INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

A Review of Convolutional Neural Network Model for Audio-Visual Features Extraction in Personality Traits Recognition

Nurrul Akma Mahamad Amin[1*], Nilam Nur Amir Sjarif[2] & Siti Sophiayati Yuhaniz[3]
Universiti Teknologi Malaysia,
Kuala Lumpur, Malaysia
Email: nurrulakma@graduate.utm.my[1]; nilamnur@utm.my[2]; sophia@utm.my[3]

Submitted: 17/10/2024. Revised edition: 27/3/2025. Accepted: 11/4/2025. Published online: 27/5/2025
DOI: https://doi.org/10.11113/ijic.v15n1.498

***Abstract*—In the field of personality computing research, the analysis on audio-visual input is predominantly used to detect human personality behaviors. With the advancement of computer vision technology, there has been significant enhancement in personality computing. Personality trait recognition is one of the applications under personality computing where the machine can analyze human behaviors and recognize personality traits via video analysis. In video input, there are different audio-visual features characteristics, consisting of visual (images) and audio (sounds) elements. Therefore, it is critical to employ appropriate deep learning techniques to most effectively extract important features from audio-visual input. The maturity of convolutional neural networks (CNNs) has been proven with promising prediction results for feature extraction and selection in image classification, sound detection, and face-emotion recognition. Thus, a variety of CNN-based techniques have been developed with different salient features and CNN layer modifications to learn and extract meaningful patterns and representations from images and videos. Due to the distinct characteristics of audio-visual features, hybrid CNN-based techniques were introduced to optimally analyze these modalities. This study aimed to explore hybrid CNN-based techniques used in video analysis for personality trait recognition systems. This study also provides an overview of current issues in the development of recognition models in personality computing using hybrid CNN-based techniques. The advantages of integrating audio and visual modalities in hybrid techniques are addressed, as well as their performance accuracy. The discussion finally summarizes the findings and potential future research directions.**

***Keywords*—Computer Vision, Personality Computing, Personality Traits Recognition, Audio-Visual Feature Extraction**

I. INTRODUCTION

Personality is defined as the stable characteristics of an individual where it is related to the characteristics of people that determine a person's personality and self-quality. It also portrays a distinctive way of thinking, feeling, and behaving [1]. Some examples include being outgoing, confident in oneself, talkative, friendly, shy, or lacking in self-confidence. Personality causes one person to be different from another in terms of thoughts, feelings, and behaviours that make a person unique. The way a person acts or responds to some stimuli reflects his or her personality. In psychology, personality models are used to describe human personality through structured traits. These traits are different according to the model and they are used to describe a person's personality. Existing research mentions that several personality models have been widely used for automatic personality detection using machine learning approaches [2]. These personality models include Myers-Briggs Type Indicator (MBTI) [3], The Sixteen Personality Factor Questionnaire (16PF) [4], Eysenck Personality Questionnaire-Revised (EPQ-R) [5], Three Traits Personality Model (PEN) [6] and Five Factor Model [7]. The interpretation of human personality represented by each of these models differs from each other. According to psychology realm, the Five Factor Model, often known as Big-5 Model, is the most widely used predictors of personality traits measured in psychology as well as affective computing domain [8]. In the Big-5 Model, human personality is described by five factors traits which are openness, conscientiousness, extraversion, agreeableness, and neuroticism. Additionally, these traits are known by the acronym OCEAN and are considered as basic traits that describe a person's personality.

Traditionally, personality testing surveys have been used as a tool to systematically measure human personality. The survey form is designed differently depending on various personality models. However, recognizing people's personality using personality testing based solely on questionnaires with closed-ended questions is insufficient. Traditional approach of personality testing using Likert scale questionnaires is open to self-distort and faking [9-10].

Furthermore, using the solely questionnaire with closed-ended question in personality tests to predict personality traits among job candidates is inadequate and not comprehensive. This traditional way is also no longer up-to-date and economical due to the advancement of current technology. The reliability of personality tests is also questionable because there is a possibility that different personality types will be revealed even when a person takes the same test repeatedly [2]. Thus, to overcome problems in the traditional approach of personality assessment, automatic personality recognition has been introduced. The ChaLearn Looking at People Challenge 2016 (ECCV2016) was a pioneering dataset for automatic personality recognition [11]. The dataset contains subjects or people talking in front of a camera for 15 seconds, focusing on first impression-based personality analysis using very short video input. However, several other datasets have also been used in personality trait recognition research, allowing a broader analysis of CNN capabilities. For example, Hemamou *et al*. (2019) have developed their own dataset of French asynchronous video interviews, which consists of only 7938 candidates applying for 475 sales positions [12]. On the other hand, Suen *et al*. (2019) developed asynchronous video interview (AVI) software, involving 120 participants from the human resources field, with a video duration of 20 minutes each [13]. The Self presentation and Induced Behavior Archive for Personality Analysis (SIAP) dataset integrates facial action units, head pose, gaze and transcribed speech for a multimodal approach to personality recognition [14]. These datasets provide varying perspectives on CNN-based personality trait recognition, enabling a better understanding of feature extraction effectiveness across different environments.

Due to certain restrictions, a self-collected data set was not publicly available for further analysis and research. Most previous studies used ChaLearn dataset and focus on developing a model to recognize a person's personality traits based on features extracted from short video-clip including combination of local face cue and audio signal [15-16], added ambient background features and text transcriptions [17-18] and added gaze distribution [19]. This study reviews research papers published between 2016 and 2023 from academic databases such as IEEE Xplore, Web of Science, Scopus, SpringerLink, ScienceDirect, and Google Scholar. The search was conducted using keywords such as 'personality computing', 'personality traits recognition', 'audio-visual feature extraction,' and 'CNN'. We identified several key papers that were strongly related to personality recognition using audio-visual features with convolutional neural network models. This paper is organized into seven sections including Introduction, Audio-Visual Features for Personality Traits Recognition (PTR), Dataset for PTR, Convolutional Neural Network Model for Audio-Visual Features, Strengths and Capabilities of CNN-based Feature Extraction, Review of Previous Related Work on CNN, and Conclusion section.

II. AUDIO-VISUAL FEATURES FOR PERSONALITY TRAITS RECOGNITION (PTR)

An image is a visual representation of something like an object or a person. In computer vision, image representation refers to the method of transforming or converting an image into a numerical form. Commonly, images are represented as a collection of pixels, with each pixel representing a different feature like colour or intensity value. This value can be easily interpreted and processed by a computer. The purpose of image representation is to extract relevant features from an image that plays the key role in image recognition. Visual features are one of the essential input data in recognition tasks including object detection and image classification. According to Zhao, Tang, and Zhang (2022), visual based input data in personality traits recognition can be divided into two categories which are static image and dynamic video sequences [20]. Static image input data enables recognition tasks to be performed on a single image, whereas dynamic video sequences employ a series of image frames. Dynamic video sequences input data not only providing series of images but additionally consists of temporal and spatial information. Both temporal and spatial information complement the series of images frames for personality traits recognition [21]. Current research in personality traits recognition used visual features from a video to automatically identify OCEAN personality traits [13][19][21]. The visual features include face images, ambient and scene backgrounds images.

In the past, detecting faces and their characteristics such as the lips, nose, eyes, gazes, and head nods was a challenging task. However, deep learning algorithms can solve this task recently. By using a simple convolutional neural network (CNN), it can easily help to detect key points on parts of faces images. Previous study has demonstrated that there is a correlation between facial key points and the Big Five personality model attributes. Cai and Liu (2022) discovered that the points from the right jawline to the chin contour showed a significant negative correlation with agreeableness in their study [23]. According to Kachur *et al*., (2020), artificial neural networks algorithm was able to reveal multidimensional personality profiles based on static morphological facial features. Morphological facial features involve the shape and structure of the front of the head from the chin to the top of the forehead where the mouth, eyes, nose and other features are located [24]. Another study found that personality traits can be reliably predicted from facial images using deep learning (CNN and Resnet) and showed an accuracy results more than 70%. Based on the study experiments, it showed that accuracy of neuroticism and extroversion was the most accurate with accuracy exceeding 90% [25].

Generally, audio is a sound wave with a frequency range of 20Hz-20kHz that can be heard by the human ear. Audio signals provide audio features such as mel-frequency cepstral coefficient (MFCC), mel-bank features (MBF), spectrogram, chromogram, spectral centroid, and spectral roll-off, which can be extracted from raw audio samples. In an accent classification task, the MFCC yielded the highest accuracy [26]. In personality trait identification, MBF features were compared to MFCC and log bank, and the results demonstrate that MBF performs better [19]. The recent trends of audio features implementation in personality traits recognition are using deep learning techniques. Zhao *et al*., (2023) used VGGish, a pretrained CNN model to capture high-level segment-level audio features to learn an audio feature

representation from video input [27]. Next, we discussed the datasets used for personality recognition, and in the following section, we reviewed the design of convolutional neural network models for audio-visual features extraction

III. DATASETS FOR PERSONALITY TRAITS RECOGNITION (PTR)

Automatic personality recognition has been studied using various datasets, each offering unique characteristics for model evaluation. These datasets differ in terms of sample size, personality trait models, labeling methods, data collection tools, and the environment used during data collection. These differences affect the performance and generalization of deep learning models in personality recognition. One of the most widely used datasets is the ChaLearn dataset [17], which consists of 10,000 YouTube videos labeled with observer rated Big Five personality traits. Due to its large sample size and uncontrolled environment, it presents challenges related to noise and variability in data including lighting conditions, face angles, background clutter, and inconsistent video quality. Similarly, the Speed Interviews Project [28] contains 8,000 videos from YouTube, also annotated with observer rated personality traits. This dataset is valuable for analyzing personality traits in real interview settings and assessing the robustness of personality recognition models in practical applications. Another large-scale dataset, the CMU-MOSEI dataset [29], expands personality recognition research with 12,300 videos labeled with six emotions. This dataset enhances personality trait analysis through multimodal emotion recognition.

In contrast to these uncontrolled datasets, several others use controlled data collection environments to reduce noise and enhance data consistency. The AVI Project database [13] includes 120 participants who completed a self-reported personality survey in a mock interview setting, allowing researchers to analyze personality traits in structured manner. Similarly, the Self-Presentation and Induced Behavior Archive for Personality Analysis (SIAP) dataset [14] consists of 60 participants and using self-reported Big Five traits for annotations in controlled interview recordings. These datasets provide more reliable personality labels, reducing the subjectivity often encountered in observer rated annotations. Additionally, other datasets focus on facial feature-based personality assessment. The IMM Face Database [30] contains 40 participants and evaluates personality traits such as extroversion, openness, agreeableness, and rigor, using facial images captured under different lighting conditions and angles. This dataset is particularly useful for studying static image-based personality prediction. The VHQ Project includes 165 video recordings of participants in controlled mock interview settings, using self-reported Big Five personality surveys for annotation [31]. Table I shows a list of existing datasets used in personality trait recognition research.

The existing studies on these datasets have leveraged CNNs to analyze facial expressions and visual cues for personality assessment by extracting meaningful facial features from images and videos. Some studies have also combined CNNs with other models, such as Long Short-Term Memory (LSTM) networks, to capture both spatial and temporal features, further enhancing recognition performance. These advancements highlight the effectiveness of CNNs in personality trait analysis and their potential for further optimization in future research. The next section will further discuss leveraging CNNs for audio-visual features.

TABLE I. EXAMPLES OF DATASETS USED IN PTR

Datasets [Ref]	***Sample Size***	***Personality Model***	***Annotation***	***Data Collection Tools / Environment***
AVI Project [13]	120	Big Five	Self-Personality Survey	Interview recording / Controlled
SIAP [14]	60	Big Five	Self-Personality Survey	Interview recording / Controlled
Chalearn [17]	10,000	Big Five	Observer Rated Personality	YouTube / Uncontrolled
Speed Interview Project [28]	8,000	Big Five	Observer Rated Personality	YouTube / Uncontrolled
CMUMOSEI [29]	12,300	6 emotions signal	Observer Rated Personality	YouTube / Uncontrolled
IMM Face [30]	40	Extroversion, Openness, Agreeable-ness, Rigor	Observer Rated Personality	Images captured / Controlled
VHQ Project [31]	165	Big Five	Self-Personality Survey	Interview recording / Controlled

IV. CONVOLUTIONAL NEURAL NETWORK MODEL FOR AUDIO-VISUAL FEATURES

Convolutional neural network (CNN) is a well-known model which is designed for image processing where it specializes for applications in image and video recognition. There are three basic types of layers in CNN includes convolutional layer, pooling layer, and fully connected layer. The convolutional layer is the core building block in CNN model which carries the responsibility for features extraction. It convolutes the input image using convolution operators and stores the convolution results to separate channels of the convolution layer. The pooling layer works to reduce the feature map's dimensionality. There will be several pooling layers scattered between convolution layers. The features become more relevant as the convolutional network moves deeper, and redundancy is also decreased. The output of the final pooling or convolutional layer is flattened and sent into the fully connected layer as the input. Fully connected layer is the last layer in CNN where image classification happens based on the extracted features. The CNN model requires an image as input, either a grayscale image (1 channel) or a color image with three channels; red, green, and blue (RGB). Thus, the approach for audio recognition using CNN is to convert the audio data into images. The common method is generating spectrograms from audio input. A spectrogram plots over time where time on the x-axis and spectrum-frequency on the y-axis. Some python libraries like librosa [32] and pyaudio [33]

are great for audio processing where the libraries can load raw audio files, extract the audio features, and visualize the audio data. Based on literature study, personality traits recognition model has been built by enhancement or modification of algorithm that based on neural network model such as convolutional neural Network (CNNs) with transformer network [27], combination of residual network (ResNet) and Long Short Term Memory (LSTM) [18-19] descriptor aggregation networks (DANs) [16], multi-task cascaded convolutional network (MTCNN) with VGG-model [22] and some more. The next paragraph will describe the details of some CNN-based models in existing study.

VGGNet is one of CNN-based models that supports up to 19 layers and is primarily concerned with the effect of convolutional neural network depth on its accuracy. VGGNet was developed with the aim of reducing the number of parameters in the convolutional layer and enhancing training time [34]. VGGNet has been used in personality traits recognition to extract facial images from the series of videos input [16]. Specifically, pre-trained deep audio CNN model called VGGish used for audio features extraction and VGG-Face used for visual features extraction [27]. The convolutional layers in VGGNet use smaller filter (3x3 or 1x1 filter) and the convolution step is fixed. VGGNet has three fully connected (FC) layers where the first two of FC layers with 4096 channels and followed by FC layers with 1000 channels to predict 1000 labels (ImageNet dataset). In addition, the pooling layer in VGGNet is not followed by each convolutional layer but the pooling layers distributed under different convolutional layers. Fig. 1 illustrates the VGGNet Model.

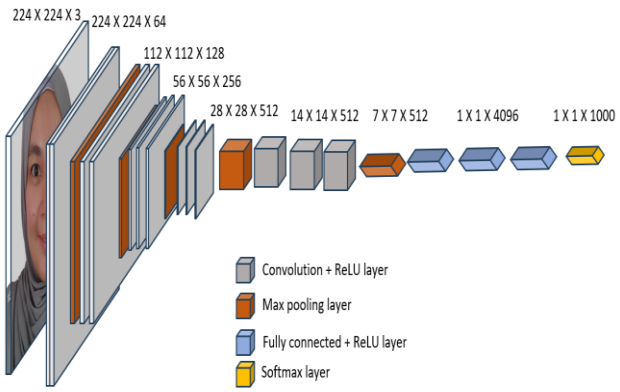

Fig. 1. Illustration of VGGNet model

Residual Network (ResNet) is another type of CNN-based model that was introduced with the basic idea of having "skip connections" built into the network [35]. The skip connections layers in ResNet aim to overcome the vanishing gradient problem by allowing gradients to flow across these levels. The ResNet model has several models where each differs in the number of layers such as ResNet18, resnet34, ResNet50 and ResNet101. For example, ResNet18 consists of 18 layers with approximately 11 million parameters while Resnet50 has 50 layers with around 23 million parameters. The more numbers of layers stacked on the network model; the more parameters produced by model. Previous studies on personality traits recognition have utilized ResNet101 model to extract face features for personality prediction and yielded prediction accuracy more than 91% [18][22]. Another study utilizes ResNet34 to extract both audio visual features including prosody features, scene background and local facial cues for personality analysis [36].

Descriptor Aggregation Networks (DANs) are modified versions of CNNs where the fully connected layers are discarded and replaced by average and max pooling layer following the last convolutional layer. Both the average and max pooling layers aggregate the deep descriptors of the last convolutionl layer before they are concatenated into the final image representation for regression [16]. By discarding the fully connected layers, several benefits offered by DANs such as reducing model size, decreasing final feature dimensionality and accelerating the model training. Another model called FaceNet was developed to perform the task of face recognition, verification and clustering using 22 deep convolutional layers. FaceNet differs from other models where it learns the mapping from the input images and builds embeddings without relying on any bottleneck layer for recognition or verification tasks [37]. Another significant point of FaceNet is its loss function where it uses triplet loss function namely anchor, positive and negative. The model should ensure that anchor image distances are closer to positive images as compared to negative images. Meanwhile, OpenFace model was inspired by FaceNet, yet OpenFace is lighter [38]. OpenFace has a weight of about 14MB, compared to FaceNet and VGG-Face which have weights of 90MB and 566MB respectively. While implementing openFace to analyze an image, the face detection library will initially create a bounding box around the face and then send each face to the neural network separately. Next, the model computes 128 dimensions face embeddings to quantify a face and then trains a Support Vector Machine (SVM) on top of the embeddings. Finally, the model successfully recognizes faces in images. In personality traits recognition OpenFace is used to handle visual modality features extraction especially facial features detection [12], [19].

Several deep learning models have been developed to address different problems and applications [39]. For example, models like autoencoders and GANs are more suited for generative tasks rather than classification-based personality inference. On the other hand, CNN-based models have demonstrated excellent performance in extracting personality-related features from images and videos due to their ability to detect spatial patterns in facial expressions and gestures. These models leverage convolutional layers to capture hierarchical feature representations, making them particularly effective for personality trait recognition. A key advantage of CNNs is their ability to utilize pre-trained models, such as VGGFace, ResNet, and EfficientNet, for transfer learning, enabling efficient feature extraction from limited datasets. CNN-based models remain the preferred approach for personality recognition due to their computational efficiency, ease of training, and adaptability to multimodal personality analysis. However, CNNs models struggle with capturing temporal dependencies in personality-related behaviors, making them less effective for long-term video-based personality analysis.

To address this, hybrid CNN-BiLSTM models have been introduced to integrate spatial and temporal feature extraction, improving the robustness of personality trait prediction. Table II summarizes the comparison of CNN-based models for personality traits recognition.

TABLE II. COMPARISON OF CNN-BASED MODELS FOR PERSONALITY TRAITS RECOGNITION

Model	*Key Characteristics*	*Advantages*	*Limitations*
VGGNet	Deep CNN model with up to 19 layers; uses small 3×3 filters	Simple architecture, effective for feature extraction	High computational cost, large model size
ResNet	Introduces skip connections to prevent vanishing gradient issues; available in ResNet18, ResNet34, ResNet50, and ResNet101 variants	Deeper networks with better gradient flow, improved accuracy	High number of parameters, increased training time
DANs	Fully connected layers replaced with global average and max pooling	Reduces model size, improves training speed	May lose fine-grained feature information
FaceNet	Uses deep CNN embeddings for face verification and clustering	High accuracy in face recognition and verification	Requires triplet loss optimization for training
Open-Face	Lightweight deep learning model inspired by FaceNet	Smaller model size, faster inference	Lower accuracy compared to larger models like VGGFace

V. STRENGTHS AND CAPABILITIES OF CNN-BASED FEATURE EXTRACTION

CNN-based algorithms are commonly adopted for training personality trait recognition models using video frames or static images due to their strong hierarchical feature learning capabilities. They can extract low-level features from image or video frames such as edges, textures, and colors, as well as high-level semantic features like facial expressions, gaze direction, and micro-expressions, which are crucial for personality analysis [19]. Other feature modalities like audios signal and transcribed texts have also been used to automatically classify a person's personality. Speech patterns, prosodic features, and textual sentiment analysis have been explored alongside CNN-based models, demonstrating the effectiveness in personality prediction. To enhance and improve classification accuracy, CNNs can be adapted to multimodal personality recognition tasks where they combine different feature modalities such as visual, audio and text. In the visual modality processing, CNNs extract spatial features from images or video frames, capturing key facial expressions, landmarks, and body posture, which are crucial for personality recognition. For audio processing, features including tone, pitch, and speech rhythm, can be extracted using spectrogram-based CNNs, where raw audio signals are transformed into MFCCs, Mel spectrograms, or Chroma features before being fed into a convolutional network [17][19]. Similarly, textual features derived from speech transcripts, captions, or social media posts can be processed using CNNs with word embeddings to capture linguistic patterns that correlate with personality traits.

Another key strength of CNN-based model in multimodal setting is their ability to learn modality-specific feature representation and enabling cross-modal fusion and interaction. This allows deeper personality traits analysis through multi-features concatenation. Advanced fusion techniques, such as attention mechanisms and hierarchical feature integration, further enhance CNNs' ability to model personality traits from multi-input sources [40]. These capabilities make CNN-based personality recognition systems valuable in applications such as automated recruitment, affective computing, and human-computer interaction, where understanding personality from multiple perspectives leads to more accurate and robust assessments. Additionally, CNNs can be integrated with pre-trained models to leverage their learned representations and improve personality trait recognition performance. Pre-trained models such as VGG16, ResNet, EfficientNet, and Vision Transformers (ViTs), which have been trained on large-scale datasets like ImageNet, provide rich feature representations that generalize well to target tasks. By employing transfer learning, these pre-trained networks can be fine-tuned on personality recognition datasets, allowing CNNs to extract meaningful spatial and semantic features from facial images or video frames without requiring extensive training from scratch. This reduced training requirement not only accelerates model adaptation but also enhances computational efficiency, as CNNs require fewer resources compared to transformer-based architectures. As a result, CNNs are well-suited for real-time personality analysis in practical applications, where speed and efficiency are crucial.

VI. REVIEW OF PREVIOUS RELATED WORK ON CNN

This paper presents a brief review on CNN models for audio-visual features extraction in personality traits recognition. Several CNN-based models have been developed in previous studies with various enhancements and modifications. Each of the models mentioned above performs well in its own way. Table III shows a summary of existing study in personality trait recognition in terms of modality, features extraction, learning model and accuracy achievement.

TABLE III. SUMMARY OF EXISTING STUDY IN PTR

Ref	*Modality*	*Features Extraction*	*Learning Model*	*Accuracy*
[13]	visual	face:OpenCV & DLib	CNN	0.9536
[15]	visual, audio	face:OpenFace	LSTM	0.9120
[16]	visual, audio	audio: pyAudio features library	DANs	0.9130
[17][a]	visual, audio	face:VGG-Face	DCNN + ELM	0.9173
[19]	visual, audio	audio: python speech features library	ResNet + LSTM	0.9207
[27]	visual, audio	face: VGG-Face	CNN + Bi-LSTM + Transformer net.	0.9167

Ref	*Modality*	*Features Extraction*	*Learning Model*	*Accuracy*
[12]	visual, audio, text	ambient: VGG-VD19	Classification-Regression Network (CR-Net)	0.6450
[18]	visual, audio, text	audio: OpenSmile	(ResNet + VGGish + ELM) + LSTM	0.9180
[22]	visual, audio, text	face, gaze signal: openFace	MTCNN + VGGish CNN + CNN	0.9143
[36]	visual, audio, text	audio: mel-bank	Classification-Regression Network (CR-Net)	0.9188

a. Baseline model for personality traits recognition

Literature also has shown that automatic personality traits recognition can be done through the ability of enhancement and modification of deep neural network models. Based on Table 1, base line model of personality traits recognition was developed using deep convolutional neural network (DCNNs) and extreme learning machine (ELM) [17]. The baseline model combines audio, scene, and facial features as data input for model development. Most of the existing study used the combination of facial features and audio signal for personality recognition. However, Sun *et al*., (2022) explored on eyes gaze features to be added on top of facial features and audio signals; and achieved accuracy performance 0.9207 outperformed other models which employed the combination of facial features and audio signal [16]. Therefore, future efforts to improve the accuracy of personality trait recognition will include considering more nonverbal visual cues such as facial landmarks, head nods, and upper body movement.

VII. CONCLUSION

Convolutional neural networks have demonstrated promising accuracy in image classification and computer vision tasks. It allows computational systems to extract meaningful information from static image or video sequences input which includes visual and audio features. Their hierarchical feature extraction capability allows them to capture both spatial and semantic features, enabling multimodal personality inference. CNNs also can be retrained for new recognition tasks using the pretrained network in which it is useful when dealing with limited input sources for training. Currently, CNN-based model increases the number depth layer to extract more relevant features and resulted the more parameters in the end layer. Whenever the network depth is deeper, the fitting ability of CNN is stronger. However, increasing the depth of network layers will not guarantee accuracy improvement. Thus, the modification on CNN-based model is not only regarding the increasing of depth layer but also on how to utilize each layer to improve the quality of extracted features [40]. Several techniques such as transfer learning, data augmentation, regularization, and hyperparameter tuning are believed to be able to improve quality of extracted features, as well as improve the accuracy performance of CNN model. Regardless of their strengths, CNNs face challenges in temporal modeling and feature fusion, limiting their ability to fully capture dynamic personality-related behaviors. In future, it is interesting to explore more on modification and enhancement of CNN-based model for audio-visual features extraction to improve personality recognition performance. Recent advancements, such as hybrid CNN-BiLSTM architectures, transformer-based models, and dataset augmentation, offer new opportunities to enhance personality trait prediction accuracy.

ACKNOWLEDGMENT

This research was not funded by a grant.

CONFLICTS OF INTEREST

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] Khan, Alam Sher, Hussain Ahmad, Muhammad Zubair Asghar, Furqan Khan Saddozai, Areeba Arif, and Hassan Ali Khalid. (2020). Personality Classification from Online Text using Machine Learning Approach. *International Journal of Advanced Computer Science and Applications*, 11(3), 460–476. Doi: https://doi.org/10.14569/IJACSA.2020.0110358.

[2] Mehta, Yash, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. (2020). Recent Trends in Deep Learning based Personality Detection. *Artificial Intelligence Review*, *53*(4), 2313–2339. Doi: https://doi.org/10.1007/s10462-019-09770-z.

[3] Quenk, Naomi L. (2009). *Essentials of Myers-Briggs Type Indicator Assessment.* John Wiley & Sons. https://books.google.com.my/books?id=th_gTxfPdlgC.

[4] Costa, P., & McCrae, R. (2008). The Revised Neo Personality Inventory (neo-pi-r). In G. J. Boyle, G. Matthews, D. H. Saklofske (Eds.). *The Revised NEO Personality inventory* (NEO-PI-R) (pp. 179-198). SAGE Publications Doi: https://doi.org/10.4135/9781849200479.n9.

[5] Eysenck, H. J., Eysenck, S. B. G. (1985). *Manual of the Revised Eysenck Personality Questionnaire.* Hodder and Stoughton, London. Doi: 10.1037/t12641-000.

[6] Eysenck, Hans Jurgen. (2012). *A Model for Personality.* Springer Science & Business Media. Doi: https://doi.org/10.1007/978-3-642-67783-0.

[7] Digman, John M. (1990). Personality Structure: Emergence of the Five-Factor Model. *Annual Review of Psychology*. 41(1), 417–440. Doi: https://doi.org/10.1146/annurev.ps.41.020190.002221.

[8] Digman, John M. (1997). Higher-order Factors of the Big Five. *Journal of Personality and Social Psychology*. *73*(6), 1246. Doi: https://doi.org/10.1037/0022-3514.73.6.1246.

[9] Monaro, Merylin, Cristina Mazza, Marco Colasanti, Stefano Ferracuti, Graziella Orrù, Alberto di Domenico, Giuseppe Sartori, and Paolo Roma. (2021). Detecting Faking-good Response Style in Personality Questionnaires with Four Choice Alternatives. *Psychological Research.* 1–14. Doi: https://doi.org/10.1007/s00426-020-01473-3.

[10] Tett, Robert P., and Daniel V. Simonet. (2021). Applicant Faking on Personality Tests: Good or Bad and Why Should We Care? *Personnel Assessment and Decisions,* 7(12). Doi: https://doi.org/10.25035/pad.2021.01.002.

[11] Gürpinar, Furkan, Heysem Kaya, and Albert Ali Salah. (2016). "Multimodal Fusion of Audio, Scene, and Face Features for First Impression Estimation. *2016 23rd International conference on pattern recognition (ICPR),* IEEE. 43–48.
Doi: https://doi.org/10.1109/ICPR.2016.7899605.

[12] Hemamou, Léo, Ghazi Felhi, Vincent Vandenbussche, Jean-Claude Martin, and Chloé Clavel. (2019). Hirenet: A Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews. *Proceedings of the AAAI conference on artificial intelligence*, *33*(01), 573–581.
DOI: https://doi.org/10.1609/aaai.v33i01.3301573.

[13] Suen, Hung-Yue, Kuo-En Hung, and Chien-Liang Lin. (2019). Tensor Flow-based Automatic Personality Recognition used in Asynchronous Video Interviews. *IEEE Access, 7*, 61018–61023.
Doi: https://doi.org/10.1109/ACCESS.2019.2902863

[14] Giritlioğlu, D., Mandira, B., Yilmaz, S. F., Ertenli, C. U., Akgür, B. F., Kınıklıoğlu, M., Kurt, A. G., Mutlu, E., Gürel, Ş. C., & Dibeklioğlu, H. (2021). Multimodal Analysis of Personality Traits on Videos of Self-presentation and Induced Behavior. *Journal on Multimodal User Interfaces*, *15*(4).
https://doi.org/10.1007/s12193-020-00347-7.

[15] Subramaniam, Arulkumar, Vismay Patel, Ashish Mishra, Prashanth Balasubramanian, and Anurag Mittal. (2016). Bi-modal First Impressions Recognition using Temporally Ordered Deep Audio and Stochastic Visual Features. *Computer Vision–ECCV 2016 Workshops*: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14, 337-348. Springer International Publishing.
Doi: https://doi.org/10.1007/978-3-319-49409-8_27.

[16] Zhang, Chen-Lin, Hao Zhang, Xiu-Shen Wei, and Jianxin Wu. (2016). Deep Bimodal Regression for Apparent Personality Analysis*. European Conference on Computer Vision*, 311–324.
Doi: https://doi.org/10.1007/978-3-319-49409-8_25.

[17] Kaya, Heysem, Furkan Gurpinar, and Albert Ali Salah. (2017), Multi-modal Score Fusion and Decision Trees for Explainable Automatic Job Candidate Screening from Video cvs. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 1–9.
DOI: https://doi.org/10.1109/CVPRW.2017.210

[18] Aslan, Süleyman, Uğur Güdükbay, and Hamdi Dibeklioğlu. (2021). Multimodal Assessment of Apparent Personality using Feature Attention and Error Consistency Constraint. *Image and Vision Computing, 110*, 104163.
Doi: https://doi.org/10.1016/j.imavis.2021.104163.

[19] Sun, Xiao, Jie Huang, Shixin Zheng, Xuanheng Rao, and Meng Wang. (2022). Personality Assessment based on Multimodal Attention Network Learning with Category-based Mean Square Error. *IEEE Transactions on Image Processing, 31*, 2162-2174.
Doi: https://doi.org/10.1109/TIP.2022.3152049.

[20] Zhao, Xiaoming, Zhiwei Tang, and Shiqing Zhang. (2022), Deep Personality Trait Recognition: A Survey. *Frontiers in Psychology, 13*, 839619.
Doi: https://doi.org/10.3389/fpsyg.2022.839619.

[21] Junior, Julio CS Jacques, Yağmur Güçlütürk, Marc Pérez, Umut Güçlü, Carlos Andujar, Xavier Baró, Hugo Jair Escalante *et al*. (2019). First Impressions: A Survey on Vision-based Apparent Personality Trait Analysis. *IEEE Transactions on Affective Computing, 13*(1), 75–95.
Doi: https://doi.org/10.1109/TAFFC.2019.2930058.

[22] Suman, Chanchal, Sriparna Saha, Aditya Gupta, Saurabh Kumar Pandey, and Pushpak Bhattacharyya. (2022). A Multi-modal Personality Prediction System. *Knowledge-based Systems, 236*, 107715.
Doi: https://doi.org/10.1016/j.knosys.2021.107715.

[23] Cai, Lei, and Xiaoqian Liu. (2022). Identifying Big Five Personality Traits based on Facial Behavior Analysis. *Frontiers in Public Health, 10*, 1001828.
Doi: https://doi.org/10.3389/fpubh.2022.1001828.

[24] Kachur, Alexander, Evgeny Osin, Denis Davydov, Konstantin Shutilov, and Alexey Novokshonov. (2020). Assessing the Big Five Personality Traits using Real-life Static Facial Images. *Scientific Reports, 10*(1), 8487.
Doi: https://doi.org/10.1038/s41598-020-65358-6.

[25] Li, Jialin, Alia Waleed, and Hanan Salam. (2023). A Survey on Personalized Affective Computing in Human-machine Interaction. arXiv preprint arXiv:2304.00377.
Doi: https://doi.org/10.48550/arXiv.2304.00377.

[26] Singh, Yuvika, Anban Pillay, and Edgar Jembere. (2020). Features of Speech Audio for Accent Recognition. 2020 *International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems* (icABCD), IEEE. 1–6.
Doi: https://doi.org/10.1109/icABCD49160.2020.9183893.

[27] Zhao, Xiaoming, Yuehui Liao, Zhiwei Tang, Yicheng Xu, Xin Tao, Dandan Wang, Guoyu Wang, and Hongsheng Lu. (2023). Integrating Audio and Visual Modalities for Multimodal Personality Trait Recognition via Hybrid Deep Learning. *Frontiers in Neuroscience*, *16*, 1107284.
Doi: https://doi.org/10.3389/fnins.2022.1107284.

[28] Gorbova, J., Avots, E., Lusi, I., Fishel, M., Escalera, S., & Anbarjafari, G. (2018). Integrating Vision and Language for First-Impression Personality Analysis. *IEEE Multimedia*, *25*(2).
https://doi.org/10.1109/MMUL.2018.023121162.

[29] Williams, J., Kleinegesse, S., Comanescu, R., & Radu, O. (2018). Recognizing Emotions in Video Using Multimodal DNN Feature Fusion. 11–19.
https://doi.org/10.18653/v1/w18-3302.

[30] Fu, J., & Zhang, H. (2021). Personality Trait Detection based on ASM Localization and Deep Learning. *Scientific Programming,* 2021.
https://doi.org/10.1155/2021/5675917.

[31] Song, S., Jaiswal, S., Sanchez, E., Tzimiropoulos, G., Shen, L., & Valstar, M. (2021). Self-supervised Learning of Person-specific Facial Dynamics for Automatic Personality Recognition. *IEEE Transactions on Affective Computing.*
https://doi.org/10.1109/TAFFC.2021.3064601

[32] McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. (2015). Librosa: Audio and Music Signal Analysis in Python. *SciPy*, 18–24.
Doi: https://doi.org/10.25080/Majora-7b98e3ed-003.

[33] Pham, Hubert. (2006). Pyaudio: Portaudio v19 Python Bindings. https://people. csail. mit. edu/hubert/pyaudio.

[34] Simonyan, Karen, and Andrew Zisserman. (2014). Very Deep Convolutional Networks for Large-scale Image Recognition. arXiv preprint arXiv:1409.1556.
Doi: https://doi.org/10.48550/arXiv.1409.1556.

[35] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
Doi: https://doi.org/10.1109/CVPR.2016.90.

[36] Li, Yunan, Jun Wan, Qiguang Miao, Sergio Escalera, Huijuan Fang, Huizhou Chen, Xiangda Qi, and Guodong Guo. (2020). Cr-net: A Deep Classification-regression

Network for Multimodal Apparent Personality Analysis. *International Journal of Computer Vision, 128,* 2763–2780. Doi: https://doi.org/10.1007/s11263-020-01309-y.

[37] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. (2015). Facenet: A Unified Embedding for Face Recognition and Clustering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 815–823. Doi: https://doi.org/10.1109/CVPR.2015.7298682.

[38] Amos, Brandon, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. (2016). Openface: A General-purpose Face Recognition Library with Mobile Applications. *CMU School of Computer Science, 6*(2), 20.

[39] Bengesi, S., El-Sayed, H., Sarker, M. K., Houkpati, Y., Irungu, J., & Oladunni, T. (2024). Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers. *IEEE Access*, 12. https://doi.org/10.1109/ACCESS.2024.3397775.

[40] Luo, Juanjuan, and Defa Hu. (2023). An Image Classification Method Based on Adaptive Attention Mechanism and Feature Extraction Network. Computational Intelligence and Neuroscience. Doi: https://doi.org/10.1155/2023/4305594.