Missing Values Imputation using Similarity Matching Method for Brainprint Authentication

Siaw-Hong Liew[1], Yun-Huoy Choo[2*], Yin Fen Low[3]
[1,2] Computational Intelligence and Technologies (CIT) Research Group
Faculty of Information and Communication Technology
[3] Machine Learning and Signal Processing (MLSP) Research Group
Faculty of Electronics and Computer Engineering
Universiti Teknikal Malaysia Melaka (UTeM), 76100 Durian Tunggal, Melaka, Malaysia.

***Abstract*—This paper proposes a similarity matching imputation method to deal with the missing values in electroencephalogram (EEG) signals. EEG signals with rather high amplitude can be considered as noise, normally they will be removed. The occurrence of missing values after this artefact rejection process increases the complexity of computational modelling due to incomplete data input for model training. The fundamental concept of the proposed similarity matching imputation method is founded on the assumption that similar stimulation on a particular subject will acquire comparable EEG signals response over the related EEG channels. Hence, we replaced the missing values using the highest similarity amplitude measure across different trials in this study. Next, wavelet phase stability (WPS) was used to evaluate the performance of the proposed method since WPS portrays better signals information as compared to amplitude measure in this situation. The statistical paired sample t-test was used to validate the performance of the proposed similarity matching imputation method and the preceding mean substitute imputation method. The lower the value of mean difference indicates the better approximation of imputation data towards its original form. The proposed method is able to treat 9.75% more missing value trials, with significantly better imputation value, than the mean substitution method. Continuity of the current study will be focusing on evaluating the robustness of the proposed method in dealing with different rate of missing data.**

***Keywords*—Similarity matching; data imputation; wavelet phase stability; missing values; artefact rejection**

I. Introduction

Brainprint authentication is catching attention recently because of their high time resolution, portability and relatively low cost [1]. Many decent non-clinical grade Electroencephalogram (EEG) acquisition devices have been introduced to the consumer market. This has greatly helped in promoting the EEG research since the data acquisition process is getting simpler and affordable. The consumer grade EEG devices are capable of providing better portability with reduced calibration time [2]. Brainprint authentication is an authentication method that using EEG signals. EEG is a popular non-invasive method which record the electrical activities of the brain on the scalp. It is normally measure in small voltage fluctuations within the brain. Human brain plays important role in controlling the coordination of nerves and muscles.

The advantage of using EEG signals as biometric modality lies on its uniqueness and confidentiality. Every individual has different brain responses towards different stimuli. Thus, the EEG is expected to have high inter-subject variability and low intra-subject variability. A good biometric modality should also have this property. EEG is outstanding than the current biometric modalities because EEG signals are hidden in our brain and non-observable physically. Other biometric modalities, such as fingerprint or face, are easily obtainable physical sensors from the body surface [3]. Besides, these biometric modalities are lack of the function of liveness detection. Nevertheless, EEG signals can be easily influenced by artefact noises. The large amplitude fluctuations in the EEG signals will be occurred when the subjects having eye blinking, body movements and etc. Therefore, pre-processing steps such as filtration and artefact rejection are necessary to improve the EEG signals quality.

The main purpose of the artefact rejection is to exclude the EEG signals with amplitude greater than 100 μV. Normal amplitude for EEG signals will not exceed 100 μV unless the amplitude come from the artefacts like body movements or eye blinking [4]. Thus, the artefact rejection will lead to missing trials for that particular channel. In order to tackle this issue, a similarity matching imputation method is proposed to deal with the missing values caused by artefact rejection.

The rest of this paper is structured as follows: Section II reviews the related works about the missing values imputation and wavelet phase stability. Section III describes the proposed similarity matching imputation method. Section IV illustrates the experimentation, which includes the data acquisition, experimental setup, data pre-processing, data preparation, wavelet phase stability (WPS) and statistical test. Section V portrays the experimental results and discussion for the proposed similarity matching imputation method and mean substitute imputation method. Section VI draws the conclusion and suggests the direction of future work.

II. Related Works

In the real-life applications, it is not easy to obtain a perfect dataset especially in signal analysis. EEG signals are having the low signal-to-noise ratio. Therefore, pre-processing is compulsory to remove the noise from the signals. Missing values will appear after the pre-processing steps.

One of the easiest ways to deal with the missing values is by ignoring the missing values in the dataset. However, it is very risky if there were large amount of missing values found in the dataset [5]. Another two important issues of large amount of missing values are leading to loss of meaningful information and distorting the result analysis. Consequently, several imputation methods have been used to deal with missing values.

From the past research, mean or mode substitute imputation methods are the most commonly used because there are simple and straightforward methods [6]. Unfortunately, the mean substitute imputation method can severely distort the distribution of the data. In EEG signals analysis, mean or mode substitution might not be appropriate due to the fluctuation of amplitude because it may lead to higher degree of standard deviation. The data structure is hardly maintained due to the high variation in the replacement values [7]. Another alternative way to deal with the missing values for EEG signals was using incremental approach proposed by Kim et al. [8], which is the incremental expectation maximization principal component analysis (iEMPCA). The estimated missing values were close to original data. However, the implementation of the mean substitution method is simpler than the iEMPCA. The EMPCA starts with initializing the mean value. Then, the data is reconstructed by using the number of predefined principal components. The processes will be repeated until convergence. Besides, the incremental approach is applied in the PCA to update the weight vector incrementally for the number of hidden variables. It is to minimize the average reconstruction error. Therefore, the expectation maximization was used to estimate the missing values.

Maximum likelihood estimation method is a famous statistical method, which finds the parameter to maximize the likelihood function. Sieluzycki and Kordowski [9] proposed an maximum likelihood estimation to improve the quality of the imputation method for auditory evoked brain responses. The proposed maximum likelihood estimation method takes into account the trial-to-trial variability on the multichannel level. The proposed algorithm was proven in reconstructing the lateralization of the trial-to-trial variability for the brain evoked responses. Yet, the algorithm needs to further improve when dealing with the nonstationary of the signals that arise from the stochastic noise.

Event-related potentials (ERPs) are commonly used in EEG signals analysis for brainprint authentication [10]–[14]. ERP is the averaging of many trials in order to enhance the signal-to-noise ratio. Averaging across the repeated responses is very useful when we are interested to look at the evoked potentials. However, when we look into the amplitude information of single trial, it tends to be fragile [15]. Other than that, the averaging approach will cause the loss of the information on the response variability across single trials. It is due to the large amplitude fluctuations can be easily produced even though there are slightly changes in measurement setup or body movements. Therefore, we can conclude that the EEG signals having large variance between one trial to another. Thus, phase information is emphasized to illustrate the similarity between the signals. Time domain analysis shows the EEG signals changes over time. On the other hand, frequency domain analysis shows the energy distributed over a range of frequencies and also includes information on the phase shift that applied to each frequency component [16]. Oppenheim and Lim [17] had stressed on the importance of phase in the EEG signals by using the Fourier representation. Other than that, the usefulness of the phase information is also interpreted in signal and image reconstruction [18]. The justification was presented from a statistical viewpoint. Wavelet phase stability (WPS) is proposed in [19] to address the issue of the ERPs. WPS makes use of the wavelet-based measure that gives the phase information. In the view of signal processing, the phase of a signal encompasses more significant information as compared to the amplitude of a signal.

III. The Proposed Similarity Matching Imputation Method

We proposed a similarity matching imputation method to treat the missing values in this case study. Our main idea for the missing values imputation is based on the similarity between two trials. The main concept of the similarity matching imputation method is shown in Fig. 1. The formula of similarity is calculated as follows:

$$similarity = 1 - \frac{|a(x)-a(y)|}{|a_{max}-a_{min}|} \quad (1)$$

Where, $a(x)$ and $a(y)$ are the object x and y which similar to attribute a; a_{max} and a_{min} are the maximal and minimal occurring value of attribute a.

The trials will be excluded if the rate of missing values is greater than 20% of the total values in a trial. In other words, we have used 21 electrodes in total, so the trials with the missing values from 5 electrodes and above will be excluded. It is because large number of missing values can lead to information loss and degraded the meaningful information. Thus, the incomplete trials must be excluded instead of performing missing values imputation. On contrary, the incomplete trials with missing values less than 20% will be treated iteratively.

The proposed similarity matching imputation method will search for the most similar complete trial as the reference point to treat the particular incomplete trial. The similarity measure is calculated by comparing the data without missing value between the complete trial and the treatment trial. Only the EEG signals of the same subject, in the same data acquisition condition (i.e. quiet, low distraction or high distraction), and of the same type of stimulus (i.e. password or non-password) are included in the searching pool. Once the highest similarity measure is found, the similarity matching imputation method will replace the missing values of the particular electrodes from the complete trial.

IV. Experimentation

In this section, we illustrate the EEG data acquisition process, experimental setup, data pre-processing steps, data preparation and implementation. Besides that, we evaluate the performance of the imputation methods based on the amplitude information and the phase information.

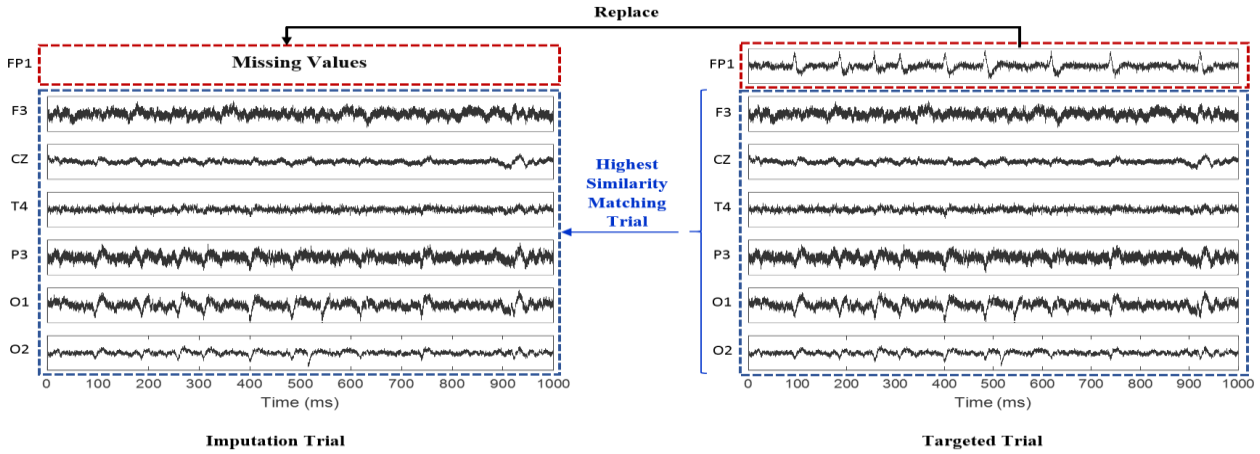

Fig. 1. Concept of Similarity Matching Imputation Method on Targetted Trial with Multiple Electrodes in a Single Subject EEG Data.

A. Data Acquisition and Experimental Setup

A new raw EEG dataset is collected from 4 healthy subjects (with mean age 29.8 years old) in Universiti Teknikal Malaysia Melaka (UTeM). All the subjects had normal vision or corrected normal vision. Beforehand, the investigator clarified the experimental procedures and the subjects were given the informed consent in prior to their participation. An ethical approval had been approved by Medical Research and Ethics Committee (MREC) from Ministry of Health Malaysia. The subject was seated on a back-rested chair to get the maximum comfort during the experiment. The computer display was located 1 meter away from the subject's eyes. EEG data acquisition began by attaching the electrodes onto the subject's scalp. The black and white pictures will display one-by-one and the subjects were asked to recognize the password picture. The subjects were required to click the mouse immediately upon the password picture is displayed. A total of 150 trials were recorded for each session, which are 60 trials of selected password picture and 90 trials with random picture from 260 pictures excluded the password picture. The 150 trials were displayed randomly to the subject. The main purpose for this is to increase the signal-to-noise ratio by averaging the total trials. The Inter-Stimulus Interval (ISI) for each trial was set to 1.5 seconds and the picture was remained on the computer screen for 1 second followed by 1.5 seconds of white blank screen.

EEG signals recording was carried out in three different environment conditions: (a) a quiet condition; (b) a low distraction condition; and (c) a high distraction condition. It is to simulate the real-world environment. For the low distraction condition, an audio clip with consistent office noise effects was played through an audio speaker and the sound level is approximately 55 decibel (dB). On the other hand, for the high distraction condition, an audio clip with inconsistent office noise effects such as noise from phone ringing, printer printing, and etc. was also played through the audio speaker with the sound level approximately 70 dB. The electrodes used to record the EEG signals were 21 electrodes by using Twente Medical Systems International (TMSi) Porti system with sampling frequency 512 Hz. The electrodes are FP1, FPZ, FP2, F7, F3, FZ, F4, F8, T3, C3, CZ, C4, T4, T5, P3, PZ, P4, T6, O1, OZ and O2. All the scalp electrodes were referred to right earlobe and grounded on right hand in the experiment.

B. Data Pre-Processing and Data Preparation

The raw EEG data are noisy, complex and highly uncertain. Thus, the pre-processing steps must be performed in prior to further analysis. The 3 basic steps for stimulus-locked EEG data are filtering, segmentation and artefact rejection. Filtering plays important role in minimizing the background noise and interference and improving the EEG signals quality. The EEG data obtained was bandpass filtered with a Finite-duration Impulse Response (FIR) filter with the cut-off frequencies of 1 – 30 Hz. In addition, the artefact rejection was used to remove the EEG signals responses with excessive body movements or other types of artefacts with amplitude greater than 100 µV. Thus, the trials with amplitude greater than 100 µV were discarded and hence cause the missing values of the particular channels.To evaluate the performance of the proposed similarity matching method, we should know the real values of the missing values. In this experiment, we generated 20% of the missing values in the original observed data to verify the efficiency of missing values imputation.

C. Wavelet Phase Stability (WPS) [19]

Wavelet phase stability (WPS) employs the wavelet-based measure that gives the phase information. It can prove that a reconstructed signal will not suffering from a degradation of the quality. WPS is used to analyze the synchronization process that is locked to the onset of the stimuli. The moving mean of WPS is defined as follows:

$$\Gamma_{s,\tau}^{m}(\mathcal{F}) = \frac{1}{m}\left|\sum_{n=1}^{m} e^{larg\left((\mathcal{W}_{\psi fm})(s,\tau)\right)}\right| \quad (2)$$

Where, $m = 1, \dots, M$ and $\Gamma_{s,\tau}^{m}(\mathcal{F})$ measures the mean of the degree of clustering of the angular distribution for certain s and τ for M trials. The value of WPS ranges from 0 to 1; where 1 indicates the perfect phase stability. The smaller the value of WPS, the poorer the phase stability. In this study, we calculated the WPS for the original data, the imputed data using similarity matching method and the imputed data by mean substitution method. Beforehand, there are some parameter setting to be set. We used the 4th derivative of the Gaussian function and the scale parameter was set to 40.

D. Statistical Test

Paired sample t-test is a statistical test which used the comparison of mean from different sources in a dataset. In this study, we used the paired sample t-test to perform the significant test on the amplitude and WPS between the similarity matching method and the mean substitution method respectively. Apart from that, we have also used the paired sample t-test to perform the statistical test on the amplitude between the original data and the imputed data. Paired sample t-test is calculated by comparing the average difference between the samples ($\bar{D}$) to the expected difference between population means (μ_D), and then takes into account the standard error of the differences ($S_D/\sqrt{N}$). The null hypothesis is true if and only if there is no difference between the population mean [20]. The statistical test shows significant different when the *p*-value is less than 0.05.

$$t = \frac{\bar{D}-\mu_D}{S_D/\sqrt{N}} \quad (3)$$

Statistical test is necessary to validate the experimental results. In this study, we compared the amplitude between the original data and the imputed data using similarity matching method; and the amplitude between the original data and the imputed data using mean substitution method. Besides, we also compared the value of WPS between the original data and the imputed data using similarity matching method, and the value of WPS between original data and the imputed data using mean substitution method. It is because the phase information is proven better than the amplitude information.

V. Experimental Results and Discussion

In this section, the experimental results are presented and discussed. The experimental results were validated from four different perspectives, i.e. (1) the comparison of amplitude between the original data and the imputed data using the proposed similarity matching method; (2) the comparison of amplitude between the original data and the imputed data using the mean substitution method; (3) the comparison of WPS between the original data and the imputed data using the similarity matching method; (4) the comparison of WPS between the original data and the imputed data using the mean substitution method. Fig. 2 shows the comparison of grand average amplitude of the original data, the imputed data using the similarity matching method, and the imputed data using mean substitution method. Meanwhile, Fig. 3 shows the statistical test of grand average in amplitude between the original data and the imputed data by using the similarity matching method and the mean substitution method. Both the similarity matching method and the mean substitution method achieved good results. In this study, non-significant different specifies a better approximation of imputation data towards its original form. The imputed data in most of the trials, 23 trials out of a total of 41 trials are significantly close to the original data. However, different methods recorded different sets of non-significantly distinct pairs in the experiment. A total of 16 pairs treated by both of the comparison methods, as shown in bold style, shared the same non-significant validation results to their original data points. On the other hands, an additional of 7 trials treated by similarity matching method, as shown in blue color style, namely trial 1, 12, 31, 33, 35, 37, and 41, are considered close to the original data. A different set of 7 trials treated by mean substitution method, as shown in green color style, namely trial 3, 6, 9, 16, 17, 18, and 30, are also close to the original data. Apart from the evaluation in amplitude, we have also evaluated the quality of the imputed data using wavelet phase stability (WPS). According to [19], the phase of a signal encompasses more significant information as compared to the amplitude. Fig. 4 and Fig. 5 show the comparison of WPS and statistical test between the original data and the imputed data treated by similarity matching method and mean substitution method respectively. We can visually observe that the imputed data by similarity matching imputation method is closer to the original data. The statistical *p*-value threshold at 0.05 is shown as the horizontal dashed line in Fig. 5. There are only 7 trials, out of 41 trials, showed non-significant different between the original data and the imputed data. A total of 6 trials treated by similarity matching method, namely trial 8, 13, 14, 15, 21 and 32, are considered close to the original data. However, only 2 trials out of 41 trials treated by the mean substitution method, namely trial 21 and 23, are considered close to the original data. The proposed similarity matching method is able to treat 9.75% more missing value trials, with significantly better imputation value, than the mean substitution method.

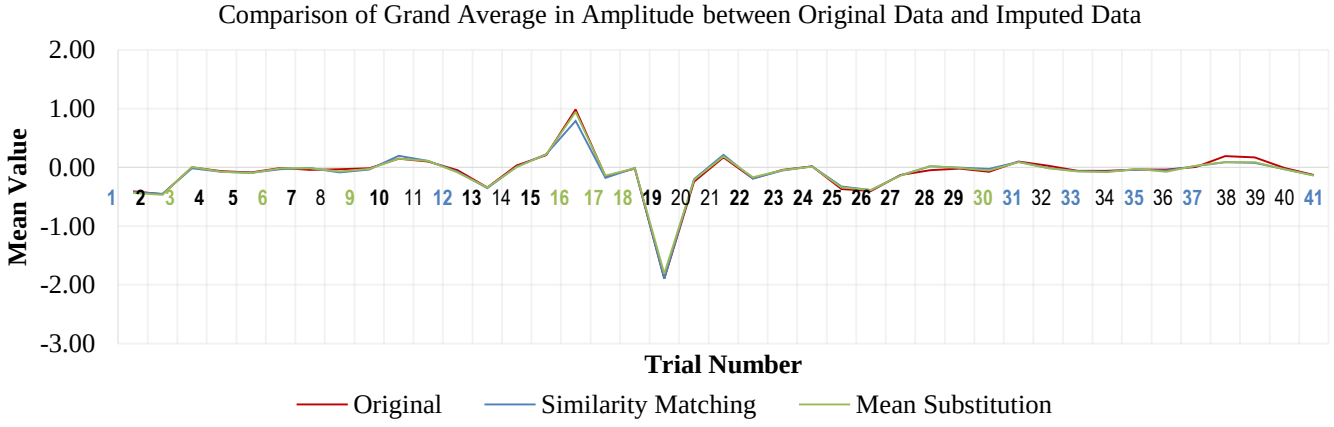

Fig. 2. Comparison of Grand Average in Amplitude between Original Data and Imputed Data.

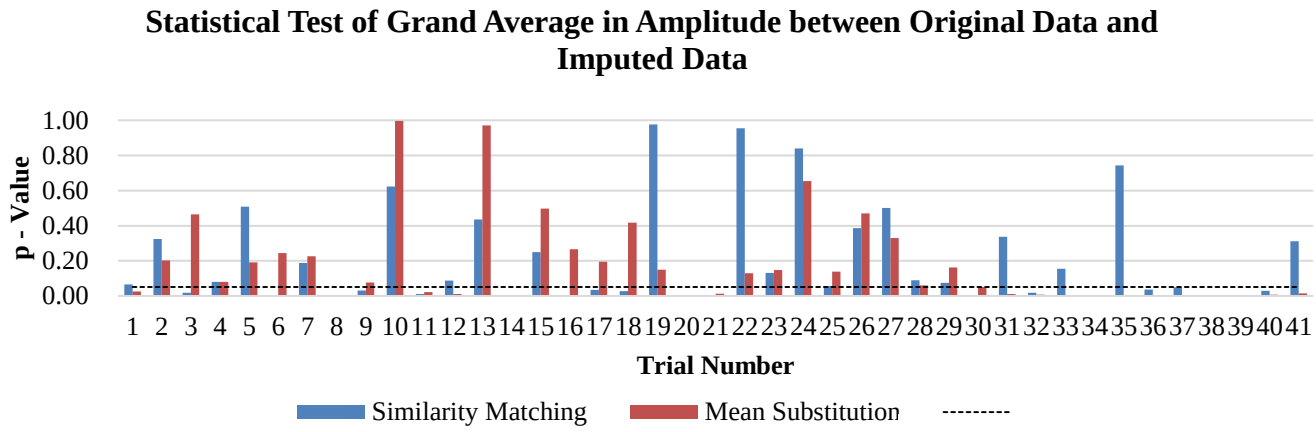

Fig. 3. Statistical Test of Amplitude between Original Data and Imputed Data. Note that the Horizontal Dashed Line in the Figure Indicates the Significant Level p – Value < 0.05.

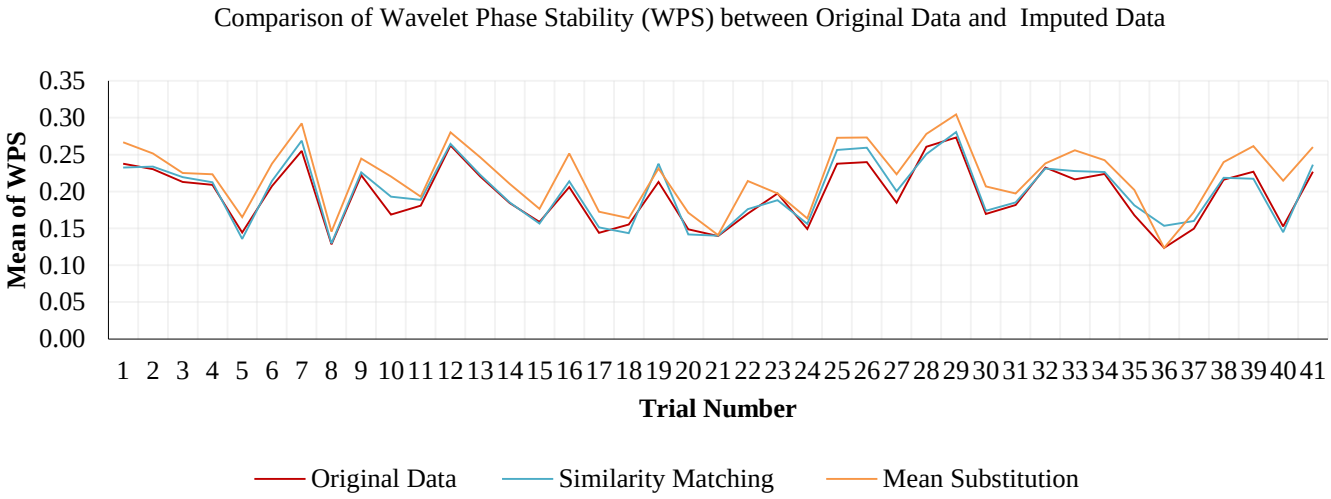

Fig. 4. Comparison of Wavelet Phase Stability (WPS) between Original Data and Imputed Data.

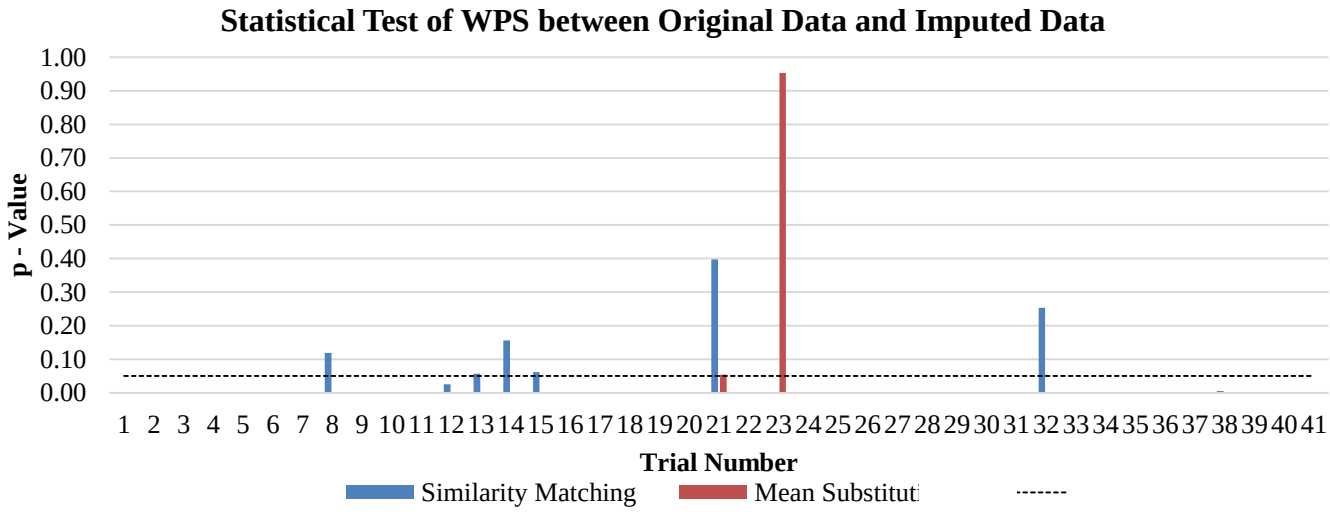

Fig. 5. Statistical Test of WPS between Original Data and Imputed Data. Note that the Horizontal Dashed Line in the Figure Indicates the Significant Level p – Value < 0.05.

TABLE I. MEAN DIFFERENCE OF WPS

Trial	Similarity Matching	Mean Substitution
1	**0.0050**	0.0292
2	**0.0037**	0.0216
3	**0.0063**	0.0120
4	**0.0032**	0.0144
5	**0.0084**	0.0210
6	**0.0068**	0.0302
7	**0.0140**	0.0373
8	**0.0016**	0.0175
9	**0.0041**	0.0224
10	**0.0246**	0.0520
11	**0.0079**	0.0122
12	**0.0023**	0.0178
13	**0.0022**	0.0262
14	**0.0011**	0.0262
15	**0.0022**	0.0176
16	**0.0078**	0.0456
17	**0.0077**	0.0289
18	0.0119	**0.0086**
19	0.0244	**0.0182**
20	**0.0069**	0.0224
21	**0.0005**	0.0010
22	**0.0064**	0.0443
23	0.0091	**0.0000**
24	**0.0068**	0.0141
25	**0.0188**	0.0352
26	**0.0194**	0.0334
27	**0.0157**	0.0387
28	**0.0100**	0.0172
29	**0.0073**	0.0311
30	**0.0043**	0.0373
31	**0.0035**	0.0154
32	**0.0010**	0.0056
33	**0.0116**	0.0396
34	**0.0029**	0.0188
35	**0.0133**	0.0342
36	0.0298	**0.0000**
37	**0.0101**	0.0224
38	**0.0025**	0.0235
39	**0.0099**	0.0346
40	**0.0077**	0.0620
41	**0.0094**	0.0337
Average	**0.0086**	**0.0250**

The mean difference of the WPS between the original data and the imputed data was evaluated to further validate the performance of both the comparison methods. Refer to Table 1. The values in bold style indicate the lower mean difference of WPS as compared to the other imputation method. The lower the mean difference between the original data and the imputed data, the better the performance of the imputation method. The similarity matching method is superior than the mean substitution method from 37 comparisons out of the total of 41 pairs. In opposite, the mean substitution method has outperformed the proposed similarity matching method in merely 4 comparison pairs from the perspective of WPS mean difference. The average of the mean difference of similarity matching method was recorded at 0.0086 only, which is 0.0164 lower than the mean substitution method.

VI. CONCLUSION

This study embarked on the motivation of treating the missing values in EEG dataset. The fundamental concept of the proposed similarity matching imputation method lies on the hypothesis of inducing the best approximated value from other complete trials as replacement. The experimental results have proven that the proposed method is better than its benchmarking mean substitute imputation method in reconstructing the EEG signals to its original form. However, we have fixed the missing data rate to 20% in this study. Hence, it is necessary to further evaluate the robustness of the proposed method in dealing with different rate of missing values in EEG dataset.

ACKNOWLEDGMENT

The authors would like to express their appreciation to Universiti Teknikal Malaysia Melaka (UTeM) for providing the UTeM Zamalah scheme scholarship. Besides, the authors would also like to thank Ministry of Higher Education Malaysia and UTeM through the Fundamental Research Grant Scheme, FRGS/1/2017/ICT/05/FTMK-CACT/F00346.

REFERENCES

[1] H. Banville and T. H. Falk, “Recent Advances and Open Challenges in Hybrid Brain-Computer Interfacing: A Technological Review of Non-Invasive Human Research,” Brain Comput. Interfaces, vol. 3, no. 1, pp. 9–46, 2016.

[2] I. Jayarathne, M. Cohen, and S. Amarakeerthi, “Survey of EEG-based biometric authentication,” in 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST), 2017, pp. 324–329.

[3] I. Nakanishi, K. Ozaki, and S. Li, “Evaluation of the Brain Wave as Biometrics in a Simulated Driving Environment,” in IEEE International Conference Proceedings in Biometrics Special Interest Group (BIOSIG), 2012, pp. 351–361.

[4] M. Teplan, “Fundamentals of EEG Measurement,” Meas. Sci. Rev., vol. 2, no. 2, pp. 1–11, 2002.

[5] R. Little and D. Rublin, Statistical Analysis with Missing Data, Second Edi. A John Wiley & Sons, Inc., 1987.

[6] I. Pratama, A. E. Permanasari, I. Ardiyanto, and R. Indrayani, “A Review of Missing Data Handling Methods on Time-Series Data,” in International Conference on Information Technology Systems and Innovation, 2016, pp. 2349–3968.

[7] M. I. R. Tokle, “Comparison of Missing Data Imputation Methods for Improving Detection of Obstructive Sleep Apnea,” 2017.

[8] S. H. Kim, H. J. Yang, and K. S. Ng, “Incremental expectation maximization principal component analysis for missing value imputation for coevolving EEG data,” J. Zhejiang Univ. Sci. C (Computers Electron., vol. 12, no. 8, pp. 687–697, 2011.

[9] C. SieluZycki and P. Kordowski, “Maximum-likelihood estimation of channel-dependent trial-to-trial variability of auditory evoked brain responses in MEG,” Biomed. Eng. Online, vol. 13, no. 1, pp. 1–19, 2014.

[10] A. Zuquete, B. Quintela, and J. P. Silva Cunha, "Biometric Authentication using Brain Responses to Visual Stimuli," in International Conference on Bio-inspired Systems and Signal Processing, 2010, pp. 103–112.

[11] S. H. Liew, Y. H. Choo, and Y. F. Low, "Fuzzy-Rough Nearest Neighbour Classifier for Person Authentication using EEG Signals," in Proceedings of 2013 International Conference on Fuzzy Theory and Its Application, 2013, pp. 316–321.

[12] I. B. Barbosa, K. Vilhelmsen, A. Van Der Meer, V. Der Weel, and T. Theoharis, "EEG Biometrics : On the Use of Occipital Cortex Based Features from Visual Evoked Potentials," in Norsk Informatikkonferanse (NIK), 2015, pp. 1–11.

[13] B. C. Armstrong, M. V Ruiz-Blondet, N. Khalifian, K. J. Kurtz, Z. Jin, and S. Laszlo, "Brainprint: Assessing the uniqueness, collectability, and permanence of a novel method for ERP biometrics," Neurocomputing, vol. 166, pp. 59–67, 2015.

[14] S. H. Liew, Y. H. Choo, Y. F. Low, and Z. I. Mohd Yusoh, "EEG-based biometric authentication modelling using incremental fuzzy-rough nearest neighbour technique," IET Biometrics, vol. 7, no. 2, pp. 145–152, 2018.

[15] J. Fell, "Cognitive neurophysiology: Beyond averaging," Neuroimage, vol. 37, no. 4, pp. 1069–1072, 2007.

[16] "Practical Introduction to Frequency-Domain Analysis," MathWorks, 2018. [Online]. Available: https://www.mathworks.com/help/signal/examples/practical-introduction-to-frequency-domain-analysis.html. [Accessed: 26-Oct-2018].

[17] A. V. Oppenheim and J. S. Lim, "The Importance of Phase in Signals," Proc. IEEE, vol. 69, no. 5, pp. 529–541, 1981.

[18] X. Ni and X. Huo, "Statistical interpretation of the importance of phase information in signal and image reconstruction," Stat. Probab. Lett., vol. 77, no. 4, pp. 447–454, 2007.

[19] Y. F. Low and D. J. Strauss, "A performance study of the wavelet-phase stability (WPS) in auditory selective attention," Brain Res. Bull., vol. 86, pp. 110–117, 2011.

[20] A. Field, Discovering Statistics using SPSS, 3rd Editio. Sage Publications, 2009.