Performance Analysis of State-of-the-Art Deep Learning Models in the Visual-Based Apparent Personality Detection

W. M. K. S. Ilmini[1,2], TGI Fernando[3*]

[1]*Faculty of Graduate Studies, University of Sri Jayewardenepura*

[2]*Intelligent Research Laboratory, Faculty of Computing, General Sir John Kotelawala Defence University*

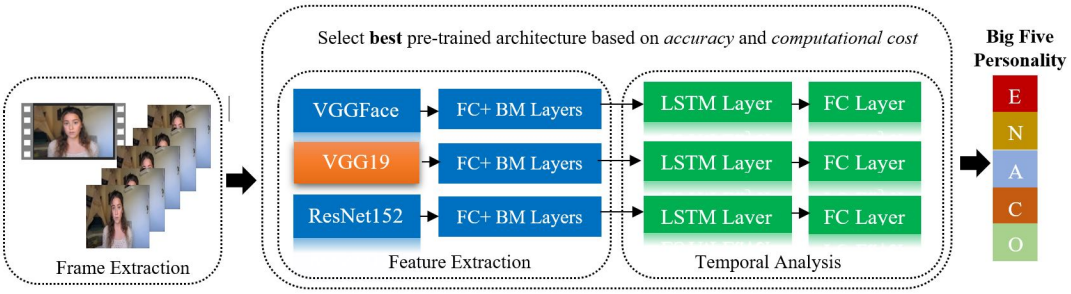
[3]*Department of Computer Science, Faculty of Applied Sciences, University of Sri Jayewardenepura*

Date Received: 07-09-2022 Date Accepted: 23-10-2022

Abstract

This paper analyses the performances of pre-trained deep learning models as feature extractors for apparent personality trait detection (APD) by utilising different statistical methods to find the best performing pre-trained model. Accuracy and computational cost were used to measure the model performance. Personality is measured using the Big Five Personality Schema. CNN-RNN networks were designed using VGG19, ResNet152, and VGGFace pre-trained models to measure the personality with scene data. The models were compared using the mean accuracy attained and the average time is taken for training and testing. Descriptive statistics, graphs, and inferential statistics were applied in model comparisons. Results convey that, ResNet152 model reported the highest mean accuracy in the test dataset (0.9077), followed by VGG19 with 0.9036; VGGFace recorded the lowest (0.8962). ResNet152 consumed more time than other architectures in model training and testing since the number of parameters is comparably higher than the other two architectures involved. Statistical test results prove no significant evidence to conclude that VGG19 and ResNet152 based CNN-RNN models performed differently. This leads to the conclusion that even with a comparably lower number of parameters VGG19 model performed well. The findings reveal that satisfactory accuracy is obtained with a limited number of frames extracted from videos since models achieved more than 90% accuracy.

**Correspondence: tgi@gmail.com*



Keywords: apparent personality traits detection (APD), convolutional neural network (CNN), long-short term memory (LSTM), recurrent neural network (RNN), statistical analysis

1. Introduction

APD draws attention to computer vision and affective computing. APD is applicable in various areas and achieves benefits: social robotics (Lee et al., 2006), (Mileounis et al., 2015), criminology (Reid, 2011), education (Hakimi et al., 2011; Jensen, 2015; ÿz, 2016), and animation movie industry (Juhan and Ismail, 2016; Zammitto et al., 2008). Apparent personality is also helpful in different situations, such as predicting social behaviour, adaptive marketing, affective interfaces, adaptive advertising, job interviews, adaptive tutoring and psychological therapies (Mehta et al., 2019). Not limited to these, computing devices can convert into intelligence devices which can act according to the user's personality. Hence, investigation of personality type/types through appearance has become popular in these fields, and the literature provides much research information. Several studies were conducted to detect the apparent personality using different machine learning algorithms and different sets of features. Before introducing deep learning, researchers used artificial neural networks (ANNs), support vector machines (SVMs), and regression techniques (Ilmini and Fernando, 2017) to detect personality traits from handwriting, speech, video, and social media data. When deep learning came into practice and showed significant improvement in accuracy, personality detection researchers also started to focus on deep learning techniques. Deep learning models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) are prevalent in APD. However, some researchers have focused on applying transfer learning in the model development process. The ChaLearn Looking At People ECCV Challenge (Ponce-López et al., 2016) vastly improved research studies in this area.

The winners of the ChaLearn Looking At People ECCV Challenge, Zhang et al. (2016), Subramaniam et al. (2016), and Güçlütürk et al. (2016), proposed different deep learning architectures to predict the apparent personality. Zhang and others proposed a bi-model deep regression model using visual and audio to predict apparent personality. For visual modality, they used DAN (Descriptor Aggregation Network) and DAN+ architectures composed of VGGFace (Parkhi et al., 2015) pre-trained model as a feature extractor. Audio modality was designed using a fully connected layer followed by a sigmoid layer. They concluded that the DAN+ model's performance is better than the VGGFace, DAN, and ResNet (He et al., 2016) architectures for visual modality. Subramaniam et al. (2016) proposed a bi-model neural network architecture. The bi-model neural network comprises two branches, an audio feature analyser and

visual feature analyser. The results from the two branches are fused at the end to predict personality traits. In the pre-processing of the visual data, the OpenFace C++ library (Baltrusaitis et al., 2016; Mahdy et al., 2019) was used to prevent possible bias from background data. Two neural network architectures were developed, i.e., a volumetric (3D) convolution model and a long short-term memory (LSTM) based model. The results showed that the convolutional model takes less time in training when compared with the LSTM model. The performance of the two models was compared using criteria given by the ChaLearn Looking at People ECCV Challenge. The findings reveal that the LSTM model performs better than the convolutional model. The underlying reason for the better performance of the LSTM model is the capability of learning temporal relationships than convolutional models. Güçlütürk et al. (2016) proposed two residual blocks based on a deep neural network to measure the personality from the audio and visual data. Both residual architectures compromise 17 layers. The visual and audio modality outputs are combined using a fully connected layer to measure the Big five personality scores.

Furthermore, researchers developed different solutions to APD with various deep learning architectures. Table 1 summarises the significant research works.

Table 1: The taxonomy of research conducted in this area using deep learning techniques with ChaLearn by looking at the People's First Impression dataset

Research Work	C.S.*	T.L.**	Modality	Network architecture	Comments
(Zhang et al., 2016)	Y	Y	Scene and Audio	Visual: ResNet VGG, Face DAN and DAN+ Audio: Linear Regressor	Data: 100 frames per video Accuracy: Visual module, VGG-Face: 0.9072, ResNet: 0.9080, DAN: 0.9100, DAN+: 0.9111 (with epoch fusion) and audio regressor: 0.8900
(Subramaniam et al., 2016)	Y	X	Scene and Audio	3D-CNN and LSTM (visual and audio data)	Data: 6 frames per video Accuracy: LSTM: 0.913355 and 3D-CNN: 0.912473
(Güçlütürk et al., 2016)	X	X	Scene and Audio	Residual Architecture with 17 layers for both modalities	Data: all frames Epochs: 900 Mean accuracy: 0.912132
(Gürpınar et al., 2016)	X	Y	Scene, facial region, and audio	CNN with VGGFace, VGG VD-19, and kernel ELM	Data: 2.45M frames for the training and 0.82M for both validation and testing Mean accuracy: 0.9094 for scene modality
(Yang and Glaser, 2017)	Y	Y	Scene and Audio	ResNet32, Bi-model LSTM	Data: 10 frames per video

*Correspondence: tgi@gmail.com

Research Work	C.S.*	T.L.**	Modality	Network architecture	Comments
				with L1 and L2 Loss	Accuracy: LSTM L2 - 0.9083, LSTM L1 - 0.8963 and ResNet - 0.8935
(Barezi et al., 2018)	X	Y	Scene, Audio, and Text	Three CNNs for three modalities with VGGFace to extract features from the visual's modality	Data: Random one frame per video Accuracy: 0.8965, Three modalities score on the validation dataset: 0.9062
(Aslan and Güdükbay, 2019)	Y	Y	Audio, Scene, facial region, and transcriptions	LSTM networks with ResNet101 and VGGish pre-trained deep learning models	Data: 2.5M frames for training Accuracy: Scene data with ResNet CNN-RNN architecture (six LSTM layers) achieved 0.9116. The Multi-feature model achieved 0.9163
(Li et al., 2020)	X	Y	Scene, facial region, audio, and transcriptions	Deep classification-regression network (CR-Net) based on ResNet34	They proposed a "Bell Loss" to overcome the drawbacks of the L1 and L2 loss in the current problem and achieved a 0.9188 mean trait score
(Aslan and Güdükbay, 2019)	X	Y	Audio, Scene, facial region, and transcriptions	LSTM networks with ResNet101 and VGGish pre-trained deep learning models	Data: One frame per second Accuracy: Ambient data model 91.1% mean accuracy, and the multi-model 91.8% average mean accuracy
(Mujtaba and Mahapatra, 2021)	X	Y	Scene, facial region, audio, and transcriptions	Multi-task deep neural network based on VGG19 and VGGFace	Accuracy: 0.9114 average mean accuracy

*Column 2: C.S. refers to a comparison study conducted. **Column 3: T.L. refers to transfer learning applied. The cell values Y – represents "Yes" and X represents "No."

Most researchers used bi-model architectures to predict the apparent personality with audio and visual modalities. Few researchers used visual, audio, and transcription data (audio streams converted to text), whereas some extracted a comparably high number of frames from the videos in the visual modality. In these works, the researchers have focused on the model's accuracy, not paying much attention to the

problem's complexity. These methods lead to a high computational cost. Out of all these works, most of them used transfer learning as a feature extraction module. Literature proves that few research works conducted in this area tend to compare the models based on accuracy. Nevertheless, a lack of work focused on comparing the models based on the computational cost of the problem. They concluded that specific pre-trained models are better than others based on the accuracy.

Automatic apparent personality detection is a combination of psychology and machine learning algorithms. Theories of psychology describe a person's appearance and behaviour depending on the person's inner aspects, that is, personality. There are various personality schemas defined in psychology from time to time, but the most acceptable one is the *Big Five Personality Schema* (Wiggins, 1996). The Big Five Personality Schema defines the personality of a human using a vector of five values: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N), abbreviated as OCEAN or CANOE. Each personality trait value has its meaning with subordinate personality types. John and Srivastava defined the Big Five Personality dimensions and correlated trait adjectives (John and Srivastava, 1999).

- Openness (openness vs closeness): Describes a person's level of imagination, feelings, and ideas. Having a high value for this trait leads to curiosity and creativity. Low scores mean they like routines and are more practical people.
- Conscientiousness (conscientiousness vs lack of direction): Describes competence, diligence, carefulness, and goal driven. A high score for this trait means hardworking and organised, while a low score leads to careless and disorganised behaviour.
- Extraversion (extraversion vs introversion): Describes a person's social behaviour, whether a person likes to be with others and enjoys being around people. People with a high score for this trait are warm and seek adventures, while those with a low score show a quiet and withdrawn nature.
- Agreeableness (agreeableness vs antagonism): Describes trustfulness and supportive nature. A high score leads to a supportive and trusting nature, and a low score reflects uncooperative and suspicious behaviour.
- Neuroticism (neuroticism vs emotional stability): Describes negative feelings such as anxiety, anger, and depression. A high score for this trait implies that the person is emotionally unstable. A low score leads to a calm nature.

In the field of deep learning major challenge is training cost. It needs a comparably high amount of time to train the model, increased processing power for training, and a large amount of data to obtain a better prediction model. To solve this vital issue, researchers in this field use *transfer learning,* which uses previously acquired knowledge in a new domain (Lu et al., 2015). The current research employed two models trained on the ILSVRC (Russakovsky et al., 2015) and VGGFace (Parkhi et al., 2015) trained on

*Correspondence: tgi@gmail.com

face recognition in the APD problem, and their performances were compared based on accuracy and computational cost.

2. Contributions

The contributions of this work to the APD field from visual data are as follows:

- This research work primarily aims to obtain a comparably high accuracy for APD with fewer features and parameters, which can reduce the computational cost while achieving high accuracy.
- Use descriptive and inferential statistics to select the suitable pre-trained model based on accuracy and computational cost.

The rest of the paper is organised as follows: Section 2 describes the experimental study conducted to find the best pre-trained deep learning model in the APD with visual data, Section 3 states the results obtained, and Section 4 includes the discussion and conclusion.

3. Methodology

This section discusses the experimental studies conducted to find the best pre-trained model as a feature extractor in CNN-RNN (LSTM) based APD in terms of accuracy and computational cost. All experiments were conducted on a precision server with Nvidia RTX 3090 24 GB.

3.1. Dataset: The First Impressions V2

This research work uses ChaLearn Looking at People First Impression dataset (Ponce-López et al., 2016) for the experiments. This is the only large publicly available dataset with personality annotations. Ten non-overlapping frames were captured from each video and stored with the ground truth values. The dataset sizes after capturing frames:

- Training Dataset: 6,000 (number of videos) x 10 (10 frames per each) = Total 60,000 frames
- Validation Dataset: 2,000 (number of videos) x 10 (10 frames per each) = Total 20,000 frames
- Test Dataset: 2,000 (number of videos) x 10 (10 frames per each) = Total 20,000 frames

The model is evaluated using one minus mean absolute error, i.e., the absolute distance between the predicted value and ground truth value. The mean accuracy of the model is calculated using Equation 1:

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

Where N is the number of videos, M = 5 number of personality traits, the target is the respective ground-truth value, and output is the predicted value from the model for a given video. We used Equation 1 to evaluate the network's performance because it is the evaluation matrix introduced at the ChaLearn First Impression Challenge (Ponce-López et al., 2016).

3.2. Network Architecture

The current study used three pre-trained deep learning models: ResNet152 (He et al., 2016), VGGFace (Parkhi et al., 2015), and VGG19 (Simonyan and Zisserman, 2014). We selected these architectures because Resnet152 achieved a 94.046% of the highest top-five accuracy in the ILSVRC out of ResNet

variants. VGG19 gained the highest (90.876%) accuracy out of all the VGG variants in the ILSVRC. VGGFace pre-trained model was used as it has received initial training on face recognition, unlike the ILSVRC-based pre-trained models, which are general-purpose classifiers. Past researchers have also proved that VGGFace-based DAN and DAN+ architectures performed well in APD.

3.3. Pre-trained deep learning models based on CNN-RNN architecture

Figure illustrates the network architecture used in this study. We have designed three separate CNN-RNN models with VGG-Face, VGG19, and ResNet152 state-of-the-art deep learning models. At the end of the CNN part, we added an RNN part for all networks by adding one LSTM layer. The main reason behind using RNN in the current research is to capture temporal information, which is essential in video analysis.

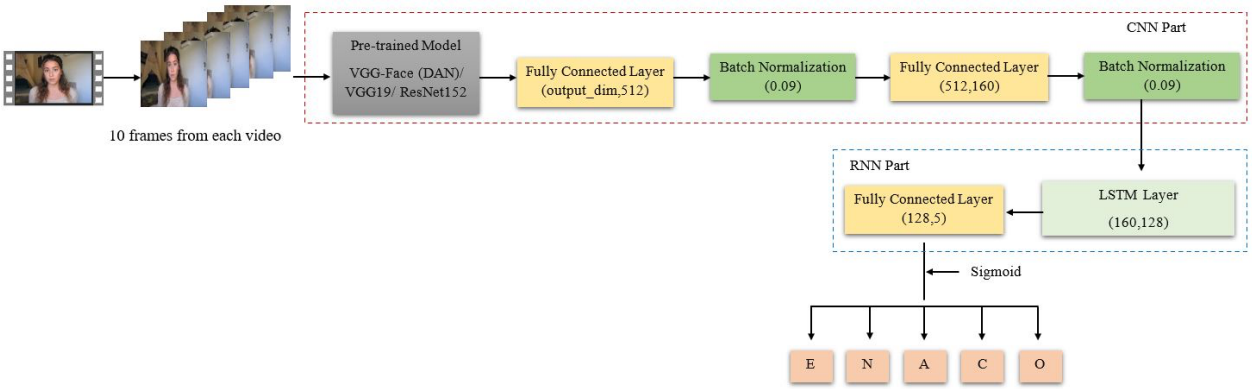


Figure 1. Network Architecture.

The total number of parameters of each model:

- VGGFace - Total parameters: 15,471,717 (until the 25th layer, freeze all the layers)
- VGG19 - Total parameters: 33,113,189 (until 49th layer, freeze all the layers)
- ResNet152 - Total parameters: 59,388,005 (until the fourth layer's last convolutional layer, set all the layers freeze)

3.4. Network Parameters

We initially experimented with these models with different configurations and found the values/methods more appropriate for the APD problem. After running a few tests, as the model tends to overfit, we set up an early stopping counter of 20 epochs for the validation loss. Since this is a regression problem, we applied the most common loss functions. All implementations were conducted using PyTorch (Paszke et al., 2019) machine learning library. The finalised parameters of the current performance study are as follows:

- Learning Rate: 1e-5

*Correspondence: tgi@gmail.com

- Epochs: 150 with an early stopping count of 20
- Optimiser: Adam
- Batch Size: 8
- Loss: L1 (Mean Absolute Error) and L2 (Mean Squared Error) Loss

The values assigned for the above parameters were finalised after conducting several experiments with the dataset. Then the values were selected based on the model's performance and the computational capability of the device used to train the model. The networks were trained ten times with the abovementioned parameters and stored to compare the performances. We used the following parameters to compare the models:

- Mean Accuracy
- The time necessary to test the models with the test dataset (2,000 videos in the test dataset)
- The average time to train the models with the training dataset (per one epoch)
- The number of epochs ran to achieve the best accuracy

Descriptive and inferential statistics were used to find the best-performed model with given parameters. The descriptive statistics used are minimum, maximum, mean, and standard deviation. Inferential methods used are the Friedman test (nonparametric multiple sample comparison test) (Friedman, 1937) and Dunn's post hoc test (Dunn, 1961). Dunn's post hoc test was used to find which models differed if the null hypothesis was rejected in the multiple comparison test. All statistical tests were conducted using the SPSS tool (IBM, 2019).

4. Results

This section summarises the results of this study. First, the mean accuracies reported by each model in the testing datasets are outlined, followed by summarising the time consumed by each model to complete the testing and training processes. Next, the trait-wise analysis is conducted to understand more about the accuracies obtained.

4.1. Statistical Analysis of Mean Accuracies Obtained in the Test Dataset

Table 1: Descriptive statistics of Mean Accuracy achieved by each architecture with two loss functions in the test dataset

Architecture	Minimum[1]	Maximum[2]	Mean[3]	Std. Deviation[4]
VGGFace_L1	0.8935	0.8962	0.8949	8.6 x e^{-4}
VGGFace_L2	0.8920	0.8955	0.8944	1.1 x e^{-3}
VGG19_L1	0.9006	0.9034	0.9022	9.0 x e^{-4}

[1] All values are truncated to four decimal places
[2] All values are truncated to four decimal places
[3] All values are truncated to four decimal places
[4] Std. deviation is in the scientific format

VGG19_L2	0.8982	0.9036	0.9018	1.5 x e^{-3}
ResNet_L1	0.9017	0.9066	0.9040	1.6 x e^{-3}
ResNet_L2	0.8992	0.9077	0.9037	3.0 x e^{-3}

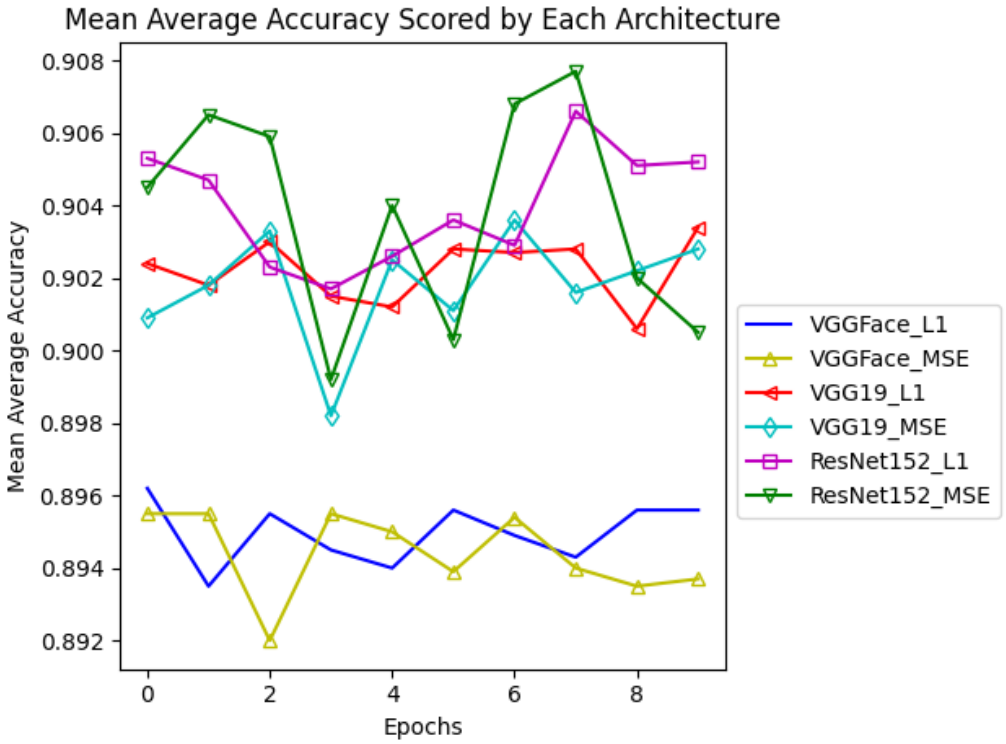


Figure 1. Mean accuracies obtained by each model with two loss functions in the test dataset

Table 1 demonstrate that ResNet152 achieved the highest average and minimum and maximum mean accuracies than VGG19 and VGGFace models. The VGGFace-based model showed the least accuracy compared to other architectures. Figure 1 also verifies that VGG19 and ResNet152 models are better than VGGFace. Scores of VGG19 and ResNet152 lie in the same range, but in some epochs, ResNet152 achieved the highest accuracy among other models (Figure 1). The Friedman test was conducted with the null hypothesis that all models are identical. Table 2 summarises the results.

*Correspondence: tgi@gmail.com

Table 2: Friedman Test Results on Mean Accuracies achieved by each architecture with two loss functions (a) Friedman Test Mean Rank Values, (b) Friedman Test Statistics

Architecture	Mean Rank
VGGFace_L1	1.60
VGGFace_L2	1.40
VGG19_L1	3.95
VGG19_L2	3.95
ResNet_L1	5.20
ResNet_L2	4.90

(a)

Test Statistics	
N	10
Chi-Square	38.037
df	5
Asymp. Sig.	$3.709 \times e^{-7}$

(b)

The Friedman test results clarify with the significance level of $3.709 \times e^{-7}$ (< 0.05) that the pre-trained models behave differently (Table 2 (a) and (b)). Furthermore, the mean rank values obtained by each model convey that the ResNet152 model has the highest mean rank value by achieving the maximum mean accuracy.

Then Dunn's post hoc test was conducted to verify which models behave differently. The results (Appendix Table A 1) indicated that VGGFace with L2 loss, and VGG19, VGGFace, and ResNet152 models perform differently. However, the post hoc tests do not convey that the VGG19 and ResNet152 models perform differently. The Friedman and post hoc test results make it challenging to conclude that loss functions behaved differently within the same architecture (Table A 1).

4.2 Statistical Analysis of Time Taken by Each Model to Test and Train

Table 3: Descriptive statistics - Time (in seconds) taken by each model to process the test and training datasets

	Architecture	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
	VGGFace_L1	**55.5**	**58.2**	**56.360**	0.9902
	VGGFace_L2	55.3	58.7	56.960	1.2989

Test Dataset[5]	VGG19_L1	64.5	92.2	68.510	8.3888
	VGG19_L2	64.0	65.9	65.300	0.6307
	ResNet_L1	**92.9**	**99.0**	**96.240**	1.8044
	ResNet_L2	91.7	97.0	94.600	1.8006
Training Dataset[6]	VGGFace_L1	**173.0**	**177.2**	**173.940**	1.2686
	VGGFace_L2	173.4	178.8	176.170	1.8080
	VGG19_L1	201.2	207.9	204.780	2.3103
	VGG19_L2	199.6	205.8	203.150	1.9271
	ResNet_L1	**747.8**	**758.7**	**751.650**	3.5177
	ResNet_L2	743.9	754.5	749.040	3.1935

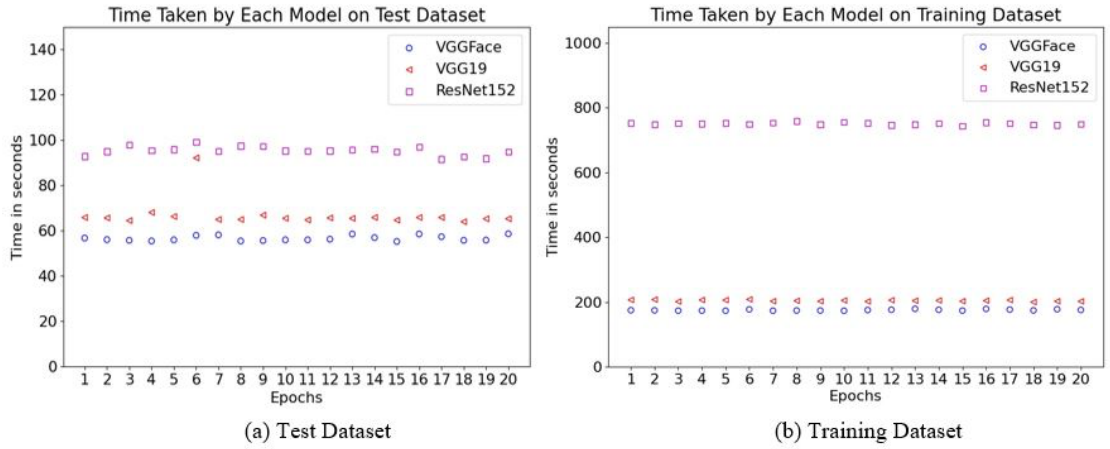


Figure 2: Time (in seconds) taken by each model to run the test and training datasets

Table 3 and Figure 2 show that ResNet152 takes comparably more time than VGG19 and VGGFace in model training and testing. The VGGFace-based model took the least time compared to the other two architectures. The Friedman test results also confirm with the significance level of $6.658 \times e^{-9}$ (<0.05) and $3.7479 \times e^{-9} < 0.05$ that the pre-trained models behave differently (a) and (b)).

[5]Per epoch average time is recorded in seconds
[6] Per epoch average time is recorded in seconds

*Correspondence: tgi@gmail.com

ResNet152 recorded the highest mean rank value. The post hoc test results (Appendix Table A 2 and Table A 3) indicate that VGGFace and ResNet152 models perform differently. Nevertheless, the post hoc tests do not reveal that VGG19 and ResNet152 models performed differently. Figure 2 shows an extreme data point in the VGG19 samples, which recorded higher time than other instances in the VGG19. This extreme data point could have resulted from the unusual data loading behaviour because, occasionally, it takes more time than usual to load data to feed the network.

Table 5: Friedman test results on time taken by each architecture with two loss functions to complete one test run on the test and training datasets (a) Friedman Test Mean Rank Values, (b) Friedman Test Statistics

Test Dataset	**Architecture**	**Mean Rank**
	VGGFace_L1	**1.30**
	VGGFace_L2	1.70
	VGG19_L1	3.75
	VGG19_L2	3.25
	ResNet_L1	**5.70**
	ResNet_L2	5.30
Training Dataset	VGGFace_L1	1.00
	VGGFace_L2	2.00
	VGG19_L1	3.80
	VGG19_L2	3.20
	ResNet_L1	**5.70**
	ResNet_L2	5.30

(a)

Test Statistics		
Test Dataset	N	10
	Chi-Square	46.662
	df	5
	Asymp. Sig.	**$6.658 \times e^{-9}$**
Training Dataset	N	10
	Chi-Square	47.886
	df	5
	Asymp. Sig.	**$3.7479 \times e^{-9}$**

(b)

4.3. Number of Epochs to Converge to the Highest Mean Accuracy

The statistical test results on the number of epochs ran to achieve the highest mean accuracy, which shows no significant difference between models. Therefore, the null hypothesis is rejected with the p-value of $(0.444) > 0.05$.

5. Discussion and Conclusion

This research aims to find the most suitable pre-trained deep learning model for CNN-RNN architecture based on the accuracy and the computational cost. Most previous research works in this area have used two (visual and audio) or three (visual, audio, and text) separate networks in the APD problem and concluded that the visual data affects more to the prediction more than audio and text data (Barezi et al., 2018). In the visual modality, few researchers have used raw frames extracted from the videos (Güçlütürk et al., 2016; Zhang et al., 2016) or face data extracted from raw frames (background removed)

(Subramaniam et al., 2016). Further, a few researchers have used raw frames and face-aligned data both in the visual modality (Aslan and Güdükbay, 2019; Gürpınar et al., 2016; Li et al., 2020). Using many features increases the problem's complexity. Hence, we focused on predicting the apparent personality with one modality.

Barezi et al. (2018) concluded that visual data are more relevant than the audio and transcription data in the APD problem, with a 0.8965 mean accuracy. The current study also emphasises the past research findings by achieving 0.9077 mean accuracy for the APD problem with scene data. The results of this study further explain that ILSVRC pre-trained models can be generalised in the current research work. The ILSVRC pre-trained models used in this study were chosen based on their performances on the ILSVRC. Past research works in this area also confirm that the ILSVRC models can solve the APD problem with a significant amount of data. Remarkably, in the current study, ten non-overlapping frames from each video achieved the best mean accuracy of 0.9077 (ResNet152 with L2 Loss).

Nevertheless, few researchers have extracted a comparably higher number of frames from the videos and fed them into the network. Zhang et al. (2016) obtained 100 frames from each video and achieved 0.9111 for visual data, while Güçlütürk et al. (2016) extracted all frames from each video and ran 900 epochs to reach 0.912132 with visual and audio features. Gürpınar et al. (2016) achieved a mean accuracy of 0.9094 for scene modality, and they used 2.45M frames for the training and 0.82M for both validation and testing. Yang and Glaser (2017) used LSTM-based architecture with L2 loss for visual and audio data and attained a 0.9083 accuracy. Aslan and Güdükbay (2019) recorded a 0.9116 mean accuracy for scene data with ResNet CNN-RNN architecture with six LSTM layers. They used 2.5M data in the training dataset.

These past research findings confirm that they achieved the highest accuracy with a comparably higher amount of data. Also, Aslan and the team (2021) used a complex architecture with six LSTM layers in the ambient modality and achieved a 91.1% average mean accuracy. Mujtaba and Mahapatra (2021) used ambient, facial, audio, and transcription data to achieve a 91.14% mean accuracy for the proposed MTDNN. This information implies that a few studies have used complex architecture with more features to achieve the best accuracy.

In the statistical analysis, the results of Dunn's test are incapable of identifying the difference in some situations, e.g., the post hoc results on time taken for training. Even though a considerable difference exists between VGG19 and ResNet152, the post hoc test does not confirm the fact. Derrac and others (2011) mentioned that the multiple comparison techniques might fail to identify a significant difference between some algorithms.

Regarding the number of parameters, the VGGFace model has the least, and ResNet152 has the highest number of parameters than VGG19. This fact affects the time complexity of model training and testing. ResNet152 consumed the highest time in model training and testing, followed by VGG19 and VGGFace

*Correspondence: tgi@gmail.com

(Table 3). Finally, to decide which model performs well, it is necessary to concentrate on computational cost and accuracy.

In terms of accuracy, ResNet152 and VGG19 models performed well. Furthermore, of those two, ResNet152 shows maximum accuracy with a value of 0.9077. However, the test time is comparably high in ResNet152 than in other architectures. However, in seconds per video, it is 0.0473, which is a small value. Also, the ResNet152 model took comparably high time to train the network than the other two architectures. VGG19-based architecture has fewer parameters than ResNet152-based architecture based on the number of network parameters. Overall, we can conclude that the VGG19 is the best-performing model in terms of accuracy and computational cost. The VGG19 with L1 loss achieved 0.9034, and VGG19 with L2 loss achieved 0.9036 mean accuracies within a comparably short time.

References

Aslan, S., Güdükbay, U., 2019. Multimodal Video-based Apparent Personality Recognition Using Long Short-Term Memory and Convolutional Neural Networks. arXiv:1911.00381 [cs].

Aslan, S., Güdükbay, U., Dibeklioğlu, H., 2021. Multimodal assessment of apparent personality using feature attention and error consistency constraint. Image and Vision Computing 110, 104163. https://doi.org/10.1016/j.imavis.2021.104163

Baltrusaitis, T., Robinson, P., Morency, L.-P., 2016. OpenFace: An open source facial behavior analysis toolkit, in: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). Presented at the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, Lake Placid, NY, USA, pp. 1–10. https://doi.org/10.1109/WACV.2016.7477553

Barezi, E.J., Kampman, O., Bertero, D., Fung, P., 2018. Investigating Audio, Visual, and Text Fusion Methods for End-to-End Automatic Personality Prediction. arXiv:1805.00705 [cs, stat].

Derrac, J., García, S., Molina, D., Herrera, F., 2011. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. Swarm and Evolutionary Computation 1, 3–18. https://doi.org/10.1016/j.swevo.2011.02.002

Dunn, O.J., 1961. Multiple comparisons among means. Journal of the American Statistical Association 56, 52–64. https://doi.org/10.2307/2282330

Friedman, M., 1937. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. Journal of the American Statistical Association 32, 675. https://doi.org/10.2307/2279372

Güçlütürk, Y., Güçlü, U., van Gerven, M.A.J., van Lier, R., 2016. Deep Impression: Audiovisual Deep Residual Networks for Multimodal Apparent Personality Trait Recognition. arXiv:1609.05119 [cs]. https://doi.org/10.1007/978-3-319-49409-8_28

Gürpınar, F., Kaya, H., Salah, A.A., 2016. Combining Deep Facial and Ambient Features for First Impression Estimation, in: Hua, G., Jégou, H. (Eds.), Computer Vision – ECCV 2016 Workshops, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 372–385. https://doi.org/10.1007/978-3-319-49409-8_30

Hakimi, S., Hejazi, E., Lavasani, M.G., 2011. The Relationships Between Personality Traits and Students' Academic Achievement. Procedia - Social and Behavioral Sciences, The 2nd International Conference on Education and Educational Psychology 2011 29, 836–845. https://doi.org/10.1016/j.sbspro.2011.11.312

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

IBM, 2019. IBM SPSS Software | IBM.

Ilmini, W.M.K.S., Fernando, T.G.I., 2017. Computational personality traits assessment: A review, in: 2017 IEEE International Conference on Industrial and Information Systems (ICIIS). Presented at the 2017 IEEE International Conference on Industrial and Information Systems (ICIIS), IEEE, Peradeniya, pp. 1–6. https://doi.org/10.1109/ICIINFS.2017.8300416

Jensen, M., 2015. Personality Traits, Learning and Academic Achievements. JEL 4, 91. https://doi.org/10.5539/jel.v4n4p91

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

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

Lee, K.M., Peng, W., Jin, S.-A., Yan, C., 2006. Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction. Journal of Communication 56, 754–772. https://doi.org/10.1111/j.1460-2466.2006.00318.x

Li, Y., Wan, J., Miao, Q., Escalera, S., Fang, H., Chen, H., Qi, X., Guo, G., 2020. CR-Net: A Deep Classification-Regression Network for Multimodal Apparent Personality Analysis. Int J Comput Vis 128, 2763–2780. https://doi.org/10.1007/s11263-020-01309-y

Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., Zhang, G., 2015. Transfer learning using computational intelligence: A survey. Knowledge-Based Systems 80, 14–23. https://doi.org/10.1016/j.knosys.2015.01.010

Mahdy, A., Hereñú, D., Sumsuddin, M., 2019. GitHub - aybassiouny/OpenFaceCpp: C++ implementation for OpenFace library by CMU.

Mehta, Y., Majumder, N., Gelbukh, A., Cambria, E., 2019. Recent Trends in Deep Learning Based Personality Detection. arXiv:1908.03628 [cs].

Mileounis, A., Cuijpers, R.H., Barakova, E.I., 2015. Creating Robots with Personality: The Effect of Personality on Social Intelligence, in: Ferrández Vicente, J.M., Álvarez-Sánchez, J.R., de la Paz López, F., Toledo-Moreo, Fco.J., Adeli, H. (Eds.), Artificial Computation in Biology and Medicine, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 119–132. https://doi.org/10.1007/978-3-319-18914-7_13

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

Parkhi, O.M., Vedaldi, A., Zisserman, A., 2015. Deep Face Recognition, in: Procedings of the British Machine Vision Conference 2015. Presented at the British Machine Vision Conference 2015, British Machine Vision Association, Swansea, p. 41.1-41.12. https://doi.org/10.5244/C.29.41

**Correspondence: tgi@gmail.com*

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. https://doi.org/10.48550/arXiv.1912.01703

Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., Baró, X., Escalante, H.J., Escalera, S., 2016. ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results, in: Hua, G., Jégou, H. (Eds.), Computer Vision – ECCV 2016 Workshops, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 400–418. https://doi.org/10.1007/978-3-319-49409-8_32

Reid, J.A., 2011. Crime and Personality: Personality Theory and Criminality Examined. Inquiries Journal 3.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 115, 211–252. https://doi.org/10.1007/s11263-015-0816-y

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs].

Subramaniam, A., Patel, V., Mishra, A., Balasubramanian, P., Mittal, A., 2016. Bi-modal First Impressions Recognition using Temporally Ordered Deep Audio and Stochastic Visual Features. arXiv:1610.10048 [cs].

Wiggins, J., 1996. The Five-factor model of personality: theoretical perspectives. Choice Reviews Online 34, 34-1846-34–1846. https://doi.org/10.5860/CHOICE.34-1846

Yang, K., Glaser, N., 2017. Prediction of Personality First Impressions With Deep Bimodal LSTM 10.

ÿz, H., 2016. The Importance of Personality Traits in Students□ Perceptions of Metacognitive Awareness. Procedia - Social and Behavioral Sciences, International Conference on Teaching and Learning English as an Additional Language, GlobELT 2016, 14-17 April 2016, Antalya, Turkey 232, 655–667. https://doi.org/10.1016/j.sbspro.2016.10.090

Zammitto, V., DiPaola, S., Arya, A., 2008. A Methodology for Incorporating Personality Modeling in Believable Game Characters. Presented at the 4th International Conference on Game Reseaarch and Development, China, p. 8.

Zhang, C.-L., Zhang, H., Wei, X.-S., Wu, J., 2016. Deep Bimodal Regression for Apparent Personality Analysis, in: Hua, G., Jégou, H. (Eds.), Computer Vision – ECCV 2016 Workshops, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 311–324. https://doi.org/10.1007/978-3-319-49409-8_25

Appendix A

Statistical Analysis Results - Pairwise Comparison Results

Pairwise comparison results in the Mean Accuracies achieved by each architecture with two loss functions

The hypothesis was tested with a significance value of 0.05:

Null Hypothesis: Distributions of all models are the same

Null Hypothesis is rejected with the p-value = 0.000 (less than 1 x e^{-3}) < 0.05

Table A 1 summarises the models which differ from each other with their significance values.

Table A 1: Dunn's post hoc test results for the Mean Accuracies achieved by each architecture with two loss functions

Architectures Considered[7]	Significance Level
VGGFace-L2-Acc AND VGG19-L1-Acc	0.035
VGGFace-L2-Acc AND VGG19-L2-Acc	0.035
VGGFace-L2-Acc AND ResNet-L2-Acc	0.000 (less than 1 x e^{-3})
VGGFace-L2-Acc AND ResNet-L1-Acc	0.000 (less than 1 x e^{-3})
VGGFace-L1-Acc AND ResNet-L2-Acc	0.001
VGGFace-L1-Acc AND ResNet-L1-Acc	0.000 (less than 1 x e^{-3})

Pairwise comparison results in time consumed by each model to assess the test dataset

The hypothesis was tested with a significance value of 0.05:

Null Hypothesis: Distributions of all models are the same

Null hypothesis is rejected with the p-value = 0.000 (less than 1 x e^{-3}) < 0.05.

Models which differ from each other are summarised with their significance values in Table A 2.

[7] Models which are statistically significant were only recorded in the table

*Correspondence: tgi@gmail.com

Table A 2: Dunn's post hoc test results on time consumed by each model to assess the test dataset

Architectures Considered[8]	Significance Level
VGGFace-L1-Time AND ResNet-L2-Time	0.000 (less than 1 x e^{-3})
VGGFace-L1-Time AND ResNet-L1-Time	0.000 (less than 1 x e^{-3})
VGGFace-L2-Time AND ResNet-L2-Time	0.000 (less than 1 x e^{-3})
VGGFace-L2-Time AND ResNet-L1-Time	0.000 (less than 1 x e^{-3})

Pairwise comparison results in the average time per epoch taken by each model to train the model

The hypothesis was tested with a significance value of 0.05:

Null Hypothesis: Distributions of all models are the same

Null Hypothesis is rejected with the p-value = 0.000 (less than 1 x e^{-3}) < 0.05

Models which differ from each other are summarised with their significance values in Table A 3.

Table A 3: Dunn's post hoc test results on the average time per epoch taken by each model to train the model

Architectures Considered[9]	Significance Level
VGGFace-L1-Time AND VGG19-L2-Time	0.012
VGGFace-L1-Time AND ResNet-L2-Time	0.000 (less than 1 x e^{-3})
VGGFace-L1-Time AND ResNet-L1-Time	0.000 (less than 1 x e^{-3})
VGGFace-L2-Time AND ResNet-L2-Time	0.001
VGGFace-L2-Time AND ResNet-L1-Time	0.000 (less than 1 x e^{-3})
VGG19-L2-Time AND ResNet-L1-Time	0.042

[8] Models which are statistically significant were only recorded in the table, time is measured in seconds

[9] Models which are statistically significant were only recorded in the table. Time is measured in seconds, and the average time per epoch is recorded.