysis and application of deep learning," in *Proc. 10th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2019, pp. 1–5.
[46] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[47] M. Seeger, "Gaussian processes for machine learning," *Int. J. Neural Syst.*, vol. 14, no. 2, pp. 69–106, 2004.
[48] A. G. Asuero, A. Sayago, and A. G. González, "The correlation coefficient: An overview," *Crit. Rev. Anal. Chem.*, vol. 36, no. 1, pp. 41–59, 2006.
[49] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, Dec. 2020.
[50] J. Camacho, A. K. Smilde, E. Saccenti, and J. A. Westerhuis, "All sparse PCA models are wrong, but some are useful. Part I: Computation of scores, residuals and explained variance," *Chemometric Intell. Lab. Syst.*, vol. 196, Jan. 2020, Art. no. 103907.
[51] S. Madan, M. Gahalawat, T. Guha, and R. Subramanian, "Head matters: Explainable human-centered trait prediction from head motion dynamics," in *Proc. Int. Conf. Multimodal Interact.*, New York, NY, USA, Oct. 2021, p. 435.
[52] S. Yan, D. Huang, and M. Soleymani, "Mitigating biases in multimodal personality assessment," in *Proc. Int. Conf. Multimodal Interact.*, New York, NY, USA, Oct. 2020, p. 361.
[53] S. R. Jannat and S. Canavan, "Expression recognition across age," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Dec. 2021, pp. 1–5.
[54] P. Barros and A. Sciutti, "Across the universe: Biasing facial representations toward non-universal emotions with the face-STN," *IEEE Access*, vol. 10, pp. 103932–103947, 2022.
[55] K. J. T. Hetterscheid, "Detecting agitated speech: A neural network approach," B.S. thesis, Dept. Comput. Sci., Univ. Twente, Enschede, The Netherlands, 2020.
[56] S. Anand, N. K. Devulapally, S. D. Bhattacharjee, and J. Yuan, "Multi-label emotion analysis in conversation via multimodal knowledge distillation," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 6090–6100.
[57] C. Xi, G. Lu, and J. Yan, "Multimodal sentiment analysis based on multi-head attention mechanism," in *Proc. 4th Int. Conf. Mach. Learn. Soft Comput.*, Jan. 2020, pp. 34–39.
[58] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2539–2544.
[59] H. Li and H. Xu, "Video-based sentiment analysis with hvnLBP-TOP feature and bi-LSTM," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 9963–9964.
[60] P. Sreevidya, O. V. R. Murthy, and S. Veni, "Sentiment analysis by deep learning approaches," *TELKOMNIKA Telecommunication Comput. Electron. Control*, vol. 18, no. 2, pp. 752–760, 2020.
[61] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 5634–5641.
[62] H. Sun, J. Liu, Y.-W. Chen, and L. Lin, "Modality-invariant temporal representation learning for multimodal sentiment classification," *Inf. Fusion*, vol. 91, pp. 504–514, Mar. 2023.

P. SREEVIDYA received the master's degree in control systems from the PSG College of Technology, India, in 2015. She is currently pursuing the Ph.D. degree with the Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham, India. Her research interests include multi-modal interactions and deep learning for applications in affective computing and recommendation systems.

J. ARAVINTH (Member, IEEE) received the B.E. degree in electronics and communication from Periyar University, in 2004, and the M.E. degree in applied electronics and the Ph.D. degree from Anna University, Chennai, in 2007 and December 2017, respectively. He is currently an Associate Professor with the Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham, India. His Ph.D. thesis was design of single and hybrid classifier-based score level fusion for multimodal biometric recognition. He is currently working on hyperspectral image compression using compressed sensing and deep learning. He has authored around 50 technical papers in reputed conferences and journals. His research interests include multimodal fusion architectures, hyperspectral image analysis, remote sensing, signal processing, and machine learning.

SATHISHKUMAR SAMIAPPAN received the master's degree from Amrita Vishwa Vidyapeetham, Tamil Nadu, India, and the Ph.D. degree in electrical and computer engineering from Mississippi University. He is currently an Associate Research Professor with Mississippi State University, USA. He is also persuading the research in remote sensing and image analysis. His research interests include machine learning, hyperspectral image analysis low altitude remote sensing, and drone remote sensing.