A method for classification NEO–FFI answers fabricated and advantageous due to psychological bias using brainwave specific brain activity networks

Yuto Ashikawa*, Takashi Ito, Shohei Ishizu and Yosuke Kurihara
* University of Aoyama Gakuin, Kanagawa, Japan
E-mail: c5624226@aoyama.jp Tel: +81-80-5899-9702
University of Aoyama Gakuin, Kanagawa, Japan
E-mail: ito@ise.aoyama.ac.jp Tel: +81-42-759-6423
University of Aoyama Gakuin, Kanagawa, Japan
E-mail: ishizu@ise.aoyama.ac.jp Tel: +81-44-953-1525
University of Aoyama Gakuin, Kanagawa, Japan
E-mail: kurihara@ise.aoyama.ac.jp Tel: +81-42-759-6371

***Abstract*— The Neuroticism Extraversion Openness–Five Factor Inventory(NEO–FFI) is a psychological scale that assesses personality through five factors. Although widely used, it is susceptible to biases, which reduces reliability. This study proposes a method to classify biased answers using brainwave measured while answering NEO–FFI questions. Brainwave networks are constructed from alpha, beta, delta, theta, and gamma waves measured at eight brain regions (frontal, temporal, parietal, and occipital lobes). Fast Fourier transform and coherence are applied to obtain feature values for nodes and edges of the network. Support vector machines are then trained to classify answers as biased or unbiased based on these features for each personality factor. An experiment with 23 participants aged 19–23 reveals an accuracy of 0.67, precision of 0.72, negative predictive value of 0.62, recall of 0.71, specificity of 0.64, and F1 score of 0.70. This indicates that the proposed method effectively classifies biased answers and improves the reliability of NEO–FFI personality assessments.**

I. INTRODUCTION

The Neuroticism Extraversion Openness‐Five Factor Inventory(NEO–FFI) is a psychological scale developed by P.T. Costa and colleagues based on L.R. Goldberg's Big Five theory[1][2]. The theory measures personality in terms of five factors (N: neuroticism, E: extraversion, O: openness, A: agreeableness, and C: conscientiousness). Each factor is evaluated by 12 questions. The answers are rated from zero to four. The total score for each factor is used to assess personality. The NEO–FFI is widely used in psychological research to quantitatively evaluate personalities [3–7].

However, it has been demonstrated that respondents can fabricate their answers owing to psychological biases to present themselves favorably[8][9]. The inclusion of such fabricated answers may compromise the reliability of personality assessments. Therefore, classify fabricated answers in NEO–FFI items can enhance the interpretative reliability of personality evaluations.

Psychological biases are induced by emotional stimuli[10][11], which activate the brain activity[12–14]. Studies have revealed that the brain blood flow can predict the scores for personality questionnaires such as the state‐trait anxiety inventory[15][16]. Additionally, studies have analyzed brain network characteristics in answer to emotional stimuli[17].

Based on these observations, this study proposes a method to classify fabricated answers in the NEO–FFI by constructing brain activity networks using brainwave data from eight brain regions for alpha, beta, theta, delta, and gamma waves and applying an SVM trained on personality factors.

II. PROPOSED METHOD

Fig. 1 shows a schematic of the signal processing conducted in the proposed method.

If the set of personality factors targeted in the NEO–FFI is $F \in \{F_{\mathrm{N}}, F_{\mathrm{E}}, F_{\mathrm{O}}, F_{\mathrm{A}}, F_{\mathrm{C}}\}$, each of the 60 questions in the questionnaire belongs to one of the factors $i \in F$. If an arbitrary question is $q = (1, 2, \dots, 60)$, the brainwaves of the respondent while answering q are measured at brain region p. Let brain region p be a set of brain regions $p \in \{\mathrm{Fpz}, \mathrm{Fz}, \mathrm{Cz}, \mathrm{Pz}, \mathrm{T3}, \mathrm{T4}, \mathrm{O1}, \mathrm{O2}\}$ based on the international 10–20 method. Let $x_n(k)$ be the signal without hum noise obtained by applying a notch filter to brainwaves measured at arbitrary $n \in p$. $k(= 1, 2, 3, \dots, N_d)$ is discretized with the sampling interval Δt. Here, N_d is the number of data points. However, because the answer times differ for each question, N_d also differs.

Based on the measured $x_n(k)$ at each region, a brain activation network is constructed with region p as a node for each of the five types of brainwaves $B \in \{\alpha\ \text{wave}, \beta\ \text{wave}, \theta\ \text{wave}, \delta\ \text{wave}, \gamma\ \text{wave}\}$. If the frequency

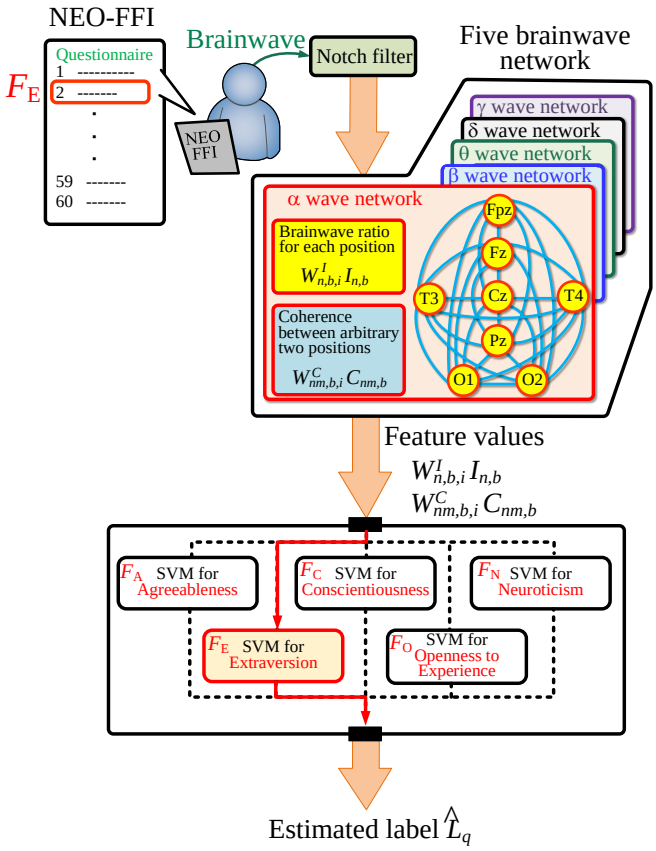

Fig.1 The signal processing flow to classify the answer is affected by psychological bias.

band of each brainwave is f_{low}– f_{high}, the frequency band of the brainwaves are as follows: α wave: 8–13 Hz, β wave: 14–29 Hz, θ wave: 4–7 Hz, δ wave: 0.5–3 Hz, and γ wave: 30 –70 Hz. A brain activation network for any brainwave $b \in B$ is defined by the value $I_{n,b}$ at a node and the value $C_{nm,b}$ of an edge between any two nodes n, $m \in p(m \neq n)$. The $I_{n,b}$ of the node at region n is calculated using Equation (1) for the amplitude spectrum $S_n(f)$ obtained by applying a fast Fourier transform to $x_n(k)$. Here, $f(=0, \Delta f, 2\Delta f, \ldots, f_N)$ is a discrete frequency, Δf is the frequency resolution, and f_N is the Nyquist frequency.

$$I_{n,b} = \sum_{f=\tilde{f}_{\text{low}}}^{\tilde{f}_{\text{high}}} S_n(f) \Big/ \Big(\sum_{f=0}^{F_N} S_n(f)\Big) \qquad (1)$$

Because f is a discrete frequency with frequency resolution Δf, the amplitude spectra corresponding to f_{low} and f_{high} in each brainwave is not necessarily obtained. Therefore, we let $\tilde{f}_{\text{low}}$ and $\tilde{f}_{\text{high}}$ be the discrete frequencies f closest to f_{low} and f_{high}, respectively, and calculate the amplitude spectrum. In addition, the edge value $C_{nm,b}$ is the average of the coherence values $C(f_{\text{low}}: f_{\text{high}})$ in the frequency– band f_{low}– f_{high} at the arbitrary region n and m. These are determined by Equation (2).

$$C(f_{\text{low}}: f_{\text{high}}) = \left|S_n(f_{\text{low}}: f_{\text{high}}) S_m(f_{\text{low}}: f_{\text{high}})\right|^2 \Big/ \left(\left|S_n(f_{\text{low}}: f_{\text{high}})\right|^2 \left|S_m(f_{\text{low}}: f_{\text{high}})\right|^2\right) \qquad (2)$$

In this proposed method, the degree of influence of factor i for $I_{n,b}$, $C_{nm,b}$ are $w^I_{n,b,i}$, $w^C_{nm,b,i} (= 0-1)$. And then, $w^I_{n,b,i} I_{n,b}$ and $w^C_{nm,b,i} C_{nm,b}$ are the feature value for the brain activation network in brainwave b.

Using an SVM, we classify whether the answer to question q was artificial from the features obtained from the network for all the brainwaves. In the learning phase of the SVM, N_s names are asked to answer the NEO–FFI. The questions answered honestly are labeled "0", and those answered intentionally are labeled "1". Here, the label assigned to the question is $L_q \in \{0, 1\}$. In the proposed method, for each personality factor i, the features are $w^I_{n,b,i} I_{n,b}$, $w^C_{nm,b,i} C_{nm,b}$. Moreover, the label constructs a learning model to determine L_q. Let M_i denote the SVM learning model for each factor i. In the classification phase of the SVM, the feature values $w^I_{n,b,i} I_{n,b}$, $w^C_{nm,b,i} C_{nm,b}$ and inputs it into the learning model M_i trained for personality factor i to determine the estimated classify label $\hat{L}_q$ for question q.

III. Verification Experiment

A. *Experimental Procedures to Obtain Dataset*

To validate the proposed method, a verification experiment was conducted with 23 participants aged 19–23 years.

The participants were asked to answer an NEO–FFI questionnaire containing 60 randomly arranged questions. The questions were displayed on a PC screen. The participants answered using a wireless keyboard and scored each question from zero to four. During this process, the participants were instructed to intentionally provide deceptive answers to questions wherein they wished to present themselves more favorably. This set of answers was labeled (A1).

To verify the proposed method, it was necessary to determine whether each question q in (A1) was answered genuinely or deceptively. Therefore, in this experiment, each participant was asked to answer another set of NEO–FFI questions in a manner similar to that for (A1), but honestly. These honest answers were labeled (A0).

For each question q, the difference in scores between (A0) and (A1) was used to determine the label L_q. For the factors F_{E}, F_{O}, F_{A}, and F_{C}, higher scores are generally considered more favorable. Meanwhile, for F_{N}, lower scores were considered more favorable [9]. Therefore, the absolute difference in scores between (A0) and (A1) was calculated. If this difference was equal to or larger than a threshold s, the answers to that question in (A1) was labeled deceptive ($L_q = 1$). If the difference was less than s, it was labeled as genuine ($L_q = 0$). The order of the (A0) and (A1) responses was randomized for

each participant. Additionally, during the (A1) answers, brainwave data from head position p was recorded using the polymate(AP1532) brainwave system from Miyuki Giken to extract feature values. This study was approved by the Ethics Review Committee of the Aoyama Gakuin University (Approval Number H23–025).

B. Hyper Parameters for SVM

In this validation experiment, the hyperparameters used for the SVM were of six types: the kernel function; degree d in the polynomial kernel; and parameters BC (adjusts the penalty for classification errors), G (adjusts the complexity of the decision boundary), C_0 (determines the importance of the negative class), and C_1 (determines the importance of the positive class).

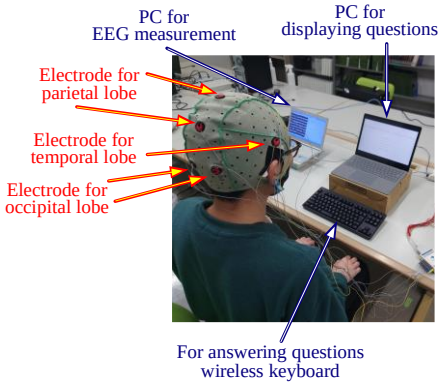

Fig.2 Experimental system.

C. Evaluation Method and Parameter Tuning by Genetic Algorithm (GA)

To evaluate the accuracy of the SVM–based classification, we performed leave–one–subject–out cross–validation using data from 22 of the 23 participants ($N_s = 22$) for training and the remaining participant's data for testing. We randomly selected data to balance the number of L_q labels (zero and one) five times in the training set and generated a confusion matrix for each test participant. This yielded 115 confusion matrices (23 participants × 5 times). We calculated the accuracy, Positive Predictive Value(PPV), Negative Predictive Value(NPV), sensitivity, specificity, and F1–score using Equations (3)–(8) and averaged these over 115 matrices for validation.

For the overall evaluation, a confusion matrix was constructed from the results of each question for each participant averaged over 115 repetitions. We optimized the weights $w^I_{n,b,i}$, $w^C_{nm,b,i}$; threshold s; kernel type (linear, RBF, polynomial); and parameters BC, G, d, C_0, C_1 using a GA. The range for $w^I_{n,b,i}$ and $w^C_{nm,b,i}$ was 0–1; that for BC, G, C_0, and C_1 was 0–3000; and that for d was 2–4. The GA fitness function is the product of the six performance metrics from Equations (3)–(8) with 10,000 generations and a number of individuals size of 40.

Table 1 Confusion matrix.

		Actual	
		Negative	Positive
Classification	Negative	TN	FN
	Positive	FP	TP

$$\text{Accuracy} = (TP + TN) / (TP + FP + FN + TN) \quad (3)$$

$$\text{PPV} = TP / (TP + FP) \quad (4)$$

$$\text{NPV} = TN / (TN + FN) \quad (5)$$

$$\text{recall} = TP / (TP + FN) \quad (6)$$

$$\text{specificity} = TN / (TN + FP) \quad (7)$$

$$\text{F1–score} = 2 \times \text{PPV} \times \text{recall} / (\text{PPV} + \text{recall}) \quad (8)$$

IV. Results

Table 1 presents the results of the verification experiments. It is evident from the table that for the factor N, the highest and lowest accuracy six performance indices were the PPV and NPV, respectively. This indicates that for this factor, the method cannot classify questions unaffected by psychological bias. Similarly, for the factors E, A, and C, the highest and lowest accuracy six performance indices were positive and negative recall, respectively. This indicates that the method cannot classify questions unaffected by psychological bias (similar to the factor N). Meanwhile, for the factor O, the highest and lowest accuracy indices were specificity and recall, respectively. This indicates that the method could not classify questions affected by psychological bias.

Overall, the highest and lowest accuracy indices were the PPV and NPV, respectively. This indicates that the method could not classify questions unaffected by psychological bias. The highest accuracy for each performance index was as follows: O, accuracy; N, PPV; O, NPV; A, recall; O, specificity; and N, F1–score. This indicates that O showed the highest accuracy for questions unaffected by psychological bias. Conversely, N and C showed high accuracies for questions affected by psychological bias, and A and E showed a balanced classification (see Table 2).

Table 2 Classification results by SVM.

	NEO–FFI measurement factors					
	N	E	O	A	C	whole
Accuracy	0.67	0.67	0.71	0.68	0.64	0.67
PPV	0.83	0.68	0.66	0.74	0.72	0.72
NPV	0.45	0.62	0.71	0.65	0.52	0.62
Recall	0.73	0.75	0.59	0.76	0.75	0.71
Specificity	0.55	0.61	0.78	0.61	0.46	0.64
F1–score	0.73	0.66	0.56	0.69	0.68	0.70

V. DISCUSSION

In the proposed method, we considered the causes of misclassification. For O, 15.1% of the questions in conditions (A0) and (A1) had an identical upper limit of four scores. Because the answers are on a five–score scale, those affected by psychological bias cannot select more than four scores. Thus, these fail to reflect the bias. This may cause misclassification.

Moreover, the balanced classification of A and E can be attributed to respondents attempting to appear favorable by considering their relationships with others (such as harmony and communication), which yields balanced results for these factors. In contrast, other factors focused more on self–perception, which may have resulted in a biased classification.

Additionally, the influence of psychological bias does not manifest immediately after starting the questions but appears after a certain time. Therefore, it is necessary to construct a brain activity network that considers the time–dependent effects. The current network does not account for these temporal variations.

VI. CONCLUSION

In this study, we proposed a method for classify fabricated answers caused by psychological bias in NEO–FFI answers. This method uses brain activity networks characterized by five types of brainwave signals for each personality factor. The validation experiments with 23 participants revealed the feasibility of classify the fabricated answers with an accuracy of 0.67, PPV of 0.72, NPV of 0.62, recall of 0.71, specificity of 0.64, and F1–score of 0.70.

Brain activity networks were constructed using the brainwave spectral content and average of coherence values from different brain regions to classify fabricated answers owing to psychological bias. Future work should consider the time required for participants to be influenced by psychological biases and incorporate response time into the features for a more accurate classification method.

REFERENCES

[1] D. Peabody and L. R. Goldberg, “Some determinants of factor structures from personality–trait descriptors,” *Journal of Personality and Social Psychology*, Vol. 57, No. 3, p.552, 1989.

[2] P. T. Costa and R. R. McCrae, “Neo personality inventory–revised (NEO PI–R),” *Psychological Assessment Resources*, 1992.

[3] R. Riemann, A. Angleitner, and J. Strelau, “Genetic and environmental influences on personality: A study of twins reared together using the self‐and peer report NEO‐FFI scales,” *Journal of Personality*, Vol. 65, No. 3, pp.449-475, 1997.

[4] N. Archer, R. G. Brown, H. Boothby, C. Foy, H. Nicholas, and S. Lovestone, “The NEO–FFI is a reliable measure of premorbid personality in patients with probable Alzheimer's disease,” *International Journal of Geriatric Psychiatry: A Journal of the Psychiatry of Late Life and Allied Sciences*, Vol. 21, No. 5, pp.477-484, 2006.

[5] M. Z. Podolska, M. Bidzan, M. Majkowicz, J. Podolski, O. S. Szmigiel, and E. R. Walknowska, “Personality traits assessed by the NEO Five–Factor Inventory (NEO–FFI) as part of the perinatal depression screening program,” *Medical Science Monitor*, Vol. 16, No. 9, pp.77-81, 2010.

[6] K. Lee and M. C. Ashton, “Prediction of self–and observer report scores on HEXACO–60 and NEO–FFI scales,” *Journal of Research in Personality*, Vol. 47, No. 5, pp.668-675, 2013.

[7] S. Sharma, “The impact of neuroticism and conscientiousness on sleep quality: exploring personality traits and sleep disturbances in young adults,” *International Journal of Interdisciplinary Approaches in Psychology*, Vol. 2, No. 5, pp.1515-1536, 2024.

[8] R. R. Holden and C. E. Lambert, “Response latencies are alive and well for identifying fakers on a self–report personality inventory: A reconsideration of van Hooft and Born 2012,” *Behavior Research Methods*, Vol. 47, pp.1436-1442, 2015.

[9] S. Horio and K. Takahashi, “Roles of social desirability scales of the Big–Five Personality Inventory in faking settings,” *Japanese Association of Industrial/Organization Psychology Journal*, Vol. 17, pp.65–77, 2004.

[10] G. Ceschi, S. Meylan, C. Rowe, and A. H. Boudoukha, “Psychological profile, emotion regulation, and aggression in police applicants: A Swiss cross-sectional study,” *Journal of Police and Criminal Psychology*, Vol. 37, pp.962–971, 2022.

[11] I. Krumpal, “Social desirability bias and context in sensitive surveys,” *Encyclopedia of quality of life and well–being research*, Cham: Springer International Publishing, pp.6527–6532, 2024.

[12] L. Xia, et al. "The brain activation of anxiety disorders during emotional stimulations: A coordinate–based activation likelihood estimation meta–analysis," 2021.

[13] K. Juliane, et al. "The effect of emotional content on brain activation and the late positive potential in a word n–back task." *PloS one*, Vol. 8, No. 9, 2013: e75598.

[14] N. A. Puccetti, et al. “Linking amygdala persistence to real–world emotional experience and psychological well–being,” *Journal of Neuroscience 21*, Vol. 41, No. 16, pp.3721-3730, April 2021.

[15] K. Hasegawa, Y. Hamada, and Y. Kurihara, “Using granger causal test associated with trait anxiety to estimate STAI scores by constructing a brain activity independence network,” The 11th IIAE International Conference on Industrial Application Engineering 2023, 2023.

[16] K. Hasegawa, Y. Hamada, and Y. Kurihara, “Verifying the effects of ceberal blood flow pulse component and dependance probability of brain activity in the prefrontal cortex on empathy estimation accuracy,” The SICE Annual Conference 2022, 2022.

[17] A. Muramatsu, S. Kobayashi, and Y. Mizuno–Matsumoto, “Complex network analysis of electroencephalography elicited by emotional stimuli presented on the smartphone,” *Journal of Affective Engineering*, Vol. 18, No. 4, pp.263-271, 2019.