KDU Journal of Multidisciplinary Studies (KJMS)
Volume 6. Issue 2 Nov 2024
DOI: https://doi.org/10.4038/kjms.v6i2.141

KJMS 2024 VOL.6 (2): 204-218

EXPLORING THE IMPACT OF FACIAL FEATURES ON APPARENT PERSONALITY TRAITS DETECTION USING DEEP LEARNING TECHNIQUES

WMKS Ilmini[1,2] and TGI Fernando[3]
Department of Computer Science, Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka[1],

Faculty of Graduate Studies, University of Sri Jayewardenepura, Nugegoda, Sri Lanka[2],

Department of Computer Science, Faculty of Applied Sciences, University of Sri Jayewardenepura, Nugegoda, Sri Lanka[3]

ABSTRACT

Apparent personality detection has emerged as a prominent research area within deep learning. While numerous deep learning solutions have been developed to predict personality accurately, the lack of transparency in how these models derive predictions based on facial features undermines trust in their results. This study focuses on identifying and differentiating facial features that contribute to the Big-Five personality traits, addressing transparency in model predictions. To conduct our experiments, we utilised the ChaLearn First Impressions V2 dataset, with background removed frames ensuring models focused more on human features than background in the learning process. We began by developing Convolutional Neural Networks architectures using pre-trained VGGFace and VGG19 models. Subsequently, we employed the Grad-CAM and Guided Grad-CAM model explainable AI techniques on the test and validation datasets, utilising the trained models. Furthermore, we employed the "SelectKBest" feature selection method to analyse the outcomes of the interpretability techniques. VGG19 achieved higher accuracy (90%) compared to VGGFace (89%). Our investigation reveals that personality prediction extends beyond facial features, with XAI techniques emphasizing non-facial aspects such as background information. Statistical analysis across deep learning architectures shows no significant correlation between features identified by XAI techniques by giving different F1-scores. Despite VGG19's superior accuracy, it exhibits a stronger inclination towards non-facial data, while VGGFace prioritizes facial features, highlighting the nuanced nature of personality prediction and suggesting avenues for further research.

KEYWORDS: *Apparent Personality Detection (APD), Convolutional Neural Networks (CNN), Explainable AI (XAI), Facial Features, Select best Feature Selection*

Corresponding Author: WMKS Ilmini, Email: kalaniilmini@kdu.ac.lk; wmksilmini@gmail.com

https://orcid.org/0000-0001-6942-6390

1. INTRODUCTION

Apparent Personality Detection (APD) has garnered significant attention in the field of deep learning research. Researchers have developed numerous deep learning solutions to predict personality traits based on facial features accurately. However, the lack of transparency regarding how these models derive their output from facial cues of-ten leads to a loss of trust in the predictions. As a result, there is a pressing need to delve deeper into identifying and differentiating facial features that influence Big-Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) (Wiggins, 1996).

APD has emerged as a prominent computer vision application with wide-ranging implications in various fields. It finds utility in diverse domains, including human resource management (Lounsbury et al., 2008; Penney, David and Witt, 2011; Alhendi, 2019), social robotics (Lee et al., 2006; Kirby, Forlizzi and Simmons, 2010; Mileounis, Cuijpers and Barakova, 2015), criminology (Reid, 2011), game development (Zammitto, DiPaola and Arya, 2008), and the animation movie industry (Juhan and Ismail, 2016). The measurement of personality encompasses various criteria, with the Big-Five personality model (Wiggins, 1996) being widely adopted and endorsed by psychologists. This model evaluates personality based on five fundamental factors: Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness. These factors are often abbreviated as "OCEAN" or "CANOE." To capture the nuances of each aspect, they are further delineated into sub-traits (John and Srivastava, 1999), collectively providing a comprehensive understanding of an individual's personality.

According to the literature, to measure the apparent personality, researchers developed various deep learning solutions such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Three-Dimensional Convolutional Neural Networks (3D-CNN). The primary objective of these studies is to design and implement a model capable of measuring apparent personality. One of the main milestones of apparent personality detection is the “Looking at People ECCV Challenge First Impression” held in 2016 (ChaLearn Looking at People - 2016 Looking at People ECCV Challenge, 2016). Researchers in this area developed different solutions to measure the personality precisely as much as possible with the ChaLearn Apparent Personality First Impressions V1 (ECCV’16) and V2 datasets (CVPR'17) (Ponce-López et al., 2016). The winners of this competition were Zhang et al. (2016), Subramaniam et al. (2016), and Güçlütürk et al. (2016). All the winners used audio and visual data to predict apparent personality. Zhang and others (2016) designed a bi-model deep regression model to predict personality from visual and audio modalities. They used CNN architectures, consisting of VGGFace (Parkhi, Vedaldi and Zisserman, 2015) pre-trained model for visual modality. These architectures are named DAN (Descriptor Aggregation Network) and DAN+. Subramaniam et al. (2016) presented a bi-model architecture with visual (background-removed) and audio features for personality judgment. They designed a 3D CNN and an LSTM to predict personality. Gucluturk et al. (2016) created bi-model deep neural network architecture with residual blocks. The maximum mean accuracies obtained by each competitor are Zhang et al. (2016) 0.9111, Subramaniam et al. (2016) 0.913355, and Güçlütürk et al. (2016) 0.912132.

After the competition, researchers developed various deep learning architectures with different modalities to predict the apparent personality with CVPR'17. Gürpınar, Kaya and Salah (2016) used visual data extracted from videos and achieved 0.9094 accuracy. Yang and Glaser (2017) proposed a Bi-model LSTM, scoring 0.9083 with L2 loss. Barezi et al. (2018) used visual, audio, and text (audio transcription) data to predict the apparent personality with tri-model deep convolutional architecture. The results concluded that the audio and text modalities are least relevant for personality detection, while visual features are more relevant, with tri-model accuracy of 0.9062. Li and others (2020) also used a multi-feature model with visual, facial, audio, and transcription data to predict the apparent personality. They proposed a deep CR-Net (classification-regression network) composed of three branches. The proposed architecture achieved 0.9188 accuracy. Mujtaba and Mahapatra (2021)

presented the multi-task deep learning approach to measure personality with visual, facial, audio, and transcription data. The proposed architecture achieved 0.9134 accuracy.

So, a significant amount of research has been done on predicting apparent personality traits, resulting in high accuracies. However, there is a noticeable dearth of studies that focus on explaining the outputs generated by these prediction models. Furthermore, the existing research in the field of explaining model outputs demonstrates varying outcomes and conclusions.

Zhang et al. (2016) used the heatmap feature visualisation XAI technique to visualise the features with the architectures ResNet, DAN, and DAN+ used in APD. According to their results, different architectures tend to focus on different image features. ResNet focused on the human body, including facial and non-facial features, while DAN/DAN+ architectures focused on human and background data. They used five random sample videos to extract this information.

Ventura et al. (2017) conducted a quantitative study with the Class Activation Map (CAM) (Zhou et al., 2015) technique and Action Unit (AU) (Ekman, Friesen and Ancoli, 1980) to gain insight into CNN-based apparent personality trait recognition. CAM was applied to find the discriminative regions in the scene data, which support personality predictions. They identified vital face regions as eyes, nose, and mouth. Then, they used the "OpenFace" library (Mahdy, Hereñú and Sumsuddin, 2019) to detect the face region and the Facial Action Coding System (Ekman, Friesen and Ancoli, 1980) to find AUs. Only 17 AUs were used in this study out of 45 AUs. These AUs were used to find the influence of emotions on personality detection. The quantitative research shows that some AUs affected personality detection. For the experiment, they selected 50 images with the highest score for personality traits.

Wei et al. (2018) used the DAN architecture initially proposed by Zhang et al. (2016) to predict the apparent personality. The results of the model interpretability techniques concluded that ResNet could identify facial regions as the primary contributors to the output. Simultaneously, DAN and DAN+ architectures were more prone to background data than ResNet. Nevertheless, with plain background data, ResNet failed, but DAN and DAN+ could identify facial features as primary contributors. They used 12 randomly selected images to interpret the output of the model.

Yang and Glaser (2017) used the saliency map visualisation technique to find the image regions contributing to the apparent personality judgment. Furthermore, they concluded that ResNet could identify facial features as the most contributing factors to the network's output.

Li et al. (2020) used the Seaborn Python library (Waskom, 2021) to calculate heatmap on scene data to find which features mostly contribute to the apparent personality judgments. The heatmap calculation on scene data found that facial features such as the eye, nose and mouth are primarily contributing elements. Also, they concluded that non-facial data such as clothing and furnishing contribute to personality judgments. They conducted a quantitative study on the relationship between heatmap features and face key points. They used face key points as two eyes, nose, corners of the mouth, and mid-distance of the two mouth corners. According to the findings, 73.96% of the highlighted points are vital. For the experiment, they used 32 frames from each video from the test dataset of CVPR'17.

Ilmini and Fernando (2022) used Grad-CAM, Guided Backpropagation, and Guided Grad-CAM XAI techniques XAI explain the outputs of apparent personality CNN models. They also concluded that facial features such as eyes, forehead, eyebrows, nose, and mouth are mainly involved in personality detection. Further, they mentioned that the background affects personality prediction.

As found in the literature, different research works tend to interpret the output of the CNN-based APD. They have used heatmap visualisation techniques such as Class Activation Map (CAM) and saliency map visualisation to find the most contributing features. They concluded that the facial region contributes more to the APD, while background data also affects the APD. Some researchers compare the XAI technique outputs with different network architectures and

conclude that they behave differently. Also, according to the literature, interpretability techniques do not convey the difference in personality traits. The summary of the literature is given in Table 1.

Table 1 :Summary of the Literature Review

Study	Methodology	Features used	Accuracy	Key Contributions	XAI Techniques Used	XAI Outputs
Zhang et al. (2016)	Bi-modal deep regression model using CNNs with VGGFace	Combined visual and audio features	0.9111	Introduced DAN/DAN+ architectures for personality prediction using visual and audio modalities	Heatmap visualization	ResNet focused on both facial and non-facial features; DAN/DAN+ emphasized background features
Subramaniam et al. (2016)	3D CNN and LSTM with bi-modal architecture	Combined visual and audio features, background removed	0.913355	Utilized background-removed frames to enhance personality prediction	N/A	N/A
Güçlütürk et al. (2016)	Bi-modal deep neural network with residual blocks	Combined audio and visual data for better accuracy	0.9109	Developed a multi-modal architecture improving personality detection with residual networks	N/A	N/A
Gürpınar et al. (2016)	Visual data extraction using pre-trained deep learning CNN model	Focused solely on visual data for personality prediction	0.9094	Showcased the effectiveness of visual-only data for personality prediction	N/A	N/A
Yang and Glaser (2017)	Bi-modal LSTM with L1 and L2 loss	Used visual and audio data to predict personality traits	0.9083	Proposed LSTM-based personality prediction model combining visual and audio data with pre-trained deep learning models	Saliency map visualization	Identified facial features as key contributors to personality judgments
Ventura et al. (2017)	CNN with DAN, DAN+ architectures	Used visual data	0.912	Used CAM and AUs to link facial emotions to personality trait detection	CAM, Action Unit (AU) analysis	Found vital face regions (eyes, nose, mouth) as key contributors; emotions affected personality detection
Barezi et al. (2018)	Tri-modal deep convolutional architecture (visual, audio, text) with ensemble techniques	Concluded that visual features are more relevant than audio and text for personality prediction	0.9062	Highlighted the limited impact of audio and text modalities compared to visual features	N/A	N/A
Wei et al. (2018)	Deep Bi-model Regressor network	Audio and Visual Data	0.9212 with epoch fusion	Demonstrated the effect of background removal on deep learning interpretability	Feature Map	ResNet highlighted facial regions, DAN/DAN+ focused more on background
Li et al. (2020)	Deep CR-Net (Classification-Regression Network) with introduced bell-loss	Combined visual, facial, audio, and transcription data for personality analysis	0.9188	Introduced a multi-feature architecture with high accuracy for personality trait prediction	Heatmap (Seaborn library)	Identified facial features like eyes, nose, and mouth as primary contributors; non-facial data like clothing also relevant
Mujtaba and Mahapatra (2021)	Multi-task deep learning with visual, facial, audio, transcription with k-fold cross validation	Used a multi-modal approach to predict apparent personality	0.9134	Applied multi-task learning for comprehensive personality detection using multiple modalities	N/A	N/A
Ilmini & Fernando (2022)	CNN with pre-trained deep learning models	Used visual features	0.9061	Emphasized the role of background information even in background-removed datasets	Grad-CAM, Guided Backpropagation, Guided Grad-CAM	Facial features (eyes, forehead, mouth) identified, background still affects personality prediction

This study addresses this critical gap by exploring the relationship between facial features and personality traits using deep learning techniques. Our primary objective is to unravel the impact of facial cues on personality prediction and shed light on the interpretability of these deep learning models. Understanding how these models make projections can enhance their transparency and build trust in their results. We leverage the CVPR'17 dataset to conduct our experiments, curated explicitly for analysing apparent personality. To minimise the potential bias arising from contextual factors, such as background information, we remove the background frames from the dataset. This approach ensures that our focus remains solely on facial features and their influence on personality prediction.

The rest of the paper is organised as follows. Section Two discusses materials and methods, Section Three includes the results, and Section Four contains the discussion and conclusion.

2. METHODOLOGY

In this section, we outline the methodology employed in this study to identify the significant facial features that influence the assessment of the Big-Five personality traits.

Figure 1 depicts the comprehensive research methodology utilised to identify the prominent facial features that impact the Big-Five personality traits.

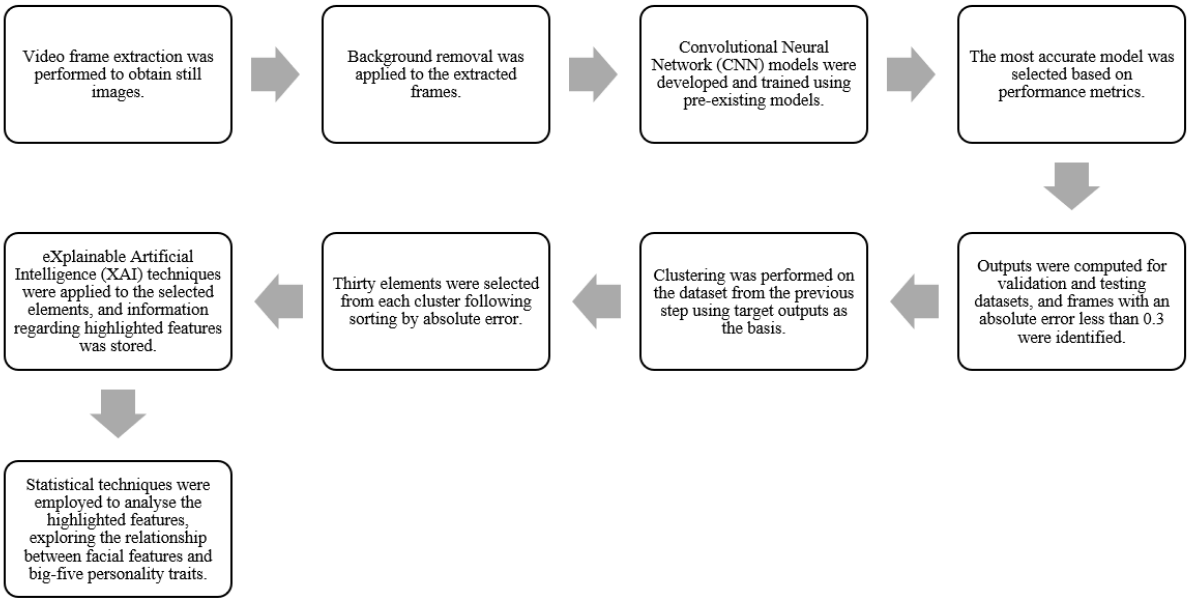

Figure 1: Methodology

Dataset and Pre-processing

The CVPR'17 dataset was employed for this study, comprising videos featuring individuals from diverse nationalities, age groups, and ethnic backgrounds. The dataset is composed of 10,000 video clips extracted from 3,000 unique videos. Among these, 6,000 video clips were assigned to the training dataset, while the validation and test datasets consisted of 2,000 video clips each. Each video clip in the dataset is associated with ground truth values representing the Big-Five personality traits, ranging from 0 to 1.

After extracting 20 frames from each video, the image dataset sizes are as follows:

- Training dataset: 6,000 X 20 = 120,000 images
- Validation dataset: 2,000 X 20 = 40,000 images

- Test dataset: 2,000 X 20 = 40,000 images

The background of each image was removed and replaced with black colour during the preprocessing stage to mitigate potential bias introduced by background details. This step was undertaken to ensure a focus on identifying significant facial features that influence personality traits in current research. The Rembg Python library (Gatis, 2022) was utilised to remove the background.

Network Architecture

This study developed the CNN models using the VGGFace (Parkhi, Vedaldi and Zisserman, 2015) and VGG19 (Simonyan and Zisserman, 2014) pre-trained deep learning models. Specifically, for the VGGFace-based CNN architecture, the DAN (Deep Aggregation Network) architecture, initially introduced by Zhang et al. (Zhang et al., 2016), was utilised. As for the VGG19-based CNN architecture, the classifier layer was removed, and two fully connected layers with output sizes of 512 and 5 were added, followed by a final sigmoid layer. These modifications were made to adapt the VGG19 model for the task of personality trait prediction.

Hyper-parameters of the network are as follows:

- Batch-size:16
- Learning rate: $1 \times e^{-5}$
- Maximum number of epochs: 200
- Early stop counter: 20
- Optimiser: RAdam=$1 \times e^{-6}$

Several test runs were conducted to determine the optimal hyperparameters mentioned above. Then, each network was trained ten times using the finalised parameters and evaluated using the test dataset. The model that achieved the highest accuracy on the test dataset was selected and saved for further analysis, specifically for feature visualisation purposes.

Prepare the Dataset for the Visualisation

The highest accuracy model was employed to evaluate the validation and test datasets. Subsequently, the images that obtained an absolute error (as defined by Equation 1) less than 0.3 were selected for further analysis. Approximately 90% of the videos from both the test and validation datasets exhibited an absolute error below this threshold.

Next, the prepared dataset was partitioned into groups based on the ground truth value (target) as outlined below:

Cluster 1: Target ≥ 0.8

Cluster 2: $0.8 >$Target ≥ 0.6

Cluster 3: $0.6 >$ Target ≥ 0.4

Cluster 4: $0.4 >$ Target ≥ 0.2

Cluster 5: Target ≤ 0.2

Absolute Error =|target - output| (1)

The target is the ground truth value, and the output is the network output.

In general, Cluster 5 tends to contain a smaller amount of data for all traits. Consequently, when selecting data, Cluster 5 is often disregarded. Each cluster is sorted in ascending order based on the absolute error (Equation 1), which is the difference between the target and the error. From each group, 30 subjects were chosen for the feature visualisation stage. As a result, 120 images with considerably lower absolute errors were utilised to interpret the network's output for each trait.

Feature Visualisation Techniques Used

Deep learning interpretability techniques have witnessed significant advancements, surpassing the limitations of earlier methods. To explain the inner process of deep learning models, a range of approaches have been developed. These techniques, commonly known as post hoc interpretability techniques, aim to interpret already trained deep learning models. While these techniques offer valuable insights, it is essential to acknowledge that they possess their own set of advantages, disadvantages, and limitations. Furthermore, interpretability techniques can be classified into model-specific and model-agnostic methods. Model-specific methods are tailored for specific models, providing explanations for their particular internal structure and operations. On the other hand, model-agnostic techniques are more generalised, applicable

across different models, and offer broader insights into model behaviour. These model-agnostic techniques enable interpretability regardless of the specific deep learning architecture employed.

Class Activation Map (CAM), GRADient-weighted CAM (Grad-CAM), and Guided Grad-CAM are widely recognised explainable artificial intelligence (XAI) techniques designed explicitly for interpreting CNN models. CAM and Grad-CAM are model-specific techniques employed to analyse CNN models concerning a specific target class. CAM requires a particular network structure in the final layer and is utilised for interpreting the last convolutional layer. On the other hand, Grad-CAM applies to any network structure, offering a generalisation of CAM. In contrast, Guided Grad-CAM stands out by its ability to generate high-resolution class discriminative visualisations, providing detailed insights into the model's decision-making process. The review study conducted by Linardatos et al. (2020) concluded that regarding CNN XAI techniques, Grad-CAM is the most influential technique according to citations per year. Hence, we used Grad-CAM and Guided Grad-CAM XAI techniques in this study.

Grad-CAM generalises the CAM for various convolutional neural network architectures (Selvaraju et al., 2020). Grad-CAM calculates a class discriminative map by taking the gradient of the score for category c y^c concerning the convolutional layer's feature map activation A^k, $\frac{\partial y^c}{\partial A^k}$. These gradients are globally averaged and pooled to obtain neuron importance (α_k^c) over width and height dimensions. Finally, to obtain the Grad-CAM, they performed a weighted combination of forward activation maps followed by the ReLU activation.

Guided Grad-CAM

Grad-CAM cannot highlight the fine-grained regions as Guided Backpropagation and Deconvolutional techniques. Because of that, the authors have obtained element-wise multiplication of Grad-CAM and Guided Backpropagation outputs. Grad-CAM's coarse map highlights the image regions, while Guided Grad-CAM identifies the object's edges (Selvaraju et al., 2020). Grad-CAM and Guided Grad-CAM techniques were employed to recognise the features that contribute positively to the output. Furthermore, additional processing was conducted on the original image and the heatmap to enhance the clarity of the visualisation output. Initially, the heatmap was utilised to identify the contours associated with it. Subsequently, these contours were used to draw polygons on the original image, allowing for the identification of the specific areas covered by the heatmap. The following pseudocode (Algorithm 1) illustrates the processing of the original image and heatmap to visualise the feature map.

```
grey_img = cv2.cvtColor(heatmap, cv2.COLOR_BGR2GRAY)
ret, thresh = cv2.threshold(grey_img, 50, 255,
cv2.THRESH_BINARY)
contours, hierarchy = cv2.findContours(thresh, 1, 2)
for item in range(len(contours)):
      cnt = contours[item]
            if len(cnt)>20:
            poly_coords = [cnt]
            cv2.polylines(img, poly_coords, True, (0,0,1), 2)
```

Algorithm 1: Pseudocode for processing the original image and heatmap to obtain a clear visualisation of the feature map.

As described earlier, the dataset consisting of 120 images was subjected to the model as mentioned earlier interpretability techniques. Building upon the methodology outlined in the original paper (Selvaraju et al., 2020), we visualised the final convolutional layer of both the VGGFace and VGG19 models.

Feature Identification

In the process of feature visualisation, a total of 11 facial features and six non-facial features were considered. The facial features encompassed the forehead, brow, eyebrow, bridge of the nose, eye, cheek, nose, nasolabial angle, mouth, mental fold, and chin. On the other hand, the non-facial features comprised the hair, neck, ear, dress/jewellery/hand gestures, and the presence of spectacles or a beard (specifically for males), as indicated by the visualisation techniques employed.

Subsequently, the dataset generated in the previous step was analysed to determine the presence of the 17 features mentioned above within each feature map. A binary approach was adopted, where a value of 1 was assigned if the visualisation techniques used highlighted a specific area or part of an area corresponding to a feature, and a value of 0 was set otherwise.

Consequently, a new dataset encompassing the target values, the network's output, and the highlighted features was curated. This process was iterated for all five traits using the VGGFace and VGG19-based models, resulting in 10 distinct datasets. Subsequently, the Python sci-kit-learn library's 'SelectKBest' module (‘scikit-learn/scikit-learn’, 2022) was applied to these datasets to determine the statistical relationship between the highlighted features and the network's output. The 'SelectKBest' algorithm, coupled with the f_regression score function (‘sklearn. feature_selection.f_regression’, 2022), was employed, which yields both an F-statistic and a P-value. A higher F-statistic signifies a significant contribution to the model's predictive ability.

3. RESULTS

This section provides a comprehensive analysis and presentation of the findings derived from the exploration of facial features and their influence on Big-Five personality traits using deep learning techniques. This section presents the outcomes of the conducted study, highlighting the key insights and observations obtained through the analysis of the collected data.

Model Accuracy

Table 2 summarises the mean accuracies achieved by each model. According to the results, the VGG19-based CNN model outperforms the VGGFace-based CNN model. Both models scored the lowest accuracy for the Neuroticism trait.

Table 2: Mean Accuracy values scored on the test dataset.

Trait	VGGFace	VGG19
Extraversion	0.9001	0.9079
Neuroticism	0.8946	0.9027
Agreeableness	0.9013	0.9087
Conscientiousness	0.9013	0.9121
Openness	0.9006	0.9074

1 Cell values represent the accuracy of each model for the personality traits on the test dataset. The Overall accuracy was calculated using Equation 2

$$Mean\ Accuracy = 1 - \frac{1}{N * M}\sum_{j=1}^{M}\sum_{i=1}^{N}|target_{ij} - output_{ij}| \quad (2)$$

N=number of videos, M = 5 (Big Five Personality Traits), target = ground truth value and output = network output.

Visualisation Techniques

VGGFace

Figure 2 displays sample outputs for various traits representing a selected instance. The analysis revealed that both facial and non-facial features, along with the background, were identified as the most influential factors. The heatmap predominantly emphasises the human body, encompassing facial and non-facial attributes. Notably, in the case of Extraversion, Neuroticism, and Agreeableness, non-facial characteristics such as dress were found to carry more

significance than facial data for the given instance. Ground truth values of the selected sample are as follows: E: 0.89719623, N: 0.8854167, A: 0.7692308, C:0.7669903, and O: 0.8333333.

VGG19

Figure 2 illustrates sample outputs obtained for the Big-Five personality traits using the VGG19-based CNN model. In these samples, both facial and non-facial features were identified as significant contributors to the predictions. However, in certain instances, the model exhibited a stronger emphasis on the background rather than the facial and non-facial features (Figure 2).

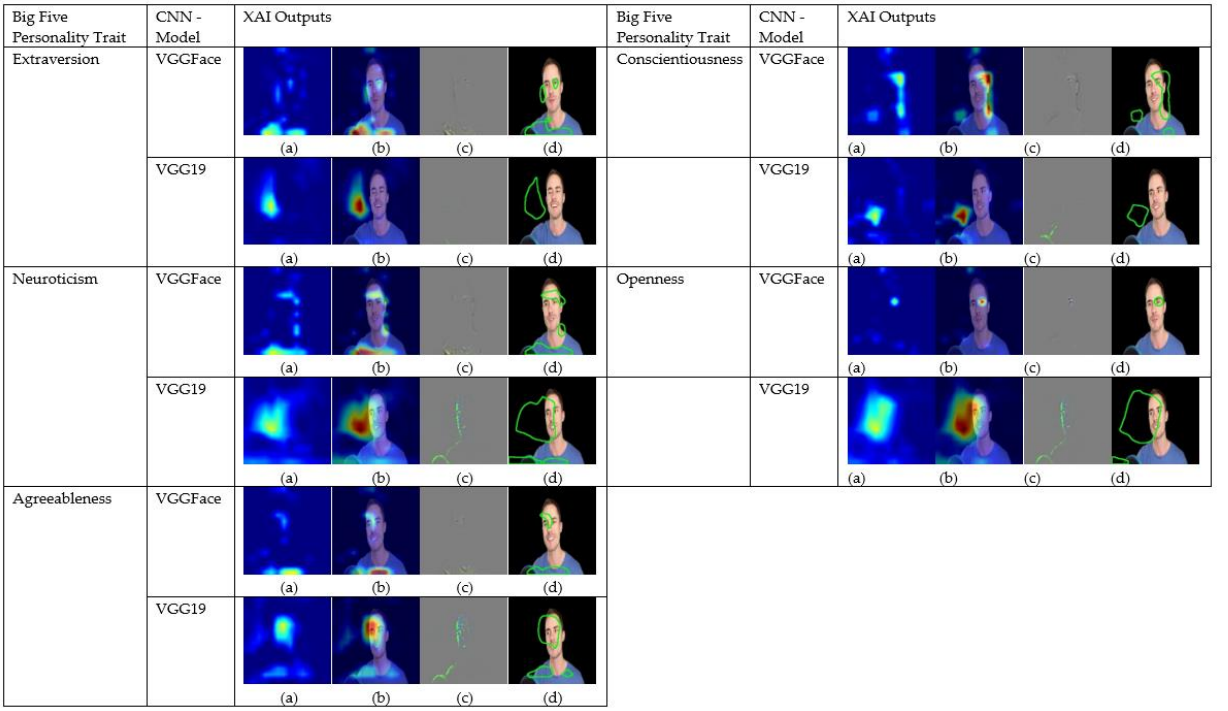

Figure 2: Visualisation Techniques output for VGGFace and VGG19 ((a) – Heatmap, (b) – Grad-CAM, (c) – Guided Grad-CAM, and (d) – Most Contributing Features)

Interpretability vs Faithfulness of CNN-based APD

Gaining insights into the performance of CNN-based APD models is crucial in understanding their underlying mechanisms. The faithfulness of a model refers to its capacity to provide accurate explanations of its internal processes. However, achieving faithfulness often involves a trade-off with interpretability, where faithful visualisations can be less straightforward to interpret and vice versa. Moreover, a trade-off exists between interpretability and the overall performance of the machine learning model (Linardatos et., 2020). As a result, researchers have introduced various machine learning interpretability techniques to address the limitations of existing methods and strike a balance between faithfulness and interpretability.

The F-statistic values obtained from the statistical tests conducted on the ten datasets are summarised in Figure 3. The highest seven F-score values for each model (corresponding to one network with one trait) are highlighted in bold. Precisely, the first column displays the F-score values obtained by each feature for the Extraversion trait using the VGGFace model, while the second column represents the Extraversion trait with the VGG19 model.

The scores demonstrate that facial features play a significant role in determining the network output

using the VGGFace CNN model. Additionally, non-facial features like hair, neck, and feature 15 (Dress, Jewellery, and Hand gestures) are identified as prominent features.

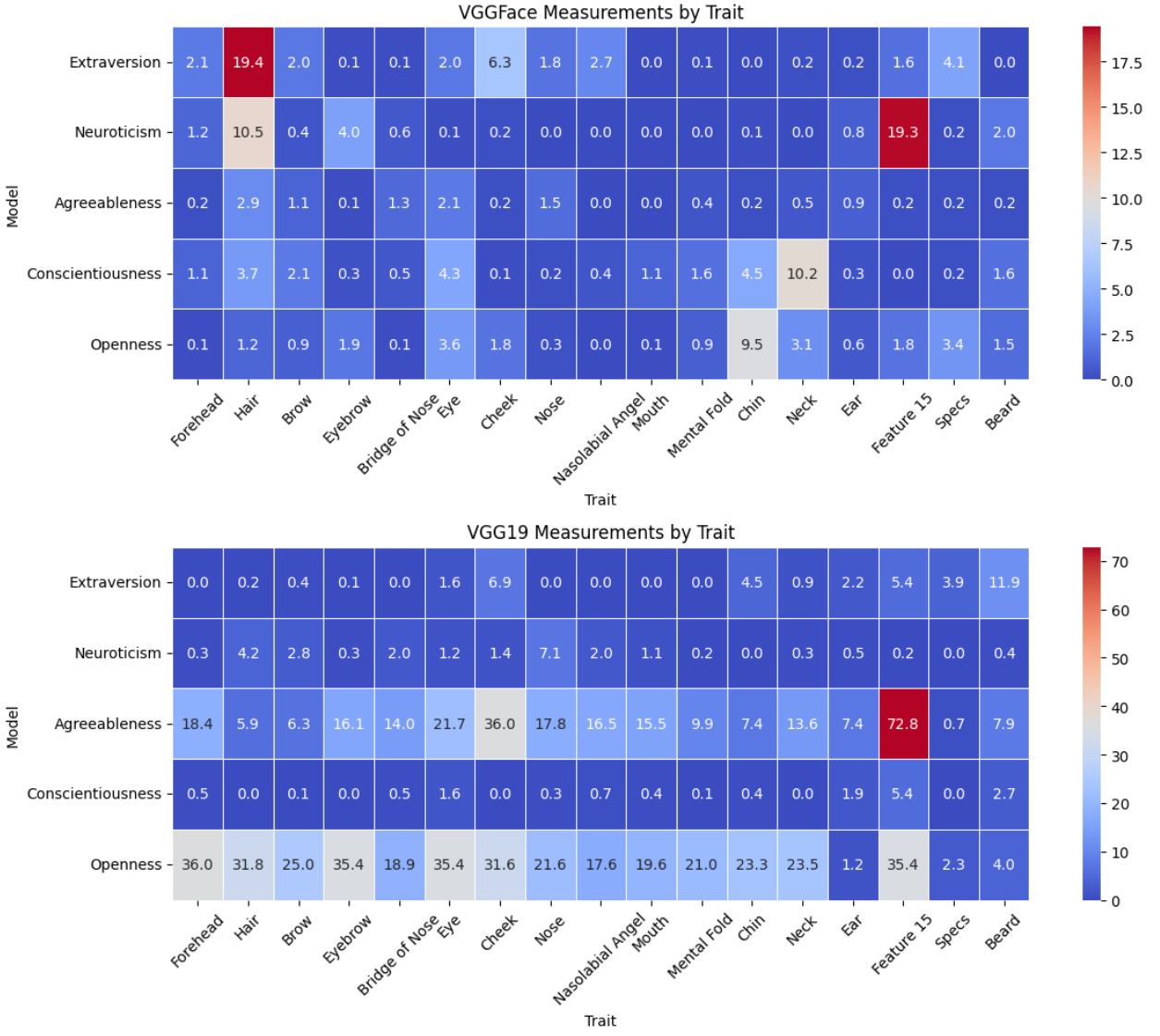

Figure 3: F-Score values which were obtained by each feature from the 'SelectKBest' technique.

The VGGFace model highlights the following features as necessary for calculating personality traits.

- Extraversion: hair, cheek, specs, nasolabial angle, forehead, brow, and eye
- Neuroticism: feature 15, hair, eyebrow, beard, forehead, ear, and bridge of the nose
- Agreeableness: hair, eye, nose, bridge of nose, brow, ear, and neck
- Conscientiousness: neck, chin, eye, hair, brow, chin, and beard
- Openness: chin, eye, specs, neck, eyebrow, cheek, and feature 15

Except for the Neuroticism trait, the eye factor is selected as one of the most contributing features for all other characteristics.

According to Figure 3, the VGG19 model emphasises non-facial features, particularly Feature 15, as indicated by its high F-statistic score. The VGG19 model identifies the following features as crucial for determining personality traits.

- Extraversion: beard, cheek, feature 15, chin, specs, ear, and eye
- Neuroticism: nose, hair, brow, bridge of the nose, nasolabial angle, cheek, and eye
- Agreeableness: feature 15, cheek, eye, forehead, nose, and nasolabial angle
- Conscientiousness: neck, chin, eye, hair, brow, mental fold, and beard features
- Openness: forehead, eyebrow, eye, feature 15, hair, cheek, and brow

In the VGG19 model for the Extraversion trait, some facial features obtained an F-score of "nan," indicating that they were not consistently highlighted across all sample elements. Additionally, the test results for the open-ness trait significantly differed from the other traits, with higher F-statistic values observed for most features. Notably, except for the Neuroticism trait with the VGGFace model, the Eye feature emerged as the main contributing feature for all traits across both architectures. Some models also showed that wearing specs and having a beard were contributing features, but this observation is dependent on the specific sample selection, where subjects need to wear specs or be male with a beard.

The two models' activation of different image regions resulted in distinct contributing features for the same trait, as reflected in Figure 3.

Sensitivity to Noise

To assess the robustness of the interpretability techniques, we introduced Gaussian noise to a sample image (Figure 4) and observed the highlighted features (Figure 5). The VGGFace model displayed minimal differences be-tween the original image and the noise-added image regarding highlighted features. However, when comparing the original image and the noise-added image, the VGG19 model exhibited noticeable variations in the highlighted features for the same trait. Moreover, the outputs of the two models showed relatively consistent results between the original image and the noise-added image (Figure 5, first row).

Figure 4: Sample instance used to check the sensitivity to noise.

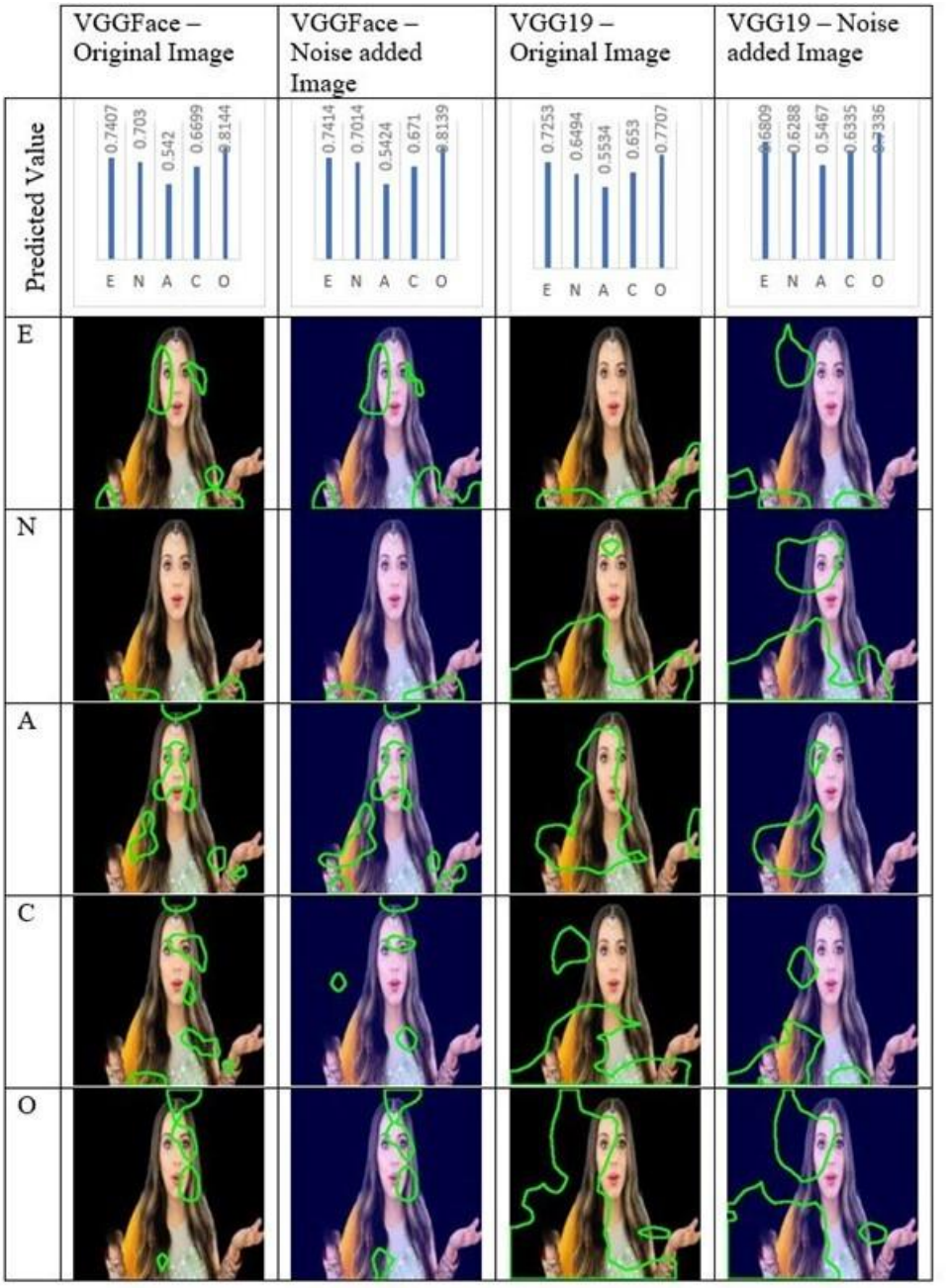

Figure 5: Features highlighted by interpretability techniques on the original image and the noise-added image.

4. DISCUSSION AND CONCLUSION

The results of this study underscore the influence of both facial and non-facial features in CNN-based apparent personality detection (APD), as highlighted by model interpretability techniques. While the VGGFace and VGG19 models identified different contributing features, it is difficult to pinpoint specific personality traits based solely on facial features. This aligns with the findings of Ventura et al. (2017), who also noted that models focused on facial regions, such

as the eyes and mouth, but found it challenging to differentiate these regions in terms of personality traits.

One of the critical gaps in the current literature is the lack of transparency in how APD models make predictions. While previous works (e.g., Wei et al., 2018) have noted that models such as ResNet, DAN, and DAN+ use different image regions for personality prediction, our study adds to this by demonstrating how models like VGGFace prioritize facial features, whereas VGG19 places more emphasis on non-facial features, including background data. This highlights the need for further investigation into how background information affects APD predictions, especially given that even with background removal, some models continue to highlight non-facial regions (Figure 2).

Our study applied Grad-CAM and Guided Grad-CAM techniques to interpret the predictions, and while the VGGFace model showed a clear focus on facial data (excluding Neuroticism), VGG19 appeared more prone to rely on non-facial elements like dress, hand gestures, and jewellery, as reflected in the F-statistic values (Figure 3). This suggests that non-facial features play a significant role in personality prediction, which is an area not sufficiently explored in previous studies. The statistical assessments, particularly the F-statistic values, further emphasize the need to account for these non-facial elements in future APD models.

Despite VGG19 outperforming VGGFace in accuracy (90% vs. 89%), our findings indicate that accuracy alone does not guarantee a more facially interpretable model. The reliance on background information rather than facial features suggests that higher accuracy models may not necessarily offer better insight into the facial cues contributing to personality detection. This opens up a discussion on the trade-offs between model accuracy and explainability.

The findings contribute to filling the gap in the literature by providing a nuanced understanding of the role of non-facial data in personality detection and by demonstrating the limitations of current XAI techniques when applied to APD models. While Grad-CAM and similar techniques are widely used in image classification, their effectiveness in explaining APD outputs may be constrained by dataset biases, as noted in the ECCV challenge dataset used in this study.

In summary, this study highlights the complexity of personality detection models and suggests that future research should focus on enhancing model transparency and explainability, particularly by exploring how non-facial features contribute to predictions. Moreover, benchmarking datasets with greater diversity and more sophisticated interpretability techniques could further advance the field and provide more robust conclusions.

5. ACKNOWLEDGEMENTS:

We would like to express special gratitude to the Faculty of Computing, General Sir John Kotelawala Defence University, for granting us to use the resources to conduct the experiments.

6. REFERENCES

Alhendi, O. (2019). 'Personality Traits and Their Validity in Predicting Job Performance at Recruitment: a Review', *Int. J. of Engineering and Management Sciences*, 4, pp. 222–231. Available at: https://doi.org/10.21791/IJEMS.2019.3.21.

Barezi, E.J. *et al.* (2018). 'Investigating Audio, Visual, and Text Fusion Methods for End-to-End Automatic Personality Prediction', *arXiv preprint*. Available at: http://arxiv.org/abs/1805.00705.

ChaLearn Looking at People - 2016 Looking at People ECCV Challenge (2016). Available at: https://gesture.chalearn.org/2016-looking-at-people-eccv-workshop-challenge.

Ekman, P., Friesen, W. and Ancoli, S. (1980). 'Facial-Sign-Of-Emotional-Experience.pdf', *J. of Personality and Social Psychology*, 39, pp. 1125–1134.

Gatis, D. (2022). 'Rembg'. Available at: https://github.com/danielgatis/rembg.

Güçlütürk, Y. *et al.* (2016). 'Deep Impression: Audiovisual Deep Residual Networks for Multimodal Apparent Personality Trait Recognition', *arXiv preprint*. Available at: https://doi.org/10.1007/978-3-319-49409-8_28.

Gürpınar, F., Kaya, H. and Salah, A.A. (2016). 'Combining Deep Facial and Ambient Features for First Impression Estimation', in G. Hua and H. Jégou (eds) *Computer Vision – ECCV 2016 Workshops*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 372–

385. Available at: https://doi.org/10.1007/978-3-319-49409-8_30.

Ilmini, W. and Fernando, T. (2022). 'Facial Feature Identification in the Deep Learning Based Apparent Personality Detection', in Proc. *2022 2nd Int. Conf. on Advanced Research in Computing (ICARC)*. pp. 49–54. Available at: https://doi.org/10.1109/ICARC54489.2022.9753897.

John, O.P. and Srivastava, S. (1999). *The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives*.

Juhan, M.S. and Ismail, N. (2016). 'Character Design towards Narrative Believability of Boboiboy in the Malaysian Animated Feature Film Boboiboy: The Movie (2016)', in Proc. *Social Sciences. 2nd International Conference on Advanced Research in Economics, Social Sciences & Trade Development*, p. 10.

Kirby, R., Forlizzi, J. and Simmons, R. (2010) 'Affective social robots', *Robotics and Autonomous Systems*, 58(3), pp. 322–332. Available at: https://doi.org/10.1016/j.robot.2009.09.015.

Lee, K.M. *et al.* (2006). 'Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction', *J. of Commun.*, 56(4), pp. 754–772. Available at: https://doi.org/10.1111/j.1460-2466.2006.00318.x.

Li, Y. *et al.* (2020). 'CR-Net: A Deep Classification-Regression Network for Multimodal Apparent Personality Analysis', *Int. J. of Computer Vision*, 128(12), pp. 2763–2780. Available at: https://doi.org/10.1007/s11263-020-01309-y.

Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S. (2020). 'Explainable AI: A Review of Machine Learning Interpretability Methods', *Entropy*, 23(1), p. 18. Available at: https://doi.org/10.3390/e23010018.

Lounsbury, J.W. *et al.* (2008). 'Personality traits and career satisfaction of human resource professionals', *Human Resource Development Int.*, 11(4), pp. 351–366. Available at: https://doi.org/10.1080/13678860802261215.

Mahdy, A., Hereñú, D. and Sumsuddin, M. (2019) 'GitHub - aybassiouny/OpenFaceCpp: C++ implementation for OpenFace library by CMU.' Available at: https://github.com/aybassiouny/OpenFaceCpp.

Mileounis, A., Cuijpers, R.H. and Barakova, E.I. (2015). 'Creating Robots with Personality: The Effect of Personality on Social Intelligence', in J.M. Ferrández Vicente et al. (eds) *Artificial Computation in Biology and Medicine*. Cham: Springer Int. Pub. (Lecture Notes in Computer Science), pp. 119–132. Available at: https://doi.org/10.1007/978-3-319-18914-7_13.

Mujtaba, D.F. and Mahapatra, N.R. (2021). 'Multi-Task Deep Neural Networks for Multimodal Personality Trait Prediction', in Proc. *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*. pp. 85–91. Available at: https://doi.org/10.1109/CSCI54926.2021.00089.

Parkhi, O.M., Vedaldi, A. and Zisserman, A. (2015). 'Deep Face Recognition', in *Proc. of the British Machine Vision Conference 2015*.Swansea: British Machine Vision Association, p. 41.1-41.12. Available at: https://doi.org/10.5244/C.29.41.

Penney, L.M., David, E. and Witt, L.A. (2011). 'A review of personality and performance: Identifying boundaries, contingencies, and future research directions', *Human Resource Management Review*, 21(4), pp. 297–310. Available at: https://doi.org/10.1016/j.hrmr.2010.10.005.

Ponce-López, V. *et al.* (2016). 'ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results', in G. Hua and H. Jégou (eds) *Computer Vision – ECCV 2016 Workshops*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 400–418. Available at: https://doi.org/10.1007/978-3-319-49409-8_32.

Reid, J.A. (2011). 'Crime and Personality: Personality Theory and Criminality Examined', *Inquiries J.*, 3(01). Available at: http://www.inquiriesjournal.com/articles/1690/crime-and-personality-personality-theory-and-criminality-examined.

'scikit-learn/scikit-learn' (2022). scikit-learn. Available at: https://github.com/scikit-learn/scikit-learn/blob/16625450b58f555dc3955d223f0c3b64a5686984/sklearn/feature_selection/_univariate_selection.py.

Selvaraju, R.R. *et al.* (2020) 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization', *Int. J. of Computer Vision*, 128(2), pp. 336–359. Available at: https://doi.org/10.1007/s11263-019-01228-7.

Simonyan, K. and Zisserman, A. (2014). 'Very Deep Convolutional Networks for Large-Scale Image Recognition', *arXiv:1409.1556 [cs]* [Preprint]. Available at: http://arxiv.org/abs/1409.1556.

'sklearn.feature_selection.f_regression' (2022). Available at: https://scikit-learn/stable/modules/generated/sklearn.feature_selection.f_regression.html.

Subramaniam, A. *et al.* (2016). 'Bi-modal First Impressions Recognition Using Temporally Ordered Deep Audio and Stochastic Visual Features', in G. Hua and H. Jégou (eds) *Computer Vision – ECCV 2016 Workshops*. Cham: Springer Int. Pub. (Lecture Notes in Computer Science), pp. 337–348. Available at: https://doi.org/10.1007/978-3-319-49409-8_27.

Ventura, C., Masip, D. and Lapedriza, A. (2017). 'Interpreting CNN Models for Apparent Personality Trait Regression', in Proc. *2017 IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Honolulu, HI, USA: IEEE, pp. 1705–1713. Available at: https://doi.org/10.1109/CVPRW.2017.217.

Waskom, M. (2021). 'seaborn: statistical data visualization', *Journal of Open Source Software*, 6(60), p. 3021. Available at: https://doi.org/10.21105/joss.03021.

Wei, X.-S. *et al.* (2018). 'Deep Bimodal Regression of Apparent Personality Traits from Short Video Sequences', *IEEE Trans. on Affective Computing*, 9(3), pp. 303–315. Available at: https://doi.org/10.1109/TAFFC.2017.2762299.

Wiggins, J. (1996). 'The Five-factor model of personality: theoretical perspectives', *Choice Reviews Online*, 34(03), pp. 34-1846-34–1846. Available at: https://doi.org/10.5860/CHOICE.34-1846.

Yang, K. and Glaser, N. (2017). 'Prediction of Personality First Impressions With Deep Bimodal LSTM', p. 10.

Zammitto, V., DiPaola, S. and Arya, A. (2008) 'A Methodology for Incorporating Personality Modeling in Believable Game Characters', in. *4th Int. Conf. on Game Research and Development*, China, p. 8.

Zhang, C.-L. *et al.* (2016). 'Deep Bimodal Regression for Apparent Personality Analysis', in G. Hua and H. Jégou (eds) *Computer Vision – ECCV 2016 Workshops*. Cham: Springer Int. Pub. (Lecture Notes in Computer Science), pp. 311–324. Available at: https://doi.org/10.1007/978-3-319-49409-8_25.

Zhou, B. *et al.* (2015). 'Learning Deep Features for Discriminative Localization', *arXiv preprint* Available at: http://arxiv.org/abs/1512.04150.