Electroencephalographic Correlates in Synthetic and Real Emotional Face Stimulation

Pietro Tarchi[1,*], Federico Calà[1], Lorenzo Frassineti[2] and Antonio Lanatà[1]

***Abstract*— This work reports on physiological electroencephalographic (EEG) correlates in cognitive and emotional processes within the discrimination between synthetic and real faces visual stimuli. Human perception of manipulated data has been addressed in the literature from several perspectives. Researchers have investigated how the use of deep fakes alters people's ability in face-processing tasks, such as face recognition. Although recent studies showed that humans, on average, are still able to correctly recognize synthetic faces, this study investigates whether those findings still hold considering the latest advancements in AI-based, synthetic image creation. Specifically, 18-channels EEG signals from 21 healthy subjects were analyzed during a visual experiment where synthetic and actual emotional stimuli were administered. According to recent literature, participants were able to discriminate the real faces from the synthetic ones, by correctly classifying about 77% of all images. Preliminary encouraging results showed statistical significant differences in brain activation in both stimuli (synthetic and real) classification and emotional response.**

***Index Terms* — EEG, deep fakes, face recognition.**

I. INTRODUCTION

Deep fakes are realistic digital media that portray false information, which can be created from scratch or by modifying authentic content. Media advanced creation technologies based on deep learning algorithms are universally acknowledged as a serious threat for person's reputation and digital identity. Recently, Generative Adversarial Networks (GANs) have attracted much attention being capable of generating realistic pictures that can be utilized in fraud schemes [1]. In the last decades, human's ability to distinguish between real and artificial intelligent (AI) generated faces has been investigated. Specifically, neuroscience research focused on how the use of synthetic stimuli affects people's capacity for face recognition tasks [2]. Multimedia forensic research investigated how much face-mixing operations (i.e., a face manipulation where two faces are mixed to create an hybrid one carrying traits of both original faces) are perceived by people. This is especially important for face authentication systems to prohibit unauthorized access to locations or services [3]. Farid et al. [4] reported that humans can still generally detect synthetic images correctly [5], [6]. Recently, studies on electroencephalographic (EEG) correlates investigated viewer's ability to distinguish familiar and unfamiliar people versus their face-swapped counterparts. Results showed that it is possible to discriminate fake videos from genuine ones when at least one face-swapped actor is known to the observer. [7]

Investigation of brain reactions to facial expressions is becoming a widespread research area, which aims at better understanding emotional processing and cognitive mechanisms. Even though traditional models suggest that facial identity and expression are processed in distinctive brain areas, the current findings highlight that emotion processing can have a strong influence on facial recognition and memory mechanisms [8]. Finally, other studies have shown that facial processing in adults is modulated by the emotional relevance of faces, especially those with expressions of fear [9]. However, current literature lacks in the analysis of EEG correlates derived from the combination of emotional and cognitive stimulation in the form of human faces. This paper investigated the electrical brain dynamics during cognitive and emotional mechanisms elicited by real and AI-generated (named ”synthetic”) stimuli representing human faces with positive, neutral, and negative expressions.

II. MATERIAL AND METHODS

A. Experimental Protocol

Healthy volunteers were subjected to visual stimuli representing human faces (both synthetic and real) expressing different types of emotion (positive, neutral, negative). The healthy group comprised 21 participants (10 males and 11 females), aged between 19 and 29 years (24.8±2.9). This study was approved by the Institutional Review Board and all participants gave written informed consent. Volunteers were set on a chair wearing an EEG helmet in front of a monitor where stimuli were presented. Experimental protocol was composed of two phases. The first phase was a baseline acquisition with 2 minutes of closed and open eyes each. In the second phase, subjects observed 3 sets of 20 images of faces, of which 10 were real and 10 synthetic. Each set contained faces associated exclusively with a polarized mood: positive (happy or smiling faces, Fig.1 a, b), neutral (relaxed faces, neutral expressions, Fig.1 c, d) or negative (sad, angry or discomforted faces, Fig.1 e, f). Both faces and sets were presented randomly for each subject, who pressed ”z” on a keyboard if the stimuli presented was considered synthetic or ”m” if real (or non synthetic). EEG acquisition was carried out using the DSI-24 helmet (Wearable Sensing, San Diego, CA, USA), with dry electrodes of the Ag/AgCl type. The helmet consists of 21 electrodes, arranged according to the International 10-20 Standard, and is wirelessly coupled with a triggering hub device to associate the neurophysiological recording at specific time intervals or tasks. EEG and trigger data are collected at a sample frequency of 300 Hz. Stimuli were composed of a set of caucasian faces (age range of 20-50 years) extracted from the CK+ face database [10]. Synthetic faces were generated through a generative-AI algorithm (i.e., FaceMix) [11], by mixing together 4 real images all expressing the same type of emotion in grayscale. Stimuli were presented only once, balanced in sex, type of emotional facial expression (positive, negative, neutral), and type of

[*] Corresponding author: pietro.tarchi@unifi.it

[1] Department of Information Engeneering, Università degli Studi di Firenze, Florence, Italy

[2] GenOMeC, Università degli Studi di Siena, Siena, Italy

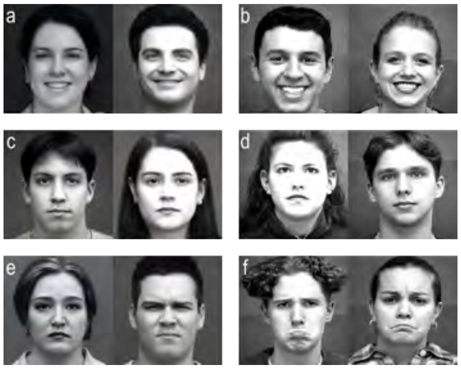

Fig. 1. Example of Synthetic (a, c, e) and Real (b, d, f) Faces Expressing Positive (a, b), Neutral (c, d) and Negative Emotions (e, f) used as Stimuli.

image, i.e., synthetic or real. The dataset comprised three classes, specifically, synthetic class, real class and emotional class, where the latter included both real and synthetic faces split for different emotional expression.

B. Signal Processing chain

EEG data were analyzed in MATLAB environment through EEGLAB [12] for continuous and event-related EEG processing. EEG signals were pre-processed following the Harvard Automated Processing Pipeline for Electroencephalography (HAPPE) [13], which is a standardized automated pipeline. Through HAPPE bandpass filtering, channel selection, electrical noise removal, bad channel rejection, wavelet-enhanced independent component analysis (W-ICA), independent component analysis (ICA), multiple artifact rejection algorithm (MARA) for independent component rejection, segmentation, interpolation, rejection, channel interpolation and re-referencing were performed.

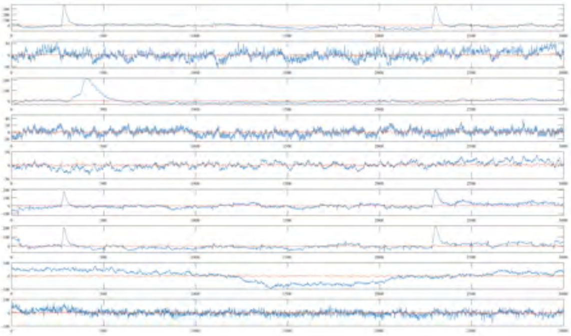

Fig. 2. Example of EEG-Channel Data Before (Blue) and After (Red) performing HAPPE.

After the pre-processing phase, EEG signals were split in epochs of 10 seconds identifying a precise stimulus. 1114 epochs were retained for further analysis (Table I). For each epoch, the average power spectrum in 5 bands of analysis (i.e., Delta 1-4 Hz, Theta 4-7 Hz, Alpha 8-12 Hz, Low-Beta 13-17 Hz, and High-Beta 18-32 Hz) and the average of Event Related Potentials (ERPs), in the 150-250 ms time interval (i.e., N200), were computed. The Power Spectrum Density (PSD) was computed with a 300 points Fast Fourier Transform (FFT) without overlap, while for ERP no baseline subtraction was performed due to the experimental design. Epochs were labelled as follows:

- tt (true-true) - in these epochs the image presented was real and subjects' answer was ”real”
- ff (false-false) - in these epochs the image presented was synthetic and subjects' answer was ”synthetic”
- tf (true-false) - in these epochs the image presented was real and subjects' answer was ”synthetic”
- ft (false-true) - in these epochs the image presented was synthetic and subjects' answer was ”real”

TABLE I

EPOCHS DIVISION ACCORDING TO LABELS.

Epochs-Label	tt	ff	tf	ft	total	percentage
positive	134	137	45	43	359	32,2%
neutral	156	147	36	42	381	34,2%
negative	144	138	47	45	374	33,6%
total	434	422	128	130	1114	100%
percentage	38,9%	37,9%	11,5&	11,7%	100%	

C. Statistical Analysis

Since the time-frequency EEG extracted features were not normally distributed, according to the Shapiro-Wilk test, surrogate tests were performed for statistical analysis [14]. Statistical comparison tests are:

- Error vs Correct (tf+ft vs tt+ff) - subjects guessed incorrectly vs correctly
- Ansfalse vs Anstrue (ff+tf vs tt+ft) - subjects answered ”synthetic” vs ”real”
- Imfalse vs Imtrue (ff+ft vs tt+ft) - presented faces belong to synthetic vs real class
- Positive vs Negative - presented faces expressed positive vs negative emotions
- Positive vs Neutral - presented faces expressed positive vs neutral emotions
- Neutral vs Negative - presented faces expressed neutral vs negative emotions

Bootstrap method was performed to estimate statistics by sampling our dataset with replacement. After performing bootstrap statistic, a paired t-test was carried out to verify whether the mean values of the parameters were statistically different at a significance level of 95% ($p<0.05$). In multiple comparisons a post-hoc Bonferroni correction was performed.

III. RESULTS

This section shows the results of statistical analysis through Scalp Topographic Maps (STMs). STMs describe the spatial distribution of extracted parameters, computed at the electrodes position, across the brain. To simplify visualization, we decided to use a false-colors map highlighting the statistical significant areas ($p<0.05$). Non-significant area is standardized with the green color. This map represents the p-value of the paired t-test in the comparison between the averages of the two different conditions under investigation. If an area of the STM assumes warm colors (yellow, orange, red), it means that the first term of the comparison is statistically greater than the second one; on the contrary, if the area assumes cold colors (cyan, light blue, blue) the second term is statistically greater than the first.

A. Subjects' Answers: Error vs Correct

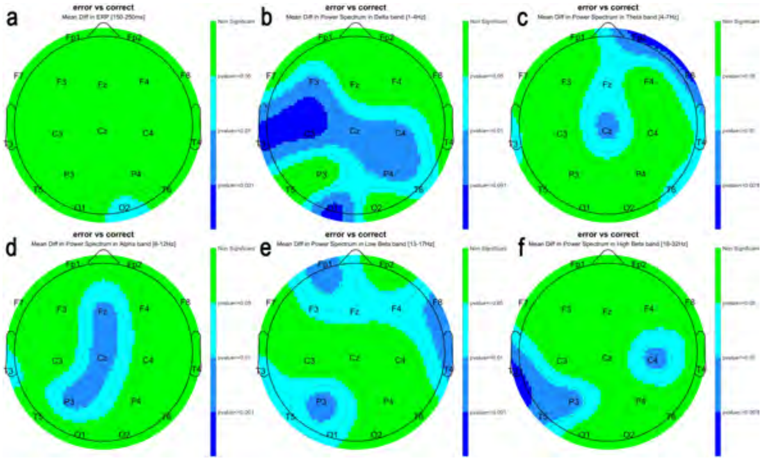
a
error vs correct
Mean Diff in ERP [150-250ms]
b
error vs correct
Mean Diff in Power Spectrum in Delta band [1-4Hz]
c
error vs correct
Mean Diff in Power Spectrum in Theta band [4-7Hz]
d
error vs correct
Mean Diff in Power Spectrum in Alpha band [8-12Hz]
e
error vs correct
Mean Diff in Power Spectrum in Low Beta band [13-17Hz]
f
error vs correct
Mean Diff in Power Spectrum in High Beta band [18-32Hz]
Non Significant
Fp1
Fp2
F7
F3
Fz
F4
F8
T3
C3
Cz
C4
T4
T5
P3
P4
T6
O1
O2

Fig. 3. False-colors STMs of statistically significant activations in subjects when guessing incorrectly (warm colors) or correctly (cold colors); a) in N200 a greater value is observed at the electrode O2 when subjects guessed correctly. b) in Delta band greater values are observed at the electrodes T3, C3, F3, Cz, C4, P4, T6 and O1 when subjects guessed correctly. c) in Theta band greater values are observed at the electrodes T3, Cz, Fz, Fp1, F8, T4 and T6 when subjects guessed correctly. d) in Alpha band greater values are observed at the electrodes T3, P3, Cz and Fz when subjects guessed correctly. e) in Low-Beta band greater values are observed at the electrodes Fp1, F3, Fz, F4, F8, T4, T3, T5, P3 and O1 when subjects guessed correctly. f) in High-Beta band greater values are observed at the electrodes T3, T5, P3 and C4 when subjects guessed correctly.

B. Subjects' Answers: Ansfalse vs Anstrue

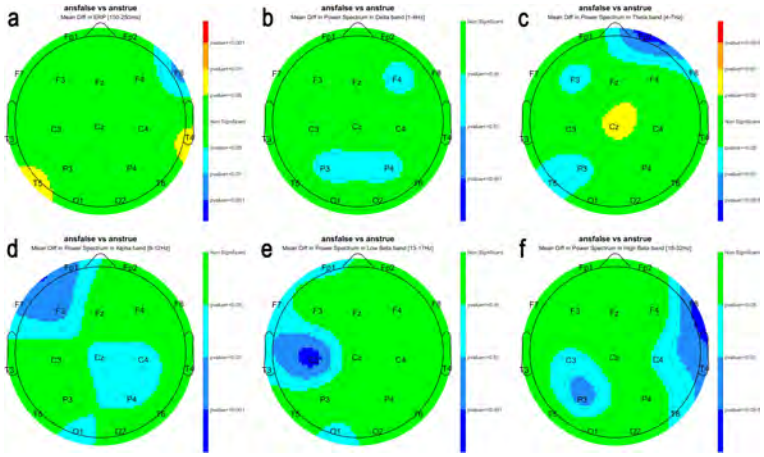
a
ansfalse vs anstrue
Mean Diff in ERP [150-250ms]
b
ansfalse vs anstrue
Mean Diff in Power Spectrum in Delta band [1-4Hz]
c
ansfalse vs anstrue
Mean Diff in Power Spectrum in Theta band [4-7Hz]
d
ansfalse vs anstrue
Mean Diff in Power Spectrum in Alpha band [8-12Hz]
e
ansfalse vs anstrue
Mean Diff in Power Spectrum in Low Beta band [13-17Hz]
f
ansfalse vs anstrue
Mean Diff in Power Spectrum in High Beta band [18-32Hz]
Non Significant

Fig. 4. False-colors STMs of statistically significant activations in subjects when answering "synthetic" (warm colors) or "real" (cold colors); a) in N200 greater values are observed at the electrodes T4 and T5 when subjects answered "synthetic"; electrode F8 shows a greater value when subjects answered "real". b) in Delta band greater values are observed at the electrodes F4, P3 and P4 when subjects answered "real". c) in Theta band a greater value is observed at the electrode Cz when the subjects answered "synthetic"; electrodes Fp2, F8, F3, T3 and P3 show greater values when subjects answered "real". d) in Alpha band greater values are observed at the electrodes Fp1, F3, F7, Cz, C4, P4 and O1 when subjects answered "real". e) in Low-Beta band greater values are observed at the electrodes Fp1, F7, C3, T3 and O1 when subjects answered "real". f) in High-Beta band greater values are observed at the electrodes C3, P3, F8, C4, T4 and T6 when subjects answered "real".

C. Image Class - Imfalse vs Imtrue

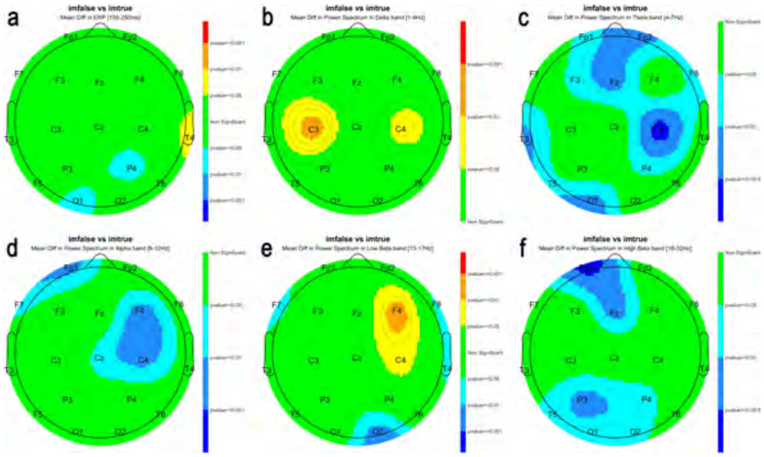
a
imfalse vs imtrue
Mean Diff in ERP [150-250ms]
b
imfalse vs imtrue
Mean Diff in Power Spectrum in Delta band [1-4Hz]
c
imfalse vs imtrue
Mean Diff in Power Spectrum in Theta band [4-7Hz]
d
imfalse vs imtrue
Mean Diff in Power Spectrum in Alpha band [8-12Hz]
e
imfalse vs imtrue
Mean Diff in Power Spectrum in Low Beta band [13-17Hz]
f
imfalse vs imtrue
Mean Diff in Power Spectrum in High Beta band [18-32Hz]
Non Significant

Fig. 5. False-colors STMs of statistically significant activations in subjects when the image presented was synthetic (warm colors) or real (cold colors); a) in N200 a greater value is observed at the electrode T4 when the image presented was synthetic; electrodes P4 and O1 show greater values when the image presented was real. b) in Delta band greater values are observed at the electrodes C3 and C4 when the image presented was synthetic. c) in Theta band greater values are observed at the electrodes Fp1, Fp2, F3, Fz, F8, C4, P4, T3, T5 and O1 when the image presented was real. d) in Alpha band greater values are observed at the electrodes Fp1, F7, Cz, C4 and F4 when the image presented was real. e) in Low-Beta band greater values are observed at the electrodes C4 and F4 when the image presented was synthetic; electrodes F7, F8, T4 and O2 show greater values when the image presented was real. f) in High-Beta band greater values are observed at the electrodes Fp1, Fz, T5, P3, P4, O1 and O2 when the image presented was real.

D. Image class - Emotions

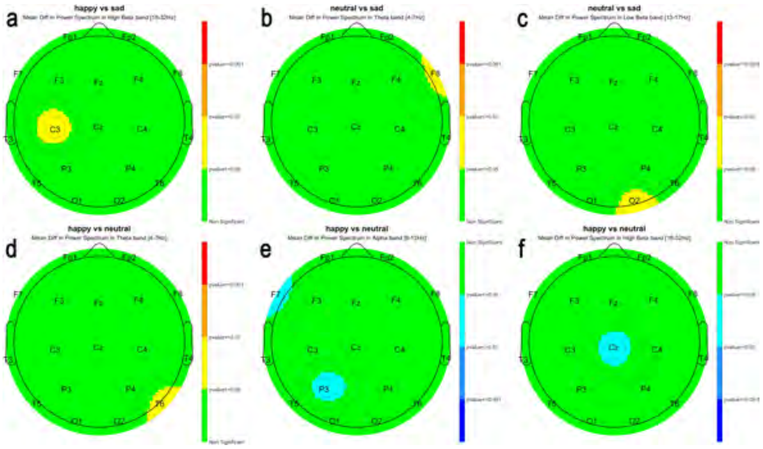
a
happy vs sad
Mean Diff in Power Spectrum in High Beta band [18-32Hz]
b
neutral vs sad
Mean Diff in Power Spectrum in Theta band [4-7Hz]
c
neutral vs sad
Mean Diff in Power Spectrum in Low Beta band [13-17Hz]
d
happy vs neutral
Mean Diff in Power Spectrum in Theta band [4-7Hz]
e
happy vs neutral
Mean Diff in Power Spectrum in Alpha band [8-12Hz]
f
happy vs neutral
Mean Diff in Power Spectrum in High Beta band [18-32Hz]
Fp1
Fp2
F7
F3
Fz
F4
F8
T3
C3
Cz
C4
T4
T5
P3
P4
T6
O1
O2
Non Significant

Fig. 6. False-colors STMs of statistically significant activations in subjects when the face presented was expressing positive (warm colors) or negative (cold colors) emotions (a), neutral (warm colors) or negative (cold colors) emotions (b & c), positve (warm colors) or neutral (cold colors) emotions (d, e & f)). a) in High-Beta band a greater value is observed at the electrode C3 when the face presented was expressing a positive emotion. b) in Theta band a greater value is observed at the electrode F8 when the face presented was expressing a neutral emotion. c) in Low-Beta band a greater value is observed at the electrode O2 when the face presented was expressing a neutral emotion. d) in Theta band a greater value is observed at the electrode T6 when the face presented was expressing a positive emotion. e) in Alpha band greater values are observed at the electrodes F7 and P3 when the face presented was expressing a neutral emotion. f) in High-Beta band a greater value is observed at the electrode Cz when the face presented was expressing a neutral emotion.

IV. DISCUSSIONS

Even though preliminary, statistical results highlighted that the face recognition process is a complex task involving the activation of the whole brain. Error vs Correct comparison, which is an unbalanced dataset, stated that subject correctly recognized 76,8% of the images while unrecognizing 23,2% of them (Table I). It is in agreement with the literature that reports on the good human's ability in recognition process of real vs synthetic stimuli. All STMs showed a greater response values toward the reals (Fig.3). In Ansfalse vs Anstrue comparison (Fig.4), we observed a prevalence of greater values towards the true answers in all 5 bands of interest, with the exception of the Cz in Theta band (Fig.4 c). These results can be due to a greater familiarity of the subject with the features of the real faces, making them more easily recognizable. Subjects, in fact, were slightly better at recognizing real faces, as shown in Table I, rather than the synthetic ones. Moreover, it is observed that the whole Beta band (Fig.4 e, f), whose activity is known to rises during tasks, showed greater values in the left and right temporal areas when the subjects responded ”real”. It may be due to the greater effort at the associative level of the subjects in familiarizing with those faces. It is also relevant that visual perception is mostly in the inferotemporal cortex (ITC): the receptive fields of its neurons include a large portion of the visual field both ipsilateral and contralateral. In a region of ITC, neurons respond particularly to faces, both in front and in profile [15]. Imfalse and Imtrue comparison showed an interesting difference between Low and High Beta bands (Fig.5 e, f). While the parieto-occipital areas responded almost similarly, a strong difference is found in the central and frontal areas. Even though they could be in contrast, a more deep investigation is needed before a conclusion. Generally, emotional comparisons showed greater values toward neutral stimuli with respect to the emotional ones (Fig.6 b, c, e, f). It suggests that since the task was more oriented in understanding synthetic vs real stimuli, participants found easier to recognize neutral vs emotions. We could hypothesize that there was a saturation effect during emotions interpretation (Table I), with a greater variance in positive and negative emotional response with respect to the neutral one. Future research should focus on clarify these contrasts by increasing the number of involved participants.

ACKNOWLEDGMENT

The research leading to these results has received partial funding from the Age-It - Challenges and Consequences of Aging — PNRR Italian Project.

REFERENCES

[1] Lago, F., Pasquini, C., Bohme, R., Dumont, H., Goffaux, V., & Boato, G. (2022). More real than real: A study on human visual perception of synthetic faces [applications corner]. IEEE Signal Process. Mag., 39(1), 109–116.

[2] Crookes, K., Ewing, L., Gildenhuys, J.-dith, Kloth, N., Hayward, W. G., Oxner, M., Pond, S., & Rhodes, G. (2015). How well do computer-generated faces tap face expertise? PLoS One, 10(11).

[3] Makrushin, A., Siegel, D., & Dittmann, J. (2020). Simulation of border control in an ongoing web-based experiment for estimating morphing detection performance of humans. IH&MMSec 2020.

[4] Farid, H., & Bravo, M. J. (2012). Perceptual discrimination of computer generated and photographic faces. Digit Investig, 8(3-4), 226–235.

[5] Holmes, O., Banks, M. S., & Farid, H. (2016). Assessing and improving the identification of computer-generated portraits. ACM TAP, 13(2), 1–12.

[6] Mader, B., Banks, M. S., & Farid, H. (2017). Identifying computer-generated portraits: The importance of training and Incentives. Perception, 46(9), 1062–1076.

[7] Tauscher, J.-P., Castillo, S., Bosse, S., & Magnor, M. (2021). EEG-based analysis of the impact of familiarity in the perception of Deepfake videos. IEEE ICIP 2021.

[8] D. Acunzo, G. MacKenzie, and M. C. van Rossum,”Spatial attention affects the early processing of neutral versus fearful faces when they are task-irrelevant: A classifier study of the EEG C1 component” CABN, vol. 19, no. 1, pp. 123-137, 2018.

[9] Leppänen, J. M., Moulson, M. C., Vogel-Farley, V. K., & Nelson, C. A. (2007). An ERP study of emotional face processing in the adult and Infant Brain. Child Dev., 78(1), 232–245.

[10] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Workshops.

[11] ”Unlimited free face mix AI generator - playform - AI art generative platform for artists and creative people. free, unlimited, easy.” Playform. [Online]. Available: https://playform.io/facemix. [Accessed: 01-Feb-2023].

[12] A. Delorme and S. Makeig,”EEGLAB: An open source toolbox for analysis of single-trial EEG Dynamics including independent component analysis,” J. Neurosci. Methods, vol. 134, no. 1, pp. 9-21, 2004.

[13] L. J. Gabard-Durnam, A. S. Mendez Leal, C. L. Wilkinson, and A. R. Levin,”The Harvard Automated Processing Pipeline for Electroencephalography (Happe): Standardized processing software for developmental and high-artifact data” Front. Neurosci., vol. 12, 2018.

[14] L. Faes, A. Porta, and G. Nollo, ”Surrogate data approaches to assess the significance of directed coherence: Application to EEG activity propagation,” Conf Proc IEEE Eng Med Biol Soc, 2009.

[15] Gross, C. G., Bender, D. B., & Rocha-Miranda, C. E. (1969). Visual receptive fields of neurons in inferotemporal cortex of the monkey. Science, 166(3910), 1303–1306.