Computers, Materials & Continua
DOI: 10.32604/cmc.2023.028333
Article

Deep Bimodal Fusion Approach for Apparent Personality Analysis

Saman Riaz[1], Ali Arshad[2], Shahab S. Band[3,*] and Amir Mosavi[4]

[1]Department of Computer Science, National University of Technology, Islamabad, 44000, Pakistan
[2]Department of Computer Science, Institute of Space Technology, Islamabad, 44000, Pakistan
[3]Future Technology Research Center, National Yunlin University of Science and Technology, Douliu, 64002, Yunlin, Taiwan
[4]Faculty of Civil Engineering, Technische Universitat Dresden, Dresden, 01069, Germany
*Corresponding Author: Shahab S. Band. Email: shamshirbands@yuntech.edu.tw
Received: 08 February 2022; Accepted: 31 March 2022

Abstract: Personality distinguishes individuals' patterns of feeling, thinking, and behaving. Predicting personality from small video series is an exciting research area in computer vision. The majority of the existing research concludes preliminary results to get immense knowledge from visual and Audio (sound) modality. To overcome the deficiency, we proposed the Deep Bimodal Fusion (DBF) approach to predict five traits of personality-agreeableness, extraversion, openness, conscientiousness and neuroticism. In the proposed framework, regarding visual modality, the modified convolution neural networks (CNN), more specifically Descriptor Aggregator Model (DAN) are used to attain significant visual modality. The proposed model extracts audio representations for greater efficiency to construct the long short-term memory (LSTM) for the audio modality. Moreover, employing modality-based neural networks allows this framework to independently determine the traits before combining them with weighted fusion to achieve a conclusive prediction of the given traits. The proposed approach attains the optimal mean accuracy score, which is 0.9183. It is achieved based on the average of five personality traits and is thus better than previously proposed frameworks.

Keywords: Apparent personality analysis; deep bimodal fusion; convolutional neural network; long short-term memory; bimodal information; fusion approach

1 Introduction

Personality significantly affects peoples' lives and impacts hone, preferences, and choices [1]. Emotions have exciting parts in making decisions, such as sharing information about pain and happiness, empowering quick decisions under time pressure, and producing commitment concerning choices. Furthermore, research recommends that the decision-making process for humans can be modeled as two frameworks, consisting of emotional and rational frameworks [1,2]. Therefore, emotions are part of each decision-making process rather than simply affecting these processes. In

like manner, personality also has a substantial impact on decision making, and it causes individual contrasts in peoples' considerations, sentiments, and inspirations. There are imperative relationships between connection styles, decision-making styles, and personality characteristics [3]. In this way, it becomes necessary to accurately perceive peoples' personalities and feelings (emotions).

In previous research, some deep bimodal approaches [4] discussed DAN and DAN+ on facial features extracted from video and linear regression on audio features. Also, in [5], DAN deep neural network (DNN) is discussed. In [1] multimodal approach is explored and proposed with the Resnet, Visual Geometry Group (VGG), and Long Short-Term Memory Networks. In [6], Deep LSTM is used for the whole video features learning.

We present a deep bimodal fusion (DBF) approach to recognize an individual's evident personality from videos to touch on the issue and improve the results compared to the previous research discussed before. Fig. 1 shows the architecture of the DBF framework.

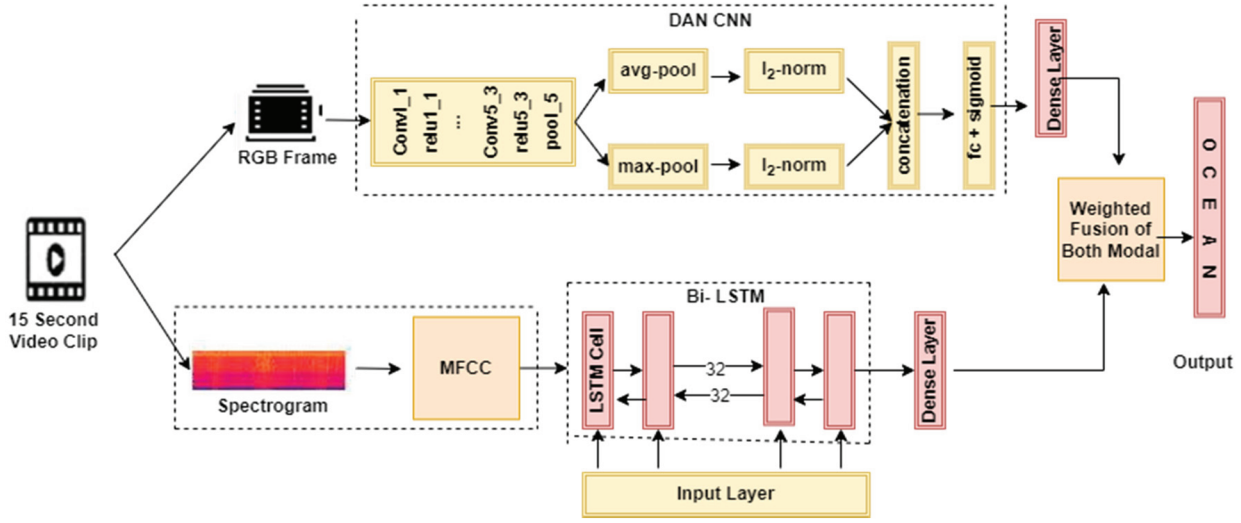

Figure 1: Architecture of the Deep Bimodal Approach

The main motivation of deep bimodal fusion (DBF) to develop a modal which will automatically decode a human behaviour by their micro expressions based on the bahaviour ecology view of facial displays (BECV). Our proposed modal will perform more efficiently as compared to state-of-the-art algorithms.

Personality Traits

A persons' personality can be characterized as the psychological factor that impacts a person's reasoning and feeling that separates the person from each other. The Five-Factor Model (FFM) is the most normal Fand comprehensively recognized framework for character among brain science analysts. This model is dependent on factors that describe a human's personality, with five estimations for an overall personality portrayal shown in Table 1. The big five traits are explained as follows:

1. Openness (O): Openness indicates a person's inclination to try new things and think out of the box.
2. Conscientiousness (C): Conscientiousness shows a persons' ability to control impulses and maintain goal-directed behavior.
3. Extraversion (E): Extraversion describes the tendency of a person to socialize.

4. Agreeableness (A): Agreeableness indicates how someone manages their relationships.
5. Neuroticism (N): Neuroticism depicts the generally enthusiastic security of a person through how they see the world [7].

Table 1: Characteristics of personality traits

Low score	Traits	High traits
Predictable, Prefer Routine, Traditional Down-To-Earth, Non-Analytical	Openness	Imaginative, Creative, Unconventional, Untraditional
Incompetent, Careless, Impulsive, Lax	Conscientiousness	Competence, Organized, Self Disciplined, Neat, Hard Working
Prefers Solitude, Reflective, Dislike, Sober, Aloof	Extraversion	Forgiving, Helpful, Opinions, Sociable, Energized
Manipulative, Rude, Irritable, Quickly And Confidently Asserts Own Rights	Agreeableness	Agrees With Other About Political Opinions, Good Natured, Forgiving,
Calm, Not Getting Irritated By Small Annoyances, Secure	Neuroticism	Constantly Worrying About Little Things, Insecure, Hypochondriacal, Anxious

2 Literature Review

We have thoroughly studied previous research conducted on the subject matter–sound representations, apparent personality analysis, and visual-based deep learning for better understanding and relevance.

As of late, for picture related undertakings, Convolutional Neural Networks (CNN) [8] permit computational models that are made out of several processing layers to better understand picture portrayals with different degrees of reflection, which is accepted as a powerful grade of models.

To comprehend frame content, providing cutting edge results on picture acknowledgment, division, recognition, and recovery. In particular, the CNN modular comprises a few convolutional layers and pooling layers piled up on the top of one another [5]. The convolution layer shares an enormous number of loads. The pooling layer sub-examples the yield of the convolution layer, which decreases the rate of information from the layer beneath. The load partaking in the convolutional layer, alongside suitable pooling plans, supplies the CNN with certain invariance properties (e.g., interpretation invariance).

In this paper, the DBF approach uses different CNN's to get familiar with the picture portrayals for the visual methodology. A while later gets the Big Five Traits forecasts by start to finish preparing.

In recent years, many sound representations have been proposed, including time area highlight and frequency domain features. A Few of the most notable sound highlights from them are Mel Frequency Cepstral Coefficients (MFCC) [9], Linear Prediction Cepstral Coefficients (LPCC) [10], and Bark Frequency Cepstral Coefficients (BFCC) [9]. Especially, the MFCC highlights have been generally utilized in the discourse acknowledgment local area [10]. MFCC alludes to momentary unearthly based highlights of a sound, which is gotten from the spectrum-of-a-spectrum of a brief snippet. It can be resolved in four stages. During these stages, the log channel bank (log bank) highlights can likewise be gotten. In our DBF structure, we separate the spectrogram and MFCC highlights from

the sound of each unusual human-centered video for American Psychological Association (APA). We utilize LSTM to become familiar with the picture portrayals (MFCC) for the sound methodology, and afterward, get the Big Five Traits forecasts by start to finish preparing.

Personality examination is an assignment that is particular to the field of psychology research. Beforehand, personality investigation requires brain science researchers to decipher the discoveries or members to finish explicit tests containing a tremendous survey that could mirror their personality. However, this technique will require a significant amount of time and money. A task similar to personality analysis in computer vision is emotion analysis [2,11]. The Emotion analysis is a different class classification study, which typically regards four feelings (happiness, sadness, anger, and neutral state) perceived by the calculation. On the other hand, in apparent personality analysis, what's predicted is the Big Five Traits (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism). These are independent of each other, and their scores are continuous values, ranging from 0 to 1. However, personality analysis tasks are realistic, comparatively harder than those relating to emotion analysis.

3 Proposed Methodology

This section introduces the deep bimodal fusion approach for the task at hand-apparent personality analysis. Our modal has three primary parts: visual modality regression, sound component learning, and ensemble process.

3.1 Visual Modality

There are three subparts: image extraction, Modal training, and score prediction in visual modal.

3.1.1 Image Extraction

Inputs of the convolutional neural network are frames or images, and for the apparent personality analysis, the input is human-focused, having facial highlights frames from videos. Thus, the Extraction of images/frames from videos is necessary. The dataset used in evaluating the proposed approach is the Cha Learn First Impressions V2 (CVPR'17) challenge dataset [12], containing a fifteen-second-long video with a frame rate of 30 fps and about 450 frames/images from every unique video. We extracted the best 100 frames from each video [5]. Then we labeled these extracted frames with personality traits values [0–1]. Based on these pictures/frames, we can prepare our modal (CNN). Fig. 2 shows the acquisition of images from videos.

Algorithm 1: Frame extraction from video

Input: HD videos in which people facing camera and speaking in English language.
Output: Frame/images having facial and ambiance feature
1. Import os, cv2 and numpy
2. Use cv2.VideoCapture(<path of video>) to read video in python.
3. Set frame limit 100 using cap.set (cv2.CAP_PROP_FRAME_COUNT, 101)
4. Check if file is opened if yes then start reading frames.
5. Then resize them using function cv2.resize(frame, (256, 256), interpolation=cv2.INTER_CUBIC) where frame size is 256 × 256 and interpolation flag.
6. Store every frame within the dataset folder using naming convention frame<i>.jpg using function cv2.imwrite(<name of file>, <frame>).
7. Continue to step 5 until videos are processed.

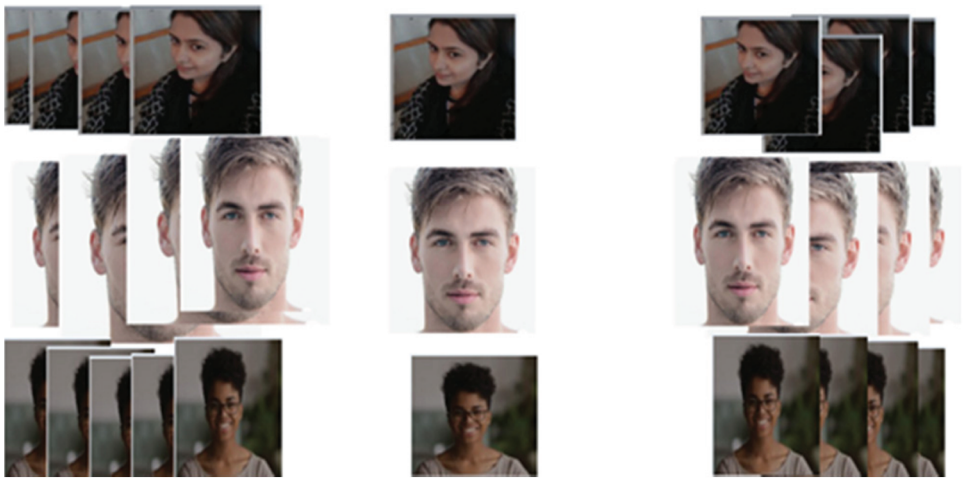

Figure 2: Acquisition of images from videos

3.1.2 Visual Modal Training

The DBF framework of visual modality, modifies the deep CNN model for fine-grained picture recovery by using selective convolutional descriptor aggregation [13]. For fine-grained image recognition, Mask-CNN is used for localizing parts and selecting descriptors [14], called Descriptor Aggregation Networks (DAN's). The main modification from the traditional CNN to DAN is disposed of the fully connected layers. After the final convolutional layers (Pool5), both average-pooling and max-pooling are replaced. Meanwhile, each pooling operation is followed by the standard '2-normalization. The two 512-d feature vectors are merged to form the final image representation. As a result, in DAN, the deep descriptors of the last convolutional layers are combined into a single visual feature. Finally, an end-to-end training regression (fc + sigmoid) layer is added because APA is a regression problem. DAN's architecture is illustrated in Fig. 3. Since DAN does not feature fully connected layers, it can provide several advantages, including reduced dimensionality of the final feature, reduced model size, and faster model training [15]. Furthermore, DAN's model performance outperforms traditional CNN with fully connected layers'fc sigmoid. We use the pre-trained VGG-Face model [16] to initialize the convolutional layers in our DAN's in the proposed DBF framework experiments.

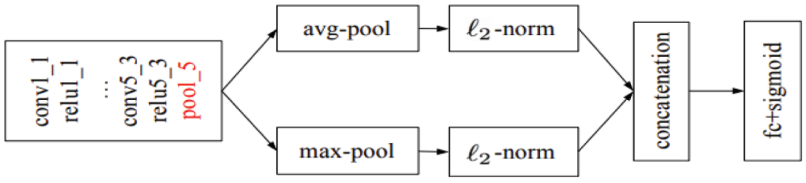

Figure 3: Design of the DAN model

In DAN architecture, we eliminated the fully connected layers. The last convolutional layer is collected by both max- and average-pooling and then connected to regression's final image representation.

Algorithm 2: Training visual modal on modifies CNN (DAN)

Input: Frames of size $256 * 256 * 3$

Output: The weight parameters in .pkl files for number of EPOCHS and BATCH.

1. Import tf, cv2, numpy,
2. Define hyper parameters that are as follow
   Learning rate = 0.5, Batch size = 25, Epochs = 2
3. Initialize placeholder.
4. Create tf session using tf.sessiontf.Session(config=config).
5. Import model and initialize with preprocessed vggface model.
6. Initialize optimizer, we used Adam optimizer and pass it with learning rate as mentioned above using tf.train.AdamOptimizer(learning_rate=LEARNING_RATE).
7. Read image data and labels from tf record reader with function tf. TFRecordReader() and decode it.
8. Shuffle the batch with function tf.train.shuffle_batch().
9. Run the tf session and store the weight parameters in .pkl files for number EPOCHS and BATCH.
10. In last save the whole session and dump it to files to use in prediction using tf.train.Saver().

3.1.3 Personality Traits Prediction

Frames are extracted from each testing video during foreseeing regression values. Then, the predicted regression scores of images are returned using the trained visual models. After that, as the predicted scores of a video, we generated an average of the images attained.

3.2 Audio Modality

In the sound methodology of DBF structure, we separated sound records from every video. Afterward, we down scaled the input Audio samples to a uniform sampling rate of 8 kHz before entering them into the model. The sound information is converted into PNG design pictures or essentially extricate the Spectrogram for each Audio. For every audio file, we used librosa python library to extract Spectrogram.

Algorithm 3: Extracted Audio from video

Input: HD videos

Output: Extracted Audio from every video for audio processing having transcript and audio feature.

1. Import moviepy.editor.
2. Read mp4 file with videoclip = VideoFileClip(<path>)
3. Get wav file from mp4 using function audioclip = videoclip.audio
4. Repeat step 2 and 3 for all videos.

The extracted features are subjected to a simplified encoding process and transferred to the LSTM network as input. The label values are also provided, and the network trains on input data based on the labels. We proposed LSTM units that capture the sequential patterns of the input data to predict the big five personality traits.

3.3 Modality Ensemble

Modality ensemble is the fusion approach used to obtain the final result after preparing the visual and sound modalities. The ensemble procedure we use in DBF is a simple yet successful averaging strategy. The predicted result of a trained regressor in APA is a five-dimensional vector addressing the

Big Five Traits values, i.e., $si = (si_1, si_2, si_3, si_4, si_5)$. We give equal weight to each of these two modalities predicted outcomes. For example, the visual modality's predicted results are si_1, si_2, and si_3, while the audio modality's predicted results are si_4, and si_5. The last ensemble results are determined as follows:

$$\text{Result} = \frac{\sum_{i=1}^{5} S_i}{5} \tag{1}$$

4 Experiments

The dataset for the personality prediction is described first in this section with a detailed description of our proposed approach. Finally, we present and discuss the experimental results.

4.1 Dataset and Evaluation Criteria

The Cha Learn First Impressions V2 (CVPR'17) challenge dataset [17], which is easily accessible, is used to evaluate the proposed approach. This challenge aims to recognize apparent personality traits as indicated by the five-factor. This challenges' dataset consists of 10000 recordings of people confronting and addressing a camera. The high-quality videos (1280720 pixels) are extracted from YouTube, with an average duration of 15 s and 30 frames per second. Individuals within the recordings conversation to the camera in a self-presentation setting, and there's a wide range of age, identity, gender, and nationality. In expansion, the RGB and sound data are given, in addition to continuous ground-truth values for each of the 5 Enormous Five Traits clarified by Amazon Mechanical Turk laborers.

The dataset was split into 6,000 videos from the training set, 2000 videos from the validation set, and 2000 videos from the evaluation set. Throughout the development stage, we train the visual and audio models of DBF framework on the training set and validate its performance on the approval/validation set.

Given video and the values of the comparing traits, the accuracy is the absolute distance. between ground truth and predicted values. The mean accuracy among all the Big Five traits' value is determined as the principal quantitative measure:

$$Mean\ accuracy = \frac{1}{5N}\sum_{j=1}^{5}\sum_{i=1}^{N} 1 - |ground\_truth_{i,j} - predicted\_value_{i,j}| \tag{2}$$

where N is the total number of predicted values.

4.2 Implementation Details

The implementation detail of our proposed DBF framework is described in this section.

4.2.1 Visual Modality

As mentioned above, we first extracted frames/images from each video in the visual modality. We have used cv2 library for the video to image acquisition. For the most part, 100 frames are extracted from most videos (about 6.1 fps). After that, we resize these frames to a resolution of $224 * 224 * 3$. As a result, 560,393 images from training videos, 188,561 images from validation videos, and 188,575 images from testing videos have been extracted. The open-source library MatConvNet [17] implements the visual DAN models in our experiments. In addition to the DAN models, we use a widely used deep convolutional network. The learning rate is 103 in the training stage. For all visual models, the weight decay is 1 to 5104, and the momentum is 0.9 [15].

4.2.2 Audio Modality

We use moviepy, editor to extract the Audio features from the original video. The waveform of sampled Audio is shown in Fig. 4. We begin by converting the audio files into PNG images (Spectrogram). To extract Spectrogram for each audio file, we use the librosa python library. We extract meaningful features from these spectrograms, such as MFCC, Spectral Centroid, Zero Crossing Rate, Chroma Frequencies, and Spectral Roll-off. Then, save NumPy to a csv file to use the deep learning algorithm. The extracted features are subjected to a simplified encoding process and transferred to the LSTM network as input. The label values are also provided, and the network is trained on input data based on the label categories.

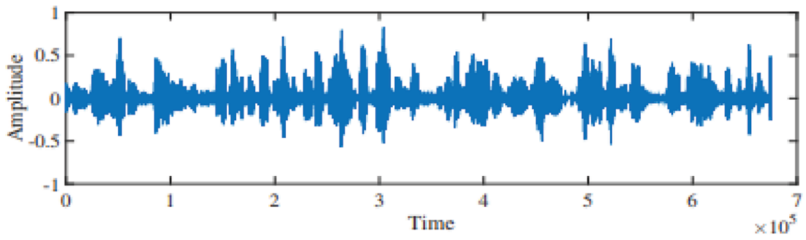

Figure 4: Audio spectrogram waveforms of a sampled audio signal

The horizontal axis represents the time in audio spectrogram waveforms of a sampled audio signal. Due to the sampling frequency of 44,100 Hz, the horizontal axis unit is 1/44100 s. The amplitude is measured on the vertical axis. The audio features are provided as input to the LSTM input layer, and the models corresponding to the audio signals are created and stored in the world dictionary. The LSTM consists of 32 units, each unit has a sigmoid and activation function. In the end, we use the SoftMax activation function. After training with extracted features, the LSTM network acts as a classifier, estimating the most likely sentence by classification. As a result, this neural network composition recognizes personality from sound. The architecture of this sub-network is depicted in Fig. 5.

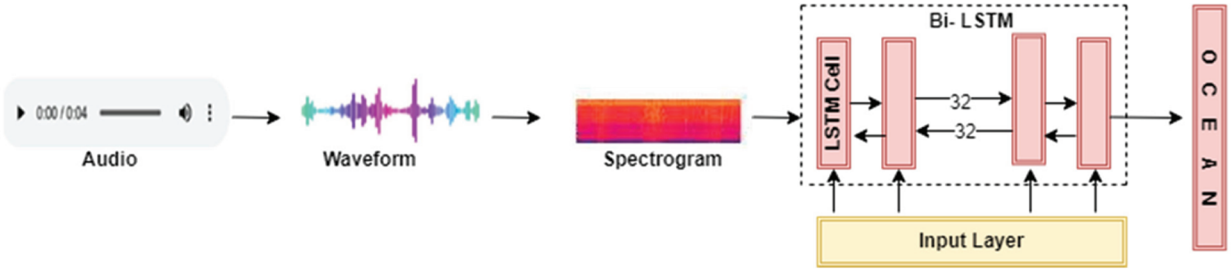

Figure 5: Audio feature-based neural network

4.3 Modal Evaluation

In the development stage, the major results of the visual and audio modalities are presented in Table 2 and Fig. 6.

Table 2: Comparison of regression mean accuracy in the development phase

Modality	Model	Param	Dim	Epoch 1	Epoch 2	Epoch fusion
Visual	VGG-Face	134.28 M	4096	0.9065	0.9060	0.9072
	ResNet	58.31 M	512	0.9072	0.9063	0.9080
	DAN-DBF)	14.71 M	1024	0.9080	0.9080	0.9100
Audio	Linear Regression	0.40 M	79534	0.8900		0.8900
	LSTM-DBF	0.80 M	79534	0.9021		0.9021

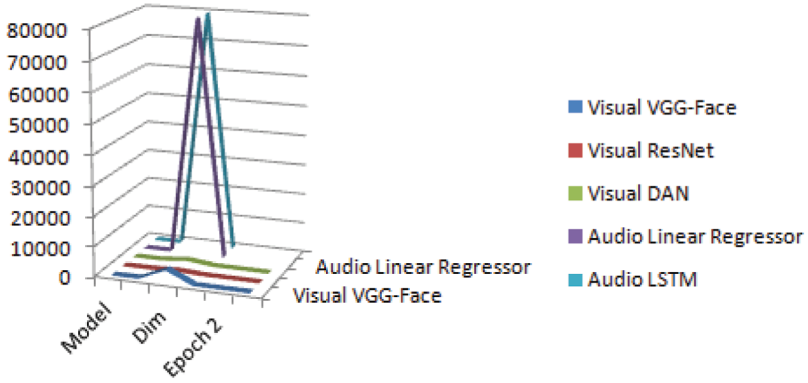

Figure 6: Graphical representation of various modalities

In Table 3, optimal weights for combining the two modalities were discovered for each trait. Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism are represented by the letters O, C, E, A, and N. Weights learned by our proposed framework for audio and visual modalities against personality five factors are shown.

Table 3: For each trait, optimal weights for combining the two modalities were discovered

Big five personality traits					
Modal	O	C	E	A	N
Video (DBF)	0.9203	0.9184	0.916	0.8996	0.9023
Audio (DBF)	0.8924	0.9021	0.9045	0.9124	0.9002

4.4 Experiment Results

In the concluding evaluation stage, we utilize the best models on the testing set from the development stage to predict the Big Five Traits values. Furthermore, our proposed DBF approach outperformed the state-of-the-art approaches. Due to the utilization of a simple weighted average method for the fusion, the proposed method demonstrates the improved performance by advanced ensemble methods, such as stacking, which is utilized to learn the appropriate weights for the late fusion. Deep audio networks must be tested to learn more discriminative representations for Audio.

We can use an ensemble of different sound models to improve personality analysis performance in our DBF framework.

The previously proposed method means and individual trait accuracies are shown in Table 4. Having traditional CNN by Evolgen [18] secured a mean accuracy 0.9134, PML [19] has achieved mean accuracy of 0.9170 using 5 support vector regressors. NJU-LAMBDA [5] have achieved 0.9130 using linear regressor and traditional CNN. Our proposed framework DBF achieved 0.9183 mean accuracy, which is better than previous methodologies, shown in Table 4 and Fig. 7.

Table 4: Comparative analysis of mean accuracy and traits accuracy

Method	Mean accuracy	O	C	E	A	N
Evolgen [16]	0.9134	0.9130	0.9136	0.9145	0.9157	0.9098
PML [17]	0.9170	0.9169	0.9166	0.9206	0.9161	0.9149
NJU-LAMDA [5]	0.9130	0.9123	0.9166	0.9133	0.9126	0.9100
DBF (Proposed Method)	**0.9183**	**0.9190**	**0.9178**	**0.9182**	**0.9203**	**0.9175**

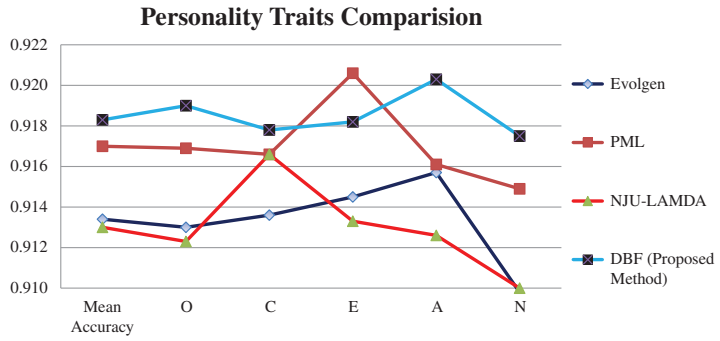

Figure 7: Graphical representation of personality traits of different modalities

5 Conclusions

This research paper presents a novel DBF framework, to capture and exploit significant features from both auditory and visual modalities. This paper proposes a Deep Bimodal Approach (DBF) for apparent personality traits from videos. The proposed approach uses bimodal deep learning with a weighted fusion. The proposed approach uses modality-specific CNN for extracting spatial features such as ambient features and facial expressions from video in the form of frames. Moreover, the LSTM network is used for extracting audio features. Our study used a labelled dataset named First Impressions V2 (CVPR'17). The existing frameworks were inadequate in achieving high accuracy with audio and visual modalities. The Evaluation of the proposed approach proves that we have achieved high accuracy compared to existing approaches by including maximum recordings features.

Future research directions include investigating the relationship between character, body movements, pose, eye-stare, and feeling/emotion to improve the performance provided by surrounding/ambient, face, audio, and transcription features.

Acknowledgement: We would like to thank Sir Syed CASE Institute of Technology for their resources and help throughout the development of this project, and our time in gaining the knowledge and tools we would need to succeed in the professional world.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: S.R. Conceptualization and Methodology; A.A. Software, Writing—review & editing; S.S. funding acquisition; and A.M. Editing.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Aslan and U. Gudukbay, "Multimodal video-based apparent personality," arXiv, 2019. [Online]. Available: https://arxiv.org/abs/1911.00381.

[2] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee *et al.,* "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. of the 6th Int. Conf. on Multimodal Interfaces (ICMI '04)*, New York, NY, USA, ACM, pp. 205–211, 2004.

[3] F. Valente, S. Kim and P. Motlicek, "Annotation and recognition of personality traits in spoken conversations," in *Idiap Research Institute*, CH-1920 Martigny, Switzerland, INTERSPEECH 2012, pp. 1–4, 2012.

[4] C. -L. Zhang, H. Zhang, X.-S. Wei and J. Wu, "Deep bimodal regression," in *European Conf. on Computer Vision*, Netherlands, pp. 14, 2016.

[5] X. Wei, C. Zhang, H. Zhang and J. Wu, "Deep bimodal regression of apparent personality traits from short video sequences," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 303–315, 2018.

[6] Y. Karen and G. Noa, "Prediction of personality first impressions with deep bimodal LSTM," *Technical Report, arXiv*, 2017. [Online]. Available: http://cs231n.stanford.edu/reports/2017/pdfs/713.pdf.

[7] Andrew Grauer, "5_personality-Conscientiousness: Conscientiousness," 2020. [Online]. Available: https://www.coursehero.com/file/84088956/5-personality/.

[8] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional," *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, vol. 1, pp. 8–9, 2012.

[9] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[10] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[11] S. Koelstra, C. Muhl, M. Soleymani, JS. Lee, A. Yazdani *et al.,* "Deap: A database for emotion analysis; Using physiological signals," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 18–31, 2011.

[12] V. Ponce-L'opez, B. Chen, M. Oliu, C. Corneanu, A. Clap'es *et al.,* "ChaLearn LAP 2016: First round challenge on first impressions-dataset and results," in *Lecture Notes in Computer Science*, Springer International Publishing: Switzerland, pp. 400–418, 2016.

[13] X. Wei, J. Luo, J. Wu and Z. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2868–2881, 2017.

[14] X. S. Wei, C. W. Xie and J. Wu, "Mask-CNN: Localizing parts and selecting descriptors for fine-grained image recognition," *arXiv*, 2016.

[15] M. Hasan and A. K. Roy-Chowdhur, "Continuous learning of human activity models," in *Proc. of the European Conf. on Computer Vision*, Zurich, pp. 15–16, 2014.

[16] O. M. Parkhi, A. Vedaldi and A. Zisserman, "Deep face recognition," in *British Machine Vision Conf.*, Seansea, UK, pp. 11–12, 2015.

[17] A. Vedaldi and K. Lenc, "MatConvNet-convolutional neural networks for MATLAB," in *ACM Int. Conf. on Multimedia*, Brisbane, Australia, pp. 689–692, 2015.
[18] A. Subramaniam, V. Patel and A. Mishra, "Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features," in *European Conf.*, Netherlands, pp. 337–348, 2016.
[19] S. E. Bekhouche, F. Dornaika, A. Ouafi and A. Taleb-Ahmed, "Personality traits and job candidate screening via analyzing facial videos," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, Hawaii, USA, pp. 1660–1663, 2017.