IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 30, 2021, accepted October 11, 2021, date of publication October 14, 2021, date of current version October 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3120083

Deep Learning Approach to Generate a Synthetic Cognitive Psychology Behavioral Dataset

JUNG-GU CHOI[1], YOONJIN NAH[2], INHWAN KO[2], AND SANGHOON HAN[1,2]
[1]Graduate Program in Cognitive Science, Yonsei University, Seoul 03722, Republic of Korea
[2]Department of Psychology, Yonsei University, Seoul 03722, Republic of Korea

Corresponding author: Sanghoon Han (sanghoon.han@yonsei.ac.kr)

This work was supported by the Yonsei Signature Research Cluster Program of 2021 under Grant 2021-22-0005.

ABSTRACT Synthetic data generation is critical in machine and deep learning research to overcome the shortage of samples or dataset sizes. Various algorithms, including the generative adversarial network and autoencoder models, have been applied to generate artificial datasets in previous studies. In this study, we propose a synthetic data generation framework for a tabular dataset collected from cognitive psychology behavioral experiments based on deep learning algorithms. Tabular datasets for the Stroop task were used to develop our framework. On account of the relatively small sample size (N=102) of the dataset used in our study, we used a pre-trained generative adversarial network model to complement the size of the dataset. Furthermore, we proposed and applied five evaluation methods with statistical tests (overlapped sample test, constraint reflection test, correlation reflection test, distribution distance test, and feature distance test) to validate generation performance based on internal levels of table structure (instance level, feature level, and whole-set level evaluations). The proposed framework with a fine-tuned generative adversarial network algorithm was compared with a random generation method to verify generation performance, including the representation of the statistical characteristics of the original datasets. We found that the generated datasets from the proposed framework exhibited more similar statistical characteristics with the original dataset than the randomly generated datasets based on five evaluation methods. The results of this study provide not only generation algorithms for cognitive psychological datasets with tabular type but also a solution to the sample size issue for researchers.

INDEX TERMS Behavioral experimental dataset, cognitive psychology, data augmentation, generative adversarial network, tabular dataset, synthetic data generation.

I. INTRODUCTION

Sample or dataset size is considered a critical factor for various data analysis methodologies, including statistical and machine learning methods [1]–[4]. In terms of statistical analysis, many statistical tests require an appropriate sample size to verify the power or reliability of the results [5], [6]. For example, Lachin suggested the importance of sample size determination and power analysis in clinical trials [7]. Additionally, Maccallum *et al.* introduced a framework to determine the minimum sample size for power in empirical behavioral research [8]. In previous studies, many researchers applied formulas for sample size calculation to support the verification of their research questions or hypotheses [9]–[11]. Moreover, an adequate dataset size is essential for machine and deep learning methodologies. Ajiboye *et al.* emphasized the size of a dataset to construct supervised learning algorithms [12]. Among the three different sizes of datasets, large datasets showed lower performance errors (mean absolute errors) than other datasets. Furthermore, Sun *et al.* suggested a relationship between dataset size and model performance in visual deep learning models [13].

However, there are several reasons for the shortage of datasets in research practice. First, in the case of structured data, including survey results, lack of follow-up or nonresponse by participants can result in missing data [14], [15]. Second, in terms of unstructured data (e.g., actigraphy or electrocardiogram as a time-series), issues of devices collecting data or participants' mistakes can influence missing or blanked data [16], [17]. For instance, Brick *et al.* attempted to handle missing data due to nonresponses in a survey

The associate editor coordinating the review of this manuscript and approving it for publication was Wentao Fan.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

dataset [18]. Further, Schlomer *et al.* suggested three handling methods for data missing from participants in counseling psychology research [19]. Angerer *et al.* replaced missing data owing to the removal of devices from the body with a median value in the analysis of circadian rhythms in patients with brain injuries [20].

Researchers have attempted a statistical approach to overcome the causes of data insufficiency [21]–[24]. Imputation methods, including both single statistics (e.g., median or mean value) and integration of multiple candidate datasets (i.e., multiple imputation) were applied to treat missing data [25]–[28]. Similarly, bootstrap methods have been utilized to reduce estimation errors in the imputation process [29]. In previous studies that used machine learning or deep learning methodologies, various methods were applied to manage missing datasets. Saqib *et al.* resampled variables, including missing values, in the analysis of electronic medical records (EMRs) to predict sepsis [30]. Furthermore, Perez and Jason suggested the effectiveness of data augmentation methods in image classification using a deep learning model [31].

In behavioral science fields, such as experimental psychology, sample size is also considered an important factor for data analysis [32]–[35]. Schweizer *et al.* focused on sample size in sports psychology research [36]. They emphasized the disadvantage small sample sizes had in improving confidence in analysis results. Furthermore, Sassenberg *et al.* compared trends in social psychology research from 2011 to 2016 [37]. In Sassenberg's research, the sample size used in the study gradually increased to complement the statistical power.

In particular, large-scale datasets are considered a promising factor in the field of cognitive psychology. Peterson *et al.* suggested that large-scale datasets and machine learning algorithms can be used to identify new cognitive or behavioral phenomena [38]. They focused on risky choices and extensively studied issues in decision theory [39], [40]. In addition, Agrawal *et al.* proposed methodologies for building models and identifying novel phenomena in large datasets [41]. To overcome noise artifacts included in the datasets, they utilized sufficiently large datasets with data-driven models.

However, some challenges remain regarding the collection of large datasets through experimental research. First, in the case of repetitive and difficult tasks, participants can select extreme or incompatible answers. Inconsistent responses or outliers in the experimental results affect the overall sample size and analysis results [42], [43]. Second, negative changes in the environment, including the Covid-19 pandemic, can affect the recruitment of study participants. Suspension of follow-up for specific groups or experiments also influences the overall study [44]. Although researchers can use alternative tools, such as Amazon's Mechanical Turk (MTurk), these have potential limitations regarding research materials and conditions [45]. Consequently, several methodologies for data augmentation or resampling need to be considered to increase sample size.

In the case of tabular datasets collected in behavioral experiments in cognitive psychology studies, several characteristics can limit the application of data augmentation methodologies proposed in previous studies. First, individual variables within a dataset are deeply associated with each other [46], [47]. For example, in the popular Stroop test, the reaction time of participants refers to the reaction of participants through cognition about the proposed material, including words and colors. It indicates that the variables of reaction time and material (e.g., words, objects, and colors) are not independent. Second, different types of variables are included in the datasets [48]–[50] that consist of categorical and countable variables, not just continuous variables. For example, datasets can include age or reaction time variables as continuous variables and specific groups and levels as categorical variables. Consequently, many characteristics of the behavioral experiment dataset face challenges in applying the proposed augmentation methods (e.g., extracting, transforming, and random sampling).

In many studies, machine and deep learning methods have been applied to propose imputation or augmentation methodologies. Lashgari *et al.* introduced a data augmentation method based on a deep learning model for electroencephalography (EEG) [51]. Jang *et al.* suggested a deep-learning-based imputation methodology for missing intervals in actigraphy data [52]. In addition, Rizos *et al.* proposed a deep learning method for short-text data augmentation in speech classification [53].

Based on diverse methods with machine and deep learning models, we attempted to suggest a data generation framework for synthetic behavioral datasets with deep learning models. Various algorithms have been used to propose data generation frameworks in previous studies. Semeniuta *et al.* utilized convolutional variational autoencoder algorithms to generate text datasets [54]. Similarly, Guan *et al.* suggested a generation method for electronic medical record datasets using generative adversarial network (GAN) algorithms [55].

In our study, we applied GAN algorithms to generate a behavioral experiment dataset with a tabular structure. To improve the performance of the framework, pre-trained GAN algorithms for tabular datasets were applied [56]. We fine-tuned a pre-trained algorithm using an open-source psychology behavioral experiment dataset with a Stroop task collected from 102 participants [57], [58]. We generated 1000 datasets by applying both deep learning algorithms and random generation and compared the results to evaluate the performance of our framework. Moreover, we proposed five evaluation tests using an internal dataset level. First, an overlapped sample test was applied to evaluate whether the deep learning model simply copied the data. Second, the constraint reflection test evaluated the reflection of the range of individual variables (minimum, median, and maximum values) in the generated dataset with the original range. The first and second methods checked differences with the original dataset at the instance and row levels (i.e., instance level evaluation). Third, the correlation between the variables in the generated dataset

was examined using the correlation reflection test. Fourth, the distances of variables between the original and generated datasets were evaluated using the distribution distance test. In the third and fourth tests, we investigated statistical characteristics in terms of variable and feature levels (i.e., feature-level evaluation). Finally, in the feature distance test, we compared the feature distance with the extracted latent features using a pre-trained AlexNet model. In the last test, latent features inherent in the dataset were compared using Euclidean and Manhattan distances (i.e., whole-set level evaluation). Based on the five aforementioned tests, we examined whether the generated dataset had statistical characteristics similar to those of the original dataset.

The objective of this study was to develop a synthetic behavioral experiment dataset generation framework based on open-source Stroop task data using GAN algorithms. The major contributions of this study are as follows:

(1) We proposed a GAN-based data generation framework for a behavioral experiment dataset in the field of cognitive psychology based on an open-source Stroop task dataset. In addition, we applied a relatively large dataset (N=102) to reflect the statistical characteristics of Stroop tasks. We also evaluated the generation performance of the framework compared to a randomly generated dataset.

(2) Advancing from generating a synthetic tabular dataset of behavioral experiment data, we proposed five individual tests based on statistical tests (overlapped sample test, constraint reflection test, correlation reflection test, distribution distance test, and feature distance test) to examine various characteristics in the generated dataset. Furthermore, we compared the generation performances at three levels for the tabular dataset (instance level, feature level, and whole-set level evaluation) based on the five tests.

(3) Based on the synthetic dataset with similar statistical characteristics, our framework can help overcome a shortage in sample size. In addition, environmental restrictions, including the Covid19 pandemic, on conducting experimental studies can be overcome with artificial datasets. Furthermore, the fatigue or physical burden of participants can be reduced by complementing with generated datasets.

The remainder of the paper is structured as follows: Section II includes a detailed description of the methodologies and Stroop task dataset used in the study. In Section III, the generation performance of the proposed deep learning-based framework is described. In Section IV, we discuss the results of the experiments and their implementation. Finally, conclusions and a summary of our study are presented in Section V.

II. METHODS

A. OVERVIEW

This study consisted of four phases. First, we collected behavioral experimental datasets from cognitive psychology research. Second, the pre-trained GAN algorithm was fine-tuned using the datasets collected in the first phase. Third, we generated 1000 datasets with the same sample size using a fine-tuned GAN model and random generation methods. Finally, five evaluation tests were conducted to examine the generated datasets. The detailed procedure is shown in Figure 1.

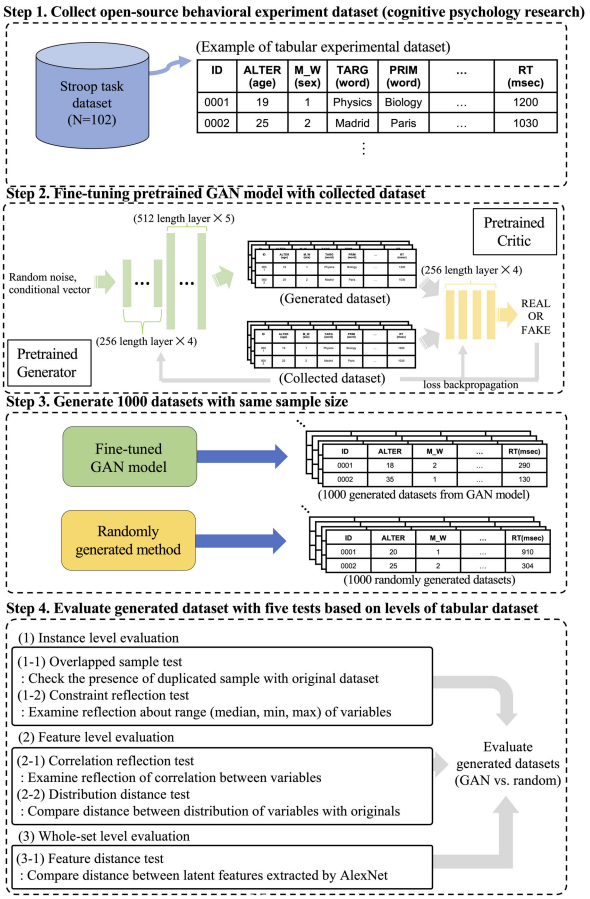

FIGURE 1. Schematic overview of this study.

B. DATA SOURCES

In this study, we used the open-source cognitive psychology dataset released by the Leibniz Institute for Psychology (ZPID) in Germany [57]. Several psychological datasets, one based on a Stroop task that is well-known as an experimental design in cognitive psychology, were selected for our experiment, collected from 102 participants (54 females and 48 males) to examine associative and affective congruency effects. Two words (priming and target words) were successively shown to evaluate the priming effect of words. After being shown the priming words, the participants were instructed to choose the terms associated with the words shown earlier. Their responses were collected vocally using a microphone. The reaction time of the participants regarding selection was recorded to evaluate the priming effects. To precisely measure their responses, several variables related to the response were stored in the dataset files. This dataset consists of 21 columns, and the descriptions of each column are listed in Table 1.

TABLE 1. Descriptions of variables in used dataset.

No.	Variables (type)	Description
1	VPID (continuous)	Participant ID
2	MAT_GR (categorical)	Material group of experimental object
3	LQ_GR (categorical)	First positions of list numbering about materials
4	B_GR (categorical)	Second digits of list numbering about materials
5	C_GR (categorical)	Third digits of list numbering about materials
6	M_W (categorical)	Sex
7	ALTER (continuous)	Age
8	HAND (categorical)	Dominant hand
9	HAND2 (categorical)	Dominant hand when writing
10	SEMESTER (continuous)	Number of semesters of participants
11	FACH (categorical)	Major of participants
12	ACTVP1 (categorical)	First position number for subject identification
13	AVTVP2 (categorical)	Second position number for subject identification
14	TARG (string)	Target word
15	PRIM (string)	Priming word
16	VBED (categorical)	Test conditions
17	SEQ (continuous)	Number of trials
18	ERR (categorical)	Selection error
19	RT_STOP (continuous)	Registered reaction time about task
20	RT1 (continuous)	Response time for response adjustment
21	RT2 (continuous)	Response time for response adjustment

Additionally, the dataset consists of two sub-datasets of experiments with different objectives. In the first sub-dataset, category-specific priming effects were examined using related words and reaction times. The priming effect is a phenomenon that affects reaction through exposure to certain stimuli (e.g., words or colored objects) [59].

In the case of category-specific priming effects, researchers wanted to confirm the effect by showing words belonging to a similar semantic category and examining participants' responses to them.

In the second sub-dataset, the color condition was added to the task design of the first sub-dataset. The dimensions of the first sub-dataset were (5184, 21) (number of rows and columns), and the second sub-dataset were (9216, 21).

C. DATA PREPROCESSING

1) SELECT VARIABLES FROM DATASET FOR EXPERIMENTS

To apply the appropriate characteristics of the Stroop task dataset, we extracted only eight variables (VPID, MAT_GR, M_W, ALTER, TARG, PRIM, RT1, and RT2) from 21. We selected variables based on the need for data analysis because of the practical applicability of the generated datasets. First, demographic information (M_W and ALTER) was needed to consider differences in age and sex in the behavioral results. Second, three variables (MAT_GR, TARG, and PRIM) for the experimental material were selected to check the words shown to the participants. Third, reaction time variables (RT1 and RT2) were selected to evaluate the effects of priming and target words. After selecting these variables, the dimensions of the dataset were changed from (5184, 21) and (9216, 21) to (5184, 8) and (9216, 8), respectively.

2) REMOVE MISSING OR EXTREME SAMPLES IN DATASET

In the Stroop task dataset, missing or extreme values in the reaction time of participants were coded with '9999' values. We confirmed the distributions of the three continuous variables (ALTER, RT1, and RT2) to check the overall distribution. Based on this confirmation, we established that the RT1 and RT2 variables in the Stroop 1 sub-dataset included extreme samples. After removing rows with extreme RT1 and RT2 values, the Stroop 1 sub-dataset had dimensions of (5180, 8) without missing or extreme samples. In the case of the Stroop 2 sub-dataset, extreme samples were not included in the dataset. Therefore, the dimensions of the Stroop 2 sub-dataset did not change. Histograms of the distribution of variables are shown in Figure 2.

D. PRE-TRAINED GAN MODEL FOR TABULAR DATASET

In this study, we attempted to generate a behavioral experimental dataset. The Stroop task dataset was relatively small to train and evaluate deep learning algorithms from scratch and achieve high performance. To complement the sample size, transfer learning and fine-tuning of a pre-trained algorithm were applied. To improve the performance of our framework, pre-trained conditional GAN models with tabular datasets were used in our study [56]. This model, similar to other general GAN algorithms, consists of two sub-modules (i.e., generator and discriminator). To generate tabular datasets with data distributions, conditional vectors were concatenated in the calculation process of the generators. In addition, normalization was applied to each feature to deal with complicated distributions. The authors named the generator module containing the conditional vector the ''conditional generator.'' A total of ten fully connected layers constructed conditional generators. Except for the input and output layers, three hidden layers in the front were composed of 256 neurons, and the five hidden layers in the back were composed of 512 neurons. ReLU activation functions and batch normalization were applied to each layer.

The discriminator module (i.e., Critics) in this model consisted of five fully connected layers. A discriminator module was constructed with 256 hidden layers. Leaky ReLU activation functions and dropout with a 0.2 ratio were used for four hidden layers without the input and output layers. For model training, the Adam optimizer and a 2×10^{-4} learning rate were used. The detailed parameters of the model are listed in Table 2.

These algorithms were validated with Census Income, KDD Cup 1999 Data, and the Online News Popularity dataset, which consists of tabular structures.

TABLE 2. Parameters of pre-trained ct-gan model.

Layer (Module)	Length of layer	Activation function	Other conditions
Conditional generator			
(input)	z ⊕ conditional vector		
1	256	ReLU	Batch normalization
2	256	ReLU	Batch normalization
3	256	ReLU	Batch normalization
4	256	ReLU	Batch normalization
5	512	Hyperbolic tangent	
6	512	Hyperbolic tangent	
7	512	Gumbel softmax	
8	512	Gumbel softmax	
9	512	Gumbel softmax	
(output)	D_i	Gumbel softmax	
Discriminator (Critic)			
(input)	output of conditional generator ⊕ conditional vector		
1	256	Leaky ReLU	Dropout (0.2)
2	256	Leaky ReLU	Dropout (0.2)
3	256	Leaky ReLU	Dropout (0.2)
4	256	Leaky ReLU	Dropout (0.2)
(output)	1		

z: noise vector; ⊕: concatenate operation; D_i: output vectors of generator (equal to length of conditional vector)

We selected and applied pre-trained conditional GAN algorithms because the size of our dataset was relatively insufficient for building a model; the pre-trained GAN algorithms were also trained with similar types of tabular datasets. To increase the usability of the proposed framework in terms of reproducibility, we used the same parameters (e.g., model architecture) and hyperparameters (e.g., optimizer or learning rate) of the pre-trained model for fine-tuning. Consequently, for fine-tuning the pre-trained algorithm, Adam optimizer, 2×10^{-4} learning rate, and 300 epochs were used as training hyperparameters.

E. GENERATED TABULAR DATASET BY PRE-TRAINED GAN MODEL

After fine-tuning the pre-trained conditional GAN model, we generated 1000 datasets with the same sample size as the original datasets to evaluate the generation performance of the framework. For example, in the case of the Stroop 1 sub-dataset, 1000 different datasets with dimensions (5180, 8) were generated. The Stroop 2 sub-dataset was applied to generate 1000 datasets with dimensions (9216, 8).

F. RANDOMIZE GENERATED DATASET

To evaluate the generation performance with fine-tuned GAN algorithms with a Stroop task dataset, we randomly generated 1000 datasets with the same sample size. Instances in the generated datasets were selected from the column values of the original dataset. For example, ALTER column values in the generated dataset were randomly selected within values of the same variable from the original datasets.

Based on this process, we generated 1000 datasets for the Stroop 1 and Stroop 2 sub-datasets. The dimensions of the randomly generated datasets were the same as those of the original dataset.

G. EVALUATION METHODS BY LEVEL OF DATASET

In our study, we proposed a deep-learning-based generation framework for a synthetic behavior experiment dataset. For a detailed evaluation of the framework, we considered the inherent levels of the tabular dataset. A total of three standards were applied. First, instance and row-level evaluations were considered. An overlapped sample was checked for an instance and row in the dataset. Second, variable and feature levels were applied. The distribution and characteristics of the variables were confirmed. Finally, a whole-set level evaluation was performed. In the case of the whole-set level, the overall characteristics and latent features of the dataset were compared. Detailed descriptions of each evaluation method are provided in the following subsections.

1) INSTANCE LEVEL EVALUATION

a: OVERLAPPED SAMPLE TEST

In this test, we attempted to confirm the overlapped samples using the original dataset. If there is an overlapped sample in the generated datasets, it is considered a copy rather than a generation. To verify the duplicated samples, we organized a test in four steps. First, both RT1 and RT2 values for the same word pair were extracted from the generated datasets based on the TARG and PRIM words in the original dataset. Second, a one-sample t-test was applied to examine the difference between the RT1 and RT2 values in the original and the generated datasets.

The null hypothesis of the test was that the RT1 and RT2 values in the generated dataset were the same as the original values. Third, the number of word pairs (TARG and PRIM) was statistically significant. Finally, the test index was calculated as the ratio of the number of statistically significant results among all the results.

An example of the calculation in the test is depicted in Figure 3.

b: CONSTRAINT REFLECTION TEST

Each variable had a range of values that required verification of whether the generated values in the row were included in the range of the original variables. We attempted to confirm the reflection of ranges from the minimum, median, and maximum values. This test consisted of three steps. First, we calculated the minimum, median, and maximum values of the continuous variables (ALTER, RT1, and RT2) in the original dataset. Second, the same values were calculated from the generated datasets. Finally, the absolute differences between the original and generated datasets were calculated

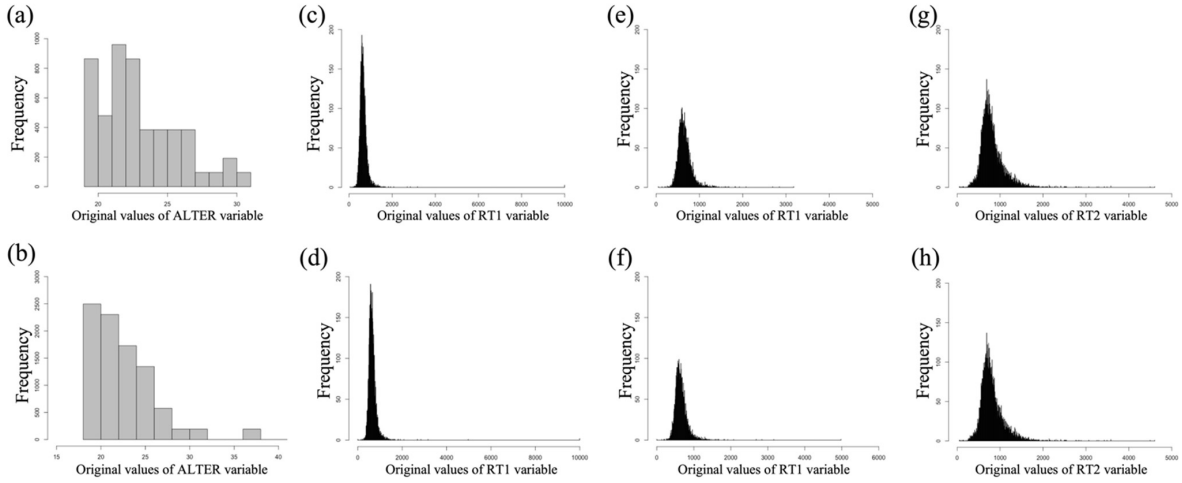

FIGURE 2. Distributions of continuous variables (ALTER, RT1, RT2). Distributions of (a) ALTER variable, (c) RT1 variable, and (g) RT2 variable in Stroop 1 sub-dataset. (b) Distributions of ALTER variable, (d) RT1 variable, and (h) RT2 variable in Stroop 2 sub-dataset. I Distributions of RT1 variables after removing extreme samples, (f) RT2 variable after removing extreme samples in Stroop 1 sub-dataset.

to evaluate the reflection status of the generated datasets. An example of the test application is shown in Figure 4.

2) FEATURE LEVEL EVALUATION

a: CORRELATION REFLECTION TEST

In the Stroop task dataset used in this study, the dataset showed a correlation between each variable. We examined the correlation reflections in the generated dataset. This test was conducted in three steps. First, we calculated the correlation coefficients for both the original and generated datasets. Second, the averaged coefficients of the variables from the generation methods were compared with the coefficients from the original dataset using absolute differences. Finally, we evaluated the reflection status of the generation methodologies by comparing the differences. An example of this test is presented in Figure 5.

b: DISTRIBUTION DISTANCE TEST

Through the preprocessing step, we confirmed that each variable in the dataset has its own distribution (Figure 2). In this test, we compared the distributions of the original and generated values of the variables. In the Stroop task dataset, five categorical variables (MAT_GR, M_W, TARG, PRIM, and ERR) and three continuous variables (ALTER, RT1, and RT2) were included. We applied the Hamming distance metric to compare the categorical variables. The Hamming distance indicates the quantified differences between two data vectors consisting of categorical data [60]. A 2-sample Kolmogorov-Smirnov (KS) test was used to compare the continuous variables.

We checked whether the two compared distributions were drowned out with the same distribution [61]. The statistical significance of the test results ($p < 0.05$) indicates that the two are drawn from the same distribution. Furthermore, KS statistics represent the quantified distance of the empirical and cumulative distribution functions between the two variables. After applying metrics for the variables, we compared the average distance values between the generated datasets. Figure 6 presents an outline of the distribution distance test.

3) WHOLE-SET LEVEL EVALUATION

a: FEATURE DISTANCE TEST

In the previous four evaluation tests, we verified the differences in fragmentary characteristics (instance and feature levels) in the dataset. Furthermore, we attempted to evaluate the inherent characteristics of the datasets using latent features. A pre-trained AlexNet model was applied to extract the latent features. Before applying the dataset to a pre-trained model, the TARG and PRIM variables were converted from word to categorical dummy values. Three conditions of the features (3, 5, and 7 feature lengths) were extracted to evaluate them.

After extracting the features, we applied the Minkowski distance metric, which indicates the generalized version of the Euclidean and Manhattan distances. The Minkowski distance was calculated using (1) [62]:

$$D(X, Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \quad (1)$$

In (1), if the value of p (power) is 1, it is the same as the Manhattan distance, which is the L1 norm. In addition, when the value of p is 2, it indicates that distance has the same meaning as Euclidean distance, which is the L2 norm. In our study, both cases, where p was 1 and 2, were evaluated. An outline of the feature distance test is shown in Figure 7.

H. STATISTICAL VERIFICATION

After we received the results of applying the five evaluation methods, we compared the characteristics and distance of the generated dataset between the fine-tuned GAN algorithm and randomized generation. To identify the differences more clearly, we used statistical tests for the evaluation results. For

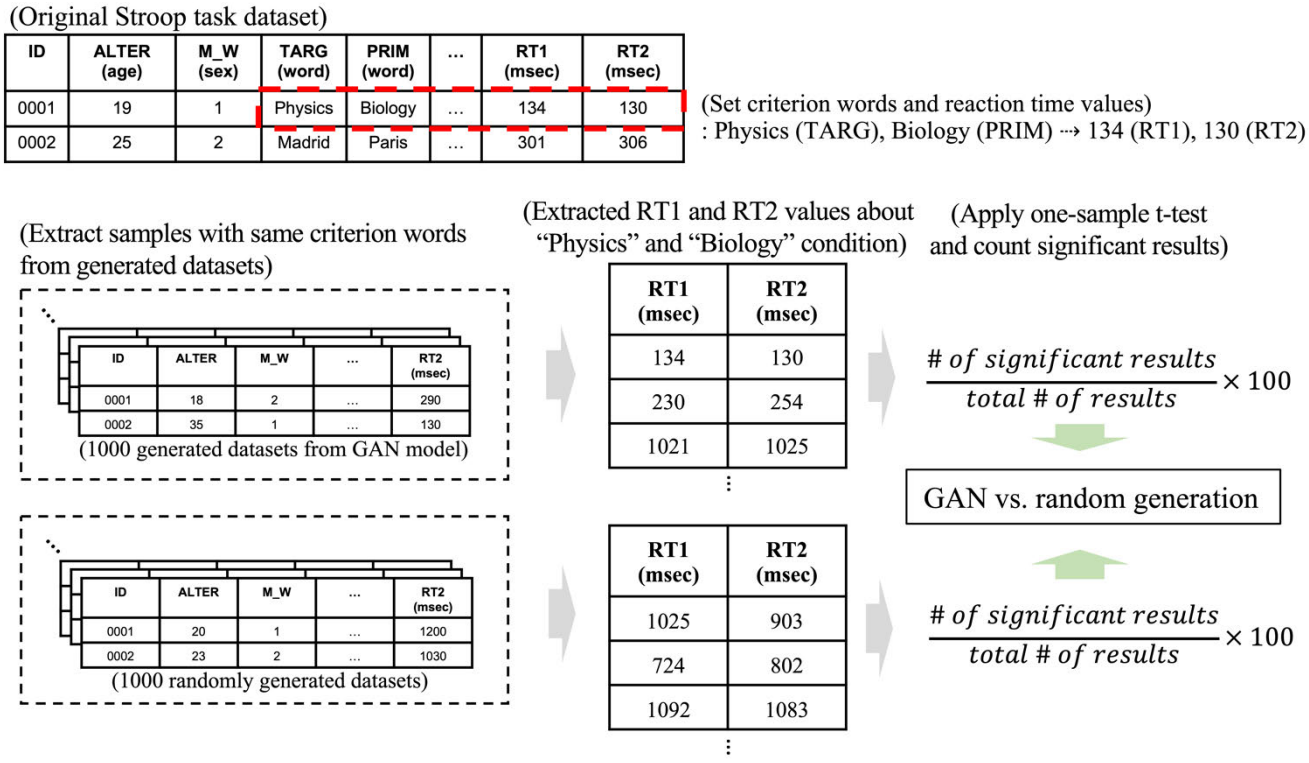

FIGURE 3. Outline of overlapped sample test.

example, we found differences in the averaged distance values between datasets from the GAN and random generation.

To confirm the differences between the two values, we used a two-sample t-test to calculate the distances. Owing to the different methodologies used for data generation, we hypothesized the independence of the two distances. The null hypothesis of the test was that the difference in the average distance values between the GAN and random generation was 0.

I. TOOLS

All code for the deep learning model and data preprocessing were written using Python (version 3.6.0) and Pytorch framework (version 10.0.1). Statistical figures are depicted using R (version 4.0.3).

III. RESULTS

We evaluated the generation performance of the proposed deep-learning-based framework using five test methods divided by the internal levels of the dataset (instance level, feature level, and whole-set level evaluation). In the case of the overlapped sample test at the instance level evaluation, we evaluated different samples in the generated dataset based on one-sample t-test results. Then, the number of samples that were validated from the p-value were calculated as a ratio of the total number of results. In the Stroop 1 sub-dataset, the GAN-based model showed 68.17% and random generation showed 57.92% of significantly different samples for the RT1 variables. Additionally, 67.47% and 65.15% were confirmed by the GAN-based and RT2 variables, respectively. In the Stroop 2 sub-dataset, we found 90.62% for the GAN-based model and 79.86% for the random generation in the RT1 variable.

For the RT2 variable, 80.38% and 84.72% were found in the GAN-based and random generation models, respectively. Table 3 lists the detailed results of the overlapped sample tests.

Additionally, in the constraint reflection test in instance level evaluation, the reflection of the range of each variable with minimum, median, and maximum values was compared. The differences between the three range values (minimum, median, and maximum) were compared with the original dataset values to evaluate the reflection status. In the Stroop 1 sub-dataset, the GAN-based model showed 2.00 (ALTER: median-minimum), 2.00 (ALTER: maximum-median), 90.31 (RT1: median-minimum), 89.93 (RT1: maximum-median), 263.66 (RT2: median-minimum), and 147.06 (RT2: maximum-median) as averaged differences. Also, we found 1.89 (ALTER: median-minimum), 1.69 (ALTER: maximum-median), 141.95 (RT1: median-minimum), 1599.38 (RT1: maximum-median), 159.41 (RT2: median-minimum), and 1711.59 (RT2: maximum-median) in randomly generated datasets. Moreover, in the case of the Stroop 2 sub-dataset, we found 0 (ALTER: median-minimum), 2.53 (ALTER: maximum-median), 133.34 (RT1: median-minimum), 200.09 (RT1: maximum-median), 313.63 (RT2: median-minimum), and 415.36 (RT2: maximum-median) in the GAN-based model. In random generation,

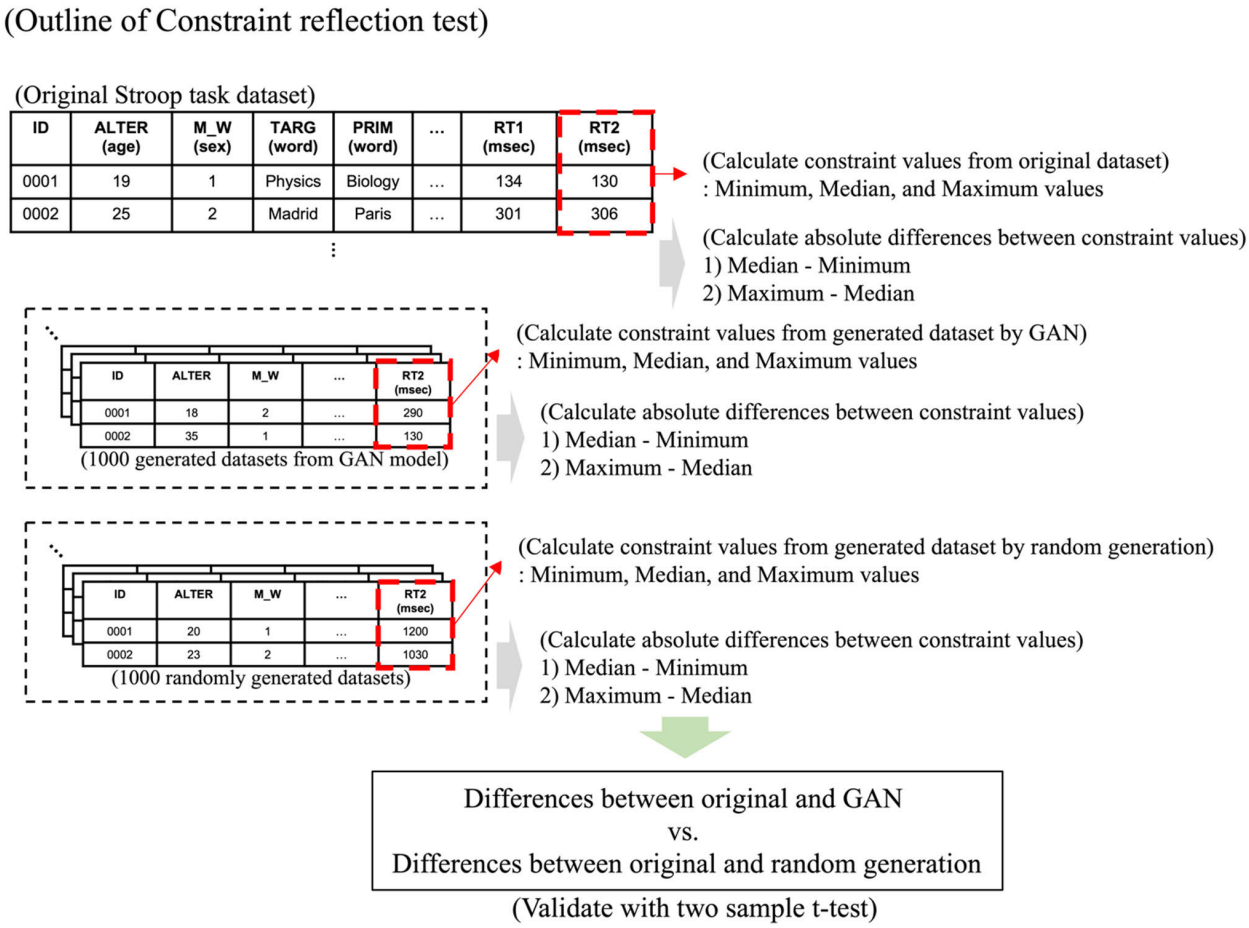

FIGURE 4. Examples of constraint reflection test (RT2 variable).

3.49 (ALTER: median-minimum), 3.49 (ALTER: maximum-median), 199.99 (RT1: median-minimum), 2064.44 (RT1: maximum-median), 415.42 (RT2: median-min), and 1928.64 (RT2: maximum-median) were checked as averaged differences. The detailed results and statistical test results are listed in Tables 4 and 5. Second, for feature level evaluation, we compared the absolute differences of the averaged correlation coefficients between the original and generated datasets using the correlation reflection test.

In the Stroop 1 sub-dataset, the GAN-based model condition showed a difference of 0.101 from the correlation coefficients of the original to the coefficients of the generated datasets by the GAN-based model.

In addition, 0.138 was observed under the random generation conditions. In the Stroop 2 sub-dataset, we checked 0.071 in the GAN-based model conditions and 0.108 in the random generation condition. The results of the correlation reflection test, statistical test, and absolute correlation coefficient value list are shown in Tables 6, 7, and 8, respectively.

Furthermore, the distances of variables' distribution were identified using the distribution distance test. In the Stroop 1 sub-dataset, the GAN-based model showed values of 0.6664 (MAT_GR), 0.4928 (M_W), 0.9964 (TARG), 0.9971 (PRIM), and 0.0809 (ERR) as a Hamming distance value.

TABLE 3. Results of overlapped sample test.

Dataset (methods)	Variables	No. of significant results (No. of total results)	Ratio (%)
Stroop 1 sub-dataset (GAN-based model)	RT1	589 (864)	68.17%
Stroop 1 sub-dataset (Random generation)	RT1	497 (858)	57.92%
Stroop 1 sub-dataset (GAN-based model)	RT2	583 (864)	67.47%
Stroop 1 sub-dataset (Random generation)	RT2	559 (858)	65.15%
Stroop 2 sub-dataset (GAN-based model)	RT1	522 (576)	90.62%
Stroop 2 sub-dataset (Random generation)	RT1	460 (576)	79.86%
Stroop 2 sub-dataset (GAN-based model)	RT2	463 (576)	80.38%
Stroop 2 sub-dataset (Random generation)	RT2	488 (576)	84.72%

The GAN-based model also showed values of 0.6669 (MAT_GR), 0.5004 (M_W), 0.9965 (TARG), 0.9974 (PRIM), and 0.5002 (ERR) for categorical variables as Hamming distance values. Additionally, 0.1765 (ALTER), 0.0647 (RT1), and 0.0508 (RT2) were found in the GAN-based models as KS statistics. Under random generation conditions,

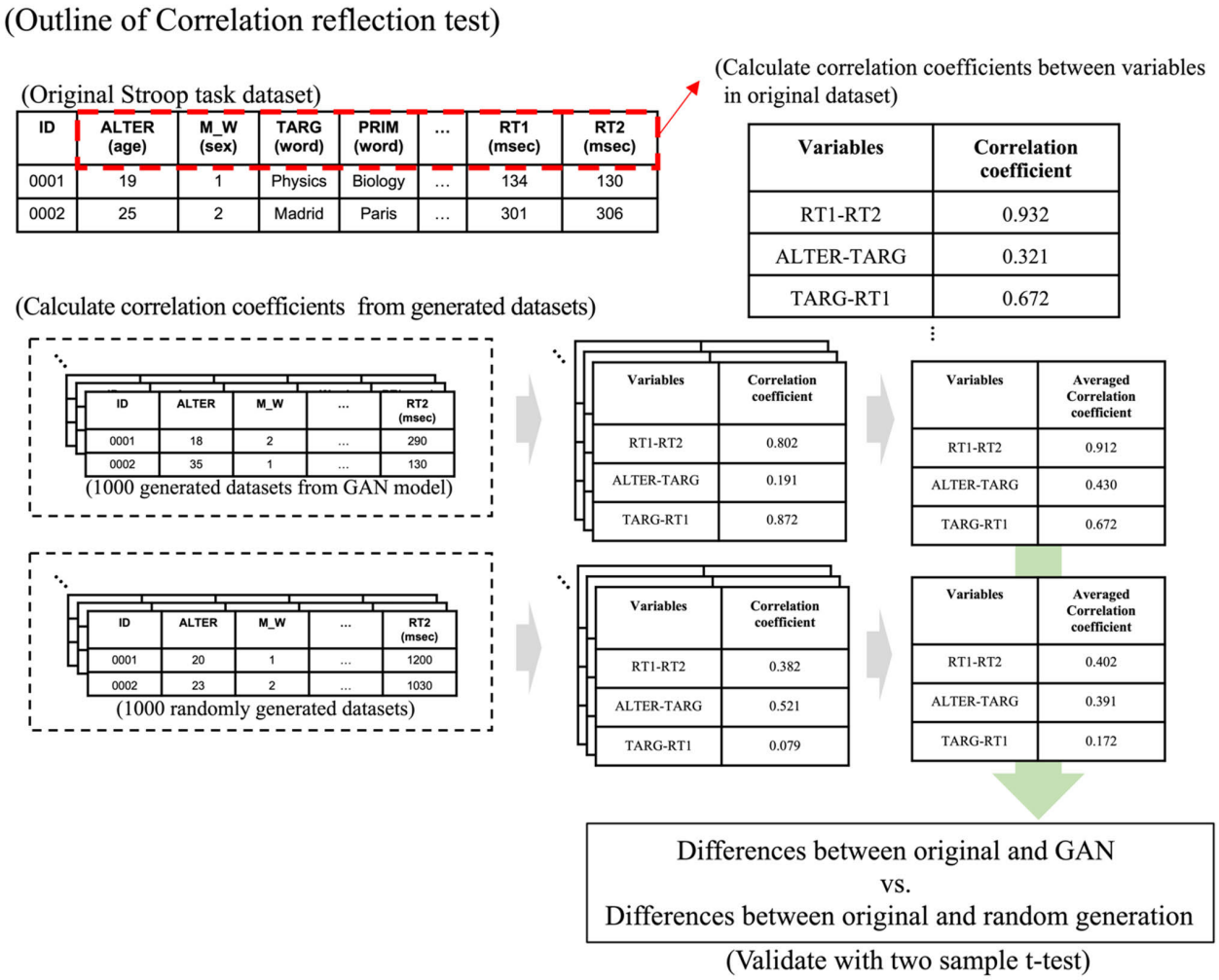

FIGURE 5. Outline of correlation reflection test (RT2 variables).

0.2274 (ALTER), 0.2717 (RT1), and 0.2705 (RT2) were checked. In the Stroop 2 sub-dataset, for categorical variables, 0.4998 (MAT_GR), 0.4999 (M_W), 0.9948 (TARG), 0.9967 (PRIM), and 0.4941 (ERR) were identified in the GAN-based model conditions; 0.5000 (MAT_GR), 0.5000 (M_W), 0.9947 (TARG), 0.9964 (PRIM), and 0.7998 (ERR) were confirmed under random generation conditions.

For continuous variables, 0.1576 (ALTER), 0.0573 (RT1), and 0.0474 (RT2) were found in the GAN-based model. In random generation conditions, 0.3125 (ALTER), 0.2934 (RT1), and 0.2934 (RT2) were checked as KS statistics. Detailed results of the distribution distance test and statistical test are listed in Tables 9 and 10, respectively. Finally, we examined the feature distance test to compare the distances of latent features between the original and generated datasets to evaluate the generated dataset at the whole-set level. Three, five, and seven length feature conditions were evaluated. In the Stroop 1 sub-dataset, in the case of seven length features, the GAN-based model showed an average distance of 34901.9; 83229.6 and 102138.0 were checked for five and three length features.

Furthermore, 73728.9, 115097.1, and 123034.9 were confirmed as averaged distance values for seven, five, and three length features, respectively.

In the Stroop 2 sub-dataset, 230547.1, 194275.8, and 190009.5 were found for seven, five, and three length feature conditions, respectively, from the GAN-based model. In addition, 348281.5, 293877.7, and 250526.9 were checked for random generation. The detailed results and statistical test results are listed in Tables 11–14.

IV. DISCUSSION

In our study, we attempted to generate a behavior experiment dataset with a tabular structure collected from cognitive psychology research and based on fine-tuned GAN algorithms. The Stroop tasks dataset was applied to verify our research agenda: artificial dataset generation for the behavioral experiment dataset using a deep learning algorithm.

To provide reasonable evidence, we reviewed several studies using ''data generation'' and ''deep learning methods'' as keywords. First, in relation to synthetic data generation, Pargas *et al.* [63] proposed data generation methods based

TABLE 4. Results of constraint reflection test.

Dataset (methods)	Variables	Differences of range	Averaged difference
Stroop 1 sub-dataset (GAN-based model)	ALTER	Median-Minimum	2.00
Stroop 1 sub-dataset (Random generation)	ALTER	Median-Minimum	1.89
Stroop 1 sub-dataset (GAN-based model)	ALTER	Maximum-Median	2.00
Stroop 1 sub-dataset (Random generation)	ALTER	Maximum-Median	1.69
Stroop 1 sub-dataset (GAN-based model)	RT1	Median-Minimum	90.30
Stroop 1 sub-dataset (Random generation)	RT1	Median-Minimum	141.95
Stroop 1 sub-dataset (GAN-based model)	RT1	Maximum-Median	89.93
Stroop 1 sub-dataset (Random generation)	RT1	Maximum-Median	1599.38
Stroop 1 sub-dataset (GAN-based model)	RT2	Median-Minimum	263.66
Stroop 1 sub-dataset (Random generation)	RT2	Median-Minimum	159.41
Stroop 1 sub-dataset (GAN-based model)	RT2	Maximum-Median	147.06
Stroop 1 sub-dataset (Random generation)	RT2	Maximum-Median	1711.59
Stroop 2 sub-dataset (GAN-based model)	ALTER	Median-Minimum	0
Stroop 2 sub-dataset (Random generation)	ALTER	Median-Minimum	3.49
Stroop 2 sub-dataset (GAN-based model)	ALTER	Maximum-Median	2.53
Stroop 2 sub-dataset (Random generation)	ALTER	Maximum-Median	3.49
Stroop 2 sub-dataset (GAN-based model)	RT1	Median-Minimum	133.34
Stroop 2 sub-dataset (Random generation)	RT1	Median-Minimum	199.99
Stroop 2 sub-dataset (GAN-based model)	RT1	Maximum-Median	200.09
Stroop 2 sub-dataset (Random generation)	RT1	Maximum-Median	2064.44
Stroop 2 sub-dataset (GAN-based model)	RT2	Median-Minimum	313.63
Stroop 2 sub-dataset (Random generation)	RT2	Median-Minimum	415.42
Stroop 2 sub-dataset (GAN-based model)	RT2	Maximum-Median	415.36
Stroop 2 sub-dataset (Random generation)	RT2	Maximum-Median	1928.64

on genetic algorithms with a population dataset. They suggested the advantages of test data generation in related studies. In addition, Tracey *et al.* [64] suggested an automatic data generation framework for structural datasets; thus, they applied dynamic optimization-based search methods to the framework. Furthermore, they demonstrated the efficiency and effectiveness of the test data generation by comparing various experimental conditions. Brissette *et al.* [65] attempted to generate synthetic weather datasets based on stochastic methods. Methodologies using the Wilks approach were used in this study to generate weather information from multiple sites. Advantages, including simplicity and complements to climate research, have been emphasized by

TABLE 5. Statistical test results of constraint reflection test.

Comparison condition (variable / dataset)	Differences of range	t statistics	*p*-value (one-tail)	*p*-value (two-tail)
Random vs. GAN-based (ALTER / Stroop 1 sub-dataset)	Median-Minimum	-10.710	1.03E-10	2.05E-10
Random vs. GAN-based (ALTER / Stroop 1 sub-dataset)	Maximum-Median	-21.186	7.23E-20	1.45E-20
Random vs. GAN-based (RT1/ Stroop 1 sub-dataset)	Median-Minimum	44.363	1.12E-23	2.24E-23
Random vs. GAN-based (RT1 / Stroop 1 sub-dataset)	Maximum-Median	522.378	1.43E-10	2.43E-10
Random vs. GAN-based (RT2 / Stroop 1 sub-dataset)	Median-Minimum	-10.520	6.45E-25	1.29E-24
Random vs. GAN-based (RT2 / Stroop 1 sub-dataset)	Maximum-Median	695.883	1.79E-15	3.29E-16
Random vs. GAN-based (ALTER / Stroop 2 sub-dataset)	Median-Minimum	221.124	1.13E-09	1.29E-09
Random vs. GAN-based (ALTER / Stroop 2 sub-dataset)	Maximum-Median	43.058	3.57E-23	7.15E-23
Random vs. GAN-based (RT1/ Stroop 2 sub-dataset)	Median-Minimum	83.468	5.32E-05	3.40E-05
Random vs. GAN-based (RT1 / Stroop 2 sub-dataset)	Maximum-Median	334.512	3.32E-11	4.09E-11
Random vs. GAN-based (RT2 / Stroop 2 sub-dataset)	Median-Minimum	10.373	2.67E-24	5.35E-24
Random vs. GAN-based (RT2 / Stroop 2 sub-dataset)	Maximum-Median	139.215	3.11E-14	2.97E-14

TABLE 6. Results of correlation reflection test.

Dataset	Generation method	Absolute differences of correlation coefficient
Stroop 1 sub-dataset	GAN-based model	0.101
Stroop 1 sub-dataset	Random generation	0.138
Stroop 2 sub-dataset	GAN-based model	0.071
Stroop 2 sub-dataset	Random generation	0.108

TABLE 7. Statistical test results of correlation reflection test.

Comparison conditions (dataset)	t statistics	*p*-value (one-tail)	*p*-value (two-tail)
GAN-based model vs. Random generation (Stroop 1 sub-dataset)	-1.5979	0.0209	0.0117
GAN-based model vs. Random generation (Stroop 2 sub-dataset)	-1.0000	0.0016	0.0325

the authors. Jones *et al.* [66] attempted to generate mutation data from protein sequences using raw mutation frequency matrices to generate and evaluate datasets. In addition, the generated datasets were validated using the SWISS-PROT database. Their study proposed the benefits of dataset generation for associated research.

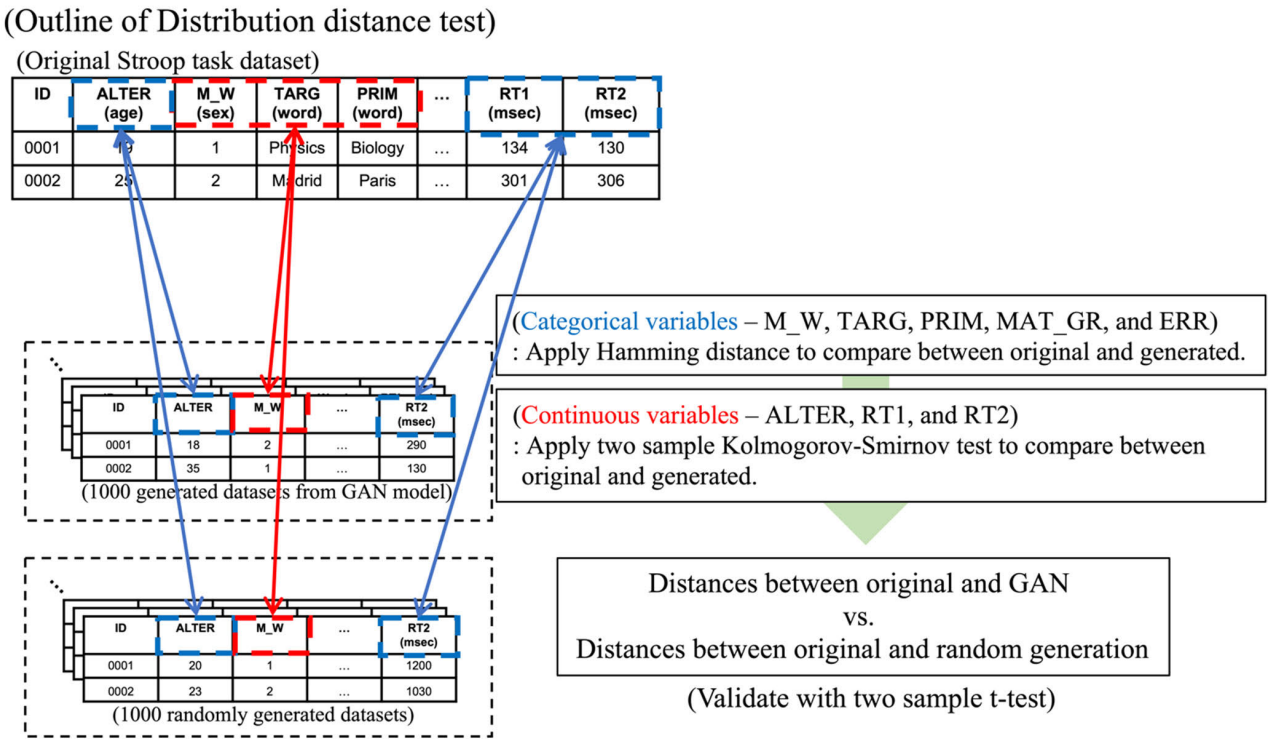

FIGURE 6. Outline of distribution distance test.

TABLE 8. Comparison of correlation coefficients ranking.

Rank	GAN-based model	Correlation coefficient	Original dataset	Correlation coefficient
Stroop 1 sub-dataset				
1	RT2-RT1	0.46108	RT2-RT1	0.96951
2	MAT_GR-M_W	0.29220	MAT_GR-TARG	0.70584
3	MAT_GR-RT1	0.22547	M_W-ALTER	0.27652
4	MAT_GR-RT2	0.19514	MAT_GR-M_W	0.13693
5	M_W-ALTER	0.14438	M_W-TARG	0.12905
6	MAT_GR-TARG	0.10034	ALTER-TARG	0.08029
7	TARG-PRIM	0.09047	MAT_GR-RT2	0.06770
8	M_W-RT1	0.08031	MAT_GR-RT1	0.06098
9	M_W-TARG	0.07570	TARG-RT2	0.05801
10	ALTER-RT2	0.07552	TARG-RT1	0.05364
Stroop 2 sub-dataset				
1	RT1-RT2	0.58019	RT2-RT1	0.99997
2	M_W-ALTER	0.09793	MAT_GR-TARG	0.86516
3	ALTER-RT1	0.09612	M_W-ALTER	0.19154
4	ALTER-RT2	0.09197	ALTER-RT2	0.09258
5	TARG-ERR	0.08454	ALTER-RT1	0.09257
6	MAT_GR-M_W	0.05297	MAT_GR-RT2	0.04569
7	MAT_GR-RT1	0.04254	MAT_GR-RT1	0.04562
8	ALTER-ERR	0.03821	TARG-RT2	0.04216
9	MAT_GR-RT2	0.03814	TARG-RT1	0.04213
10	MAT-GR-PRIM	0.03733	ERR-RT2	0.03333

TABLE 9. Results of distribution distance test.

Dataset	Variables (distance metric)	GAN-based model	Random generation
Stroop 1 sub-dataset			
	MAT_GR (Hamming distance)	0.6664	0.6669
	M_W (Hamming distance)	0.4928	0.5004
	TARG (Hamming distance)	0.9964	0.9965
	PRIM (Hamming distance)	0.9971	0.9974
	ERR (Hamming distance)	0.0809	0.5002
	ALTER (2-sample KS test)	0.1765	0.2274
	RT1 (2-sample KS test)	0.0647	0.2717
	RT2 (2-sample KS test)	0.0508	0.2705
Stroop 2 sub-dataset			
	MAT_GR (Hamming distance)	0.4998	0.5000
	M_W (Hamming distance)	0.4999	0.5000
	TARG (Hamming distance)	0.9948	0.9947
	PRIM (Hamming distance)	0.9967	0.9964
	ERR (Hamming distance)	0.4941	0.7998
	ALTER (2-sample KS test)	0.1576	0.3125
	RT1 (2-sample KS test)	0.0573	0.2934
	RT2 (2-sample KS test)	0.0474	0.2934

Second, as mentioned above, methodologies with statistical or mathematical approaches have been applied to various datasets (e.g., protein structure, climate dataset, and population dataset) for synthetic data generation.

Similarly, in terms of machine and deep learning algorithms, diverse algorithms have been utilized for data generation. Guo and Herna [67] suggested boosting and generation methodologies to complement imbalanced data using boosting and ensemble-based learning algorithms. Researchers have evaluated improvements in data generation with respect to the prediction power of classification algorithms using synthetic datasets. In Bloice *et al.* [68], augmentation methods were based on machine-learning models

TABLE 10. Statistical test results of distribution distance test.

Comparison condition (variable)	t statistics	*p*-value (one-tail)	*p*-value (two-tail)
Stroop 1 sub-dataset			
GAN-based vs. Random (MAT_GR)	-1.5408	0.0061	0.0123
GAN-based vs. Random (M_W)	-25.2818	1.04E-09	2.07E-09
GAN-based vs. Random (TARG)	-0.5726	0.0283	0.0366
GAN-based vs. Random (PRIM)	-8.2200	3.16E-16	6.32E-16
GAN-based vs. Random (ERR)	-1758.8501	1.25E-03	2.97E-05
GAN-based vs. Random (ALTER)	-178.3190	1.62E-09	1.21E-08
GAN-based vs. Random (RT1)	-843.7316	6.25E-11	6.92E-11
GAN-based vs. Random (RT2)	-871.5910	9.21E-07	3.40E-08
Stroop 2 sub-dataset			
GAN-based vs. Random (MAT_GR)	-0.8338	0.0202	0.0404
GAN-based vs. Random (M_W)	-0.6316	0.0263	0.0277
GAN-based vs. Random (TARG)	2.7389	0.0031	0.0062
GAN-based vs. Random (PRIM)	12.8170	3.14E-05	6.29E-05
GAN-based vs. Random (ERR)	-1614.2179	3.02E-05	2.01E-05
GAN-based vs. Random (ALTER)	-740.4648	1.74E-08	2.31E-08
GAN-based vs. Random (RT1)	-1191.7880	2.30E-05	9.32E-05
GAN-based vs. Random (RT2)	-1387.8429	5.45E-07	3.29E-06

TABLE 11. Results of feature distance test in stroop 1 sub-dataset.

Dataset / length of features	Features (distance metric)	GAN-based model	Random generation
7 Features	Feature 1 (Manhattan)	68556.6	144869.0
	Feature 1 (Euclidean)	1229.6	2653.2
	Feature 2 (Manhattan)	68584.1	144785.3
	Feature 2 (Euclidean)	1229.8	2652.3
	Feature 3 (Manhattan)	68567.3	144730.0
	Feature 3 (Euclidean)	1229.4	2651.1
	Feature 4 (Manhattan)	68566.3	144863.9
	Feature 4 (Euclidean)	1229.6	2652.3
	Feature 5 (Manhattan)	68607.9	144781.9
	Feature 5 (Euclidean)	1230.2	2651.5
	Feature 6 (Manhattan)	68553.3	144786.2
	Feature 6 (Euclidean)	1229.2	2652.1
	Feature 7 (Manhattan)	68583.7	144822.9
	Feature 7 (Euclidean)	1230.2	2652.1
5 Features	Feature 1 (Manhattan)	163348.9	226011.1
	Feature 1 (Euclidean)	2903.3	4050.8
	Feature 2 (Manhattan)	164439.3	226849.3
	Feature 2 (Euclidean)	2911.7	4058.5
	Feature 3 (Manhattan)	162656.7	225360.0
	Feature 3 (Euclidean)	2882.2	4033.2
	Feature 4 (Manhattan)	162895.1	225658.3
	Feature 4 (Euclidean)	2890.9	4041.3
	Feature 5 (Manhattan)	164444.7	226844.0
	Feature 5 (Euclidean)	2923.0	4064.9
3 Features	Feature 1 (Manhattan)	201609.0	242507.6
	Feature 1 (Euclidean)	3576.8	4285.2
	Feature 2 (Manhattan)	200214.3	241260.8
	Feature 2 (Euclidean)	3563.9	4273.4
	Feature 3 (Manhattan)	200302.2	241609.6
	Feature 3 (Euclidean)	3561.7	4273.0

for image datasets. In their methodologies, various traditional augmentation methods (e.g., rotation and resize) and machine learning models have been used to generate augmented image datasets. Ekbatani *et al.* [69] generated synthetic images, including pedestrians and objects on a load, using deep learning algorithms. Among various deep learning models, convolutional neural networks (CNNs) have been applied for image generation. Norgaard *et al.* [70] applied supervised learning deep learning algorithms to generate synthetic sensor datasets. GAN algorithms with supervised learning characteristics were used in their research. The effectiveness of the proposed framework was validated using a human activity dataset with similar time-series characteristics. Chen *et al.* [71] proposed a deep learning framework for artificial CT image generation.

U-net, which is constructed using a symmetric convolutional neural network, was applied to the generated image datasets. The proposed framework was developed and evaluated using a Cone-beam computed tomography (CBCT) image dataset. Based on studies, including those mentioned above, we concluded that our research topic was well founded.

Among the various algorithms used to generate synthetic datasets, we applied GAN models to generate a behavior experiment dataset with a tabular shape. Many researchers used GAN algorithms to generate structured datasets, including tabular datasets, and complement insufficiency within datasets.

Zhou *et al.* [72] used GAN algorithms to efficiently deal with an imbalanced dataset. To complement the imbalance in the dataset, they improved the framework using two methods. First, generated artificial samples were added to the minority class to optimize the loss function of the algorithms.

Second, a fully connected network module was utilized to improve the performance of the framework. After developing the two methods, we evaluated the framework using two open-source structured datasets. First, the Alibaba-MIFD dataset was tested using trained models. This dataset was composed of 69 variables. including medical information about people (e.g., cost of medicine and time of hospital stay). Second, the JD-RPLI dataset was used to evaluate the proposed framework. Information about the user (e.g., login time and user account) was contained in this dataset. Both the

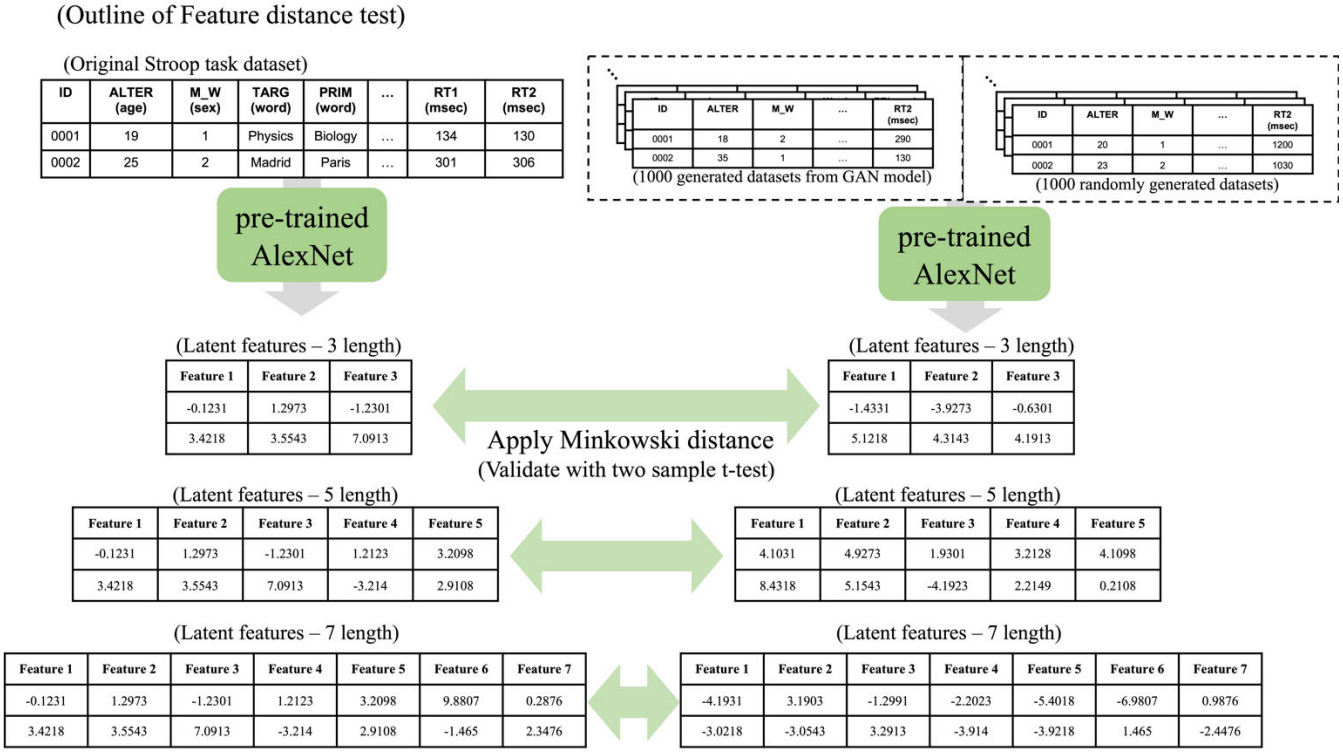

FIGURE 7. Outline of feature distance test.

aforementioned datasets are tabular-type, and the proposed framework performed better than the other algorithms in these datasets. From this previous study, we confirmed that GAN algorithms have the potential to handle tabular datasets in addition to their usefulness in the analysis of an imbalanced dataset with a shortage of class samples.

Yan *et al.* [73] used GAN algorithms to generate synthetic electronic health record (EHR) datasets for sharing datasets between research groups. To address the challenges of the complexity of EHR dataset in GAN-based generative models, several modules were added to the framework. First, the penalization module was applied to the learning process of the algorithms. Based on this module, violated values for the original values in the dataset were removed from the generated datasets. Second, the modified generator and discriminator models in the framework influenced both efficient model training and the generation of artificial instances. Third, several methods for generating datasets have been proposed for performance evaluation. The EHR dataset used in this study was composed of several variables (e.g., ICD code, BMI index, and blood pressure) with a tabular structure. The authors evaluated generation performance by comparing the distribution and statistical characteristics (e.g., correlation coefficient and Bernoulli success probabilities) of the original and generated datasets.

Xu and Veeramachaneni [74] suggested a GAN model to generate tabular datasets based on medical and educational datasets. In their framework, both discrete and continuous variables are considered in the training algorithms. Researchers have used two methods to handle variables to improve generation performance. First, in the case of numerical variables, multimodal distributions were normalized to improve the processing datasets. Normalized numerical variables showed values in the range of -1 to $+1$. Second, the softmax function was used to smooth the distributions of the categorical variables. In addition, one-hot encoding was applied to categorical variables. A total of three open-source datasets with tabular structures were used to evaluate the performance of the framework. Moreover, the synthetic datasets were evaluated using the original dataset through machine learning classifiers.

Based on the aforementioned studies, we considered steps similar to the experimental design in our study. First, the statistical characteristics of the dataset were confirmed before developing the generation framework to handle the challenges related to training algorithms. Second, the GAN algorithm was trained using a tabular dataset. Third, evaluation methods were proposed for generation performance using statistics from the generated dataset and the application of machine learning tasks.

However, many researchers using the GAN model in their work have pointed out the ambiguity of evaluation methods for generated datasets [75]–[77]. Because of the components of the algorithms (i.e., GAN models consist of generator and discriminator models), we cannot consider it the gold standard of evaluation, unlike other algorithms [75]. For this reason, in the case of image generation tasks, the generated dataset was evaluated based on human

TABLE 12. Results of feature distance test in stroop 2 sub-dataset.

Dataset / length of features	Features (distance metric)	GAN-based model	Random generation
7 Features	Feature 1 (Manhattan)	457062.2	688565.3
	Feature 1 (Euclidean)	6106.6	9365.6
	Feature 2 (Manhattan)	451071.9	684089.2
	Feature 2 (Euclidean)	6027.6	9313.2
	Feature 3 (Manhattan)	455763.6	688323.5
	Feature 3 (Euclidean)	6114.9	9375.8
	Feature 4 (Manhattan)	455875.4	688011.0
	Feature 4 (Euclidean)	6105.5	9366.5
	Feature 5 (Manhattan)	453993.2	686413.4
	Feature 5 (Euclidean)	6066.2	9340.7
	Feature 6 (Manhattan)	454476.9	686641.3
	Feature 6 (Euclidean)	6081.1	9346.6
	Feature 7 (Manhattan)	456790.1	688415.3
	Feature 7 (Euclidean)	6124.1	9374.0
5 Features	Feature 1 (Manhattan)	383847.7	580280.4
	Feature 1 (Euclidean)	5180.3	7912.9
	Feature 2 (Manhattan)	383484.9	579934.0
	Feature 2 (Euclidean)	5159.7	7904.4
	Feature 3 (Manhattan)	382080.4	578987.1
	Feature 3 (Euclidean)	5156.3	7899.2
	Feature 4 (Manhattan)	383619.3	579833.5
	Feature 4 (Euclidean)	5168.2	7905.7
	Feature 5 (Manhattan)	383890.8	580210.8
	Feature 5 (Euclidean)	5170.4	7908.6
3 Features	Feature 1 (Manhattan)	372750.5	492835.9
	Feature 1 (Euclidean)	5036.6	6602.7
	Feature 2 (Manhattan)	374157.9	493850.2
	Feature 2 (Euclidean)	5066.8	6624.7
	Feature 3 (Manhattan)	377934.7	496591.8
	Feature 3 (Euclidean)	5110.6	6656.1

TABLE 13. Statistical test results of feature distance test in stroop 1 sub-dataset.

Comparison condition (feature)	t statistics (Man[a]/Euc[b])	*p*-value (one-tail) (Man/Euc)	*p*-value (two-tail) (Man/Euc)
7 length of features			
GAN-based vs. Random (Feature 1)	-52.10 / -52.80	2.40E-28 / 6.75E-29	4.80E-28 / 1.35E-29
GAN-based vs. Random (Feature 2)	-52.10 / -52.90	2.72E-28 / 6.10E-29	5.44E-28 / 1.22E-29
GAN-based vs. Random (Feature 3)	-52.00 / -52.80	6.67E-27 / 1.25E-21	1.33E-26 / 2.49E-21
GAN-based vs. Random (Feature 4)	-52.00 / -52.80	5.02E-08 / 1.26E-29	1.00E-28 / 2.53E-29
GAN-based vs. Random (Feature 5)	-52.00 / -52.80	5.03E-27 / 9.80E-22	1.01E-26 / 1.96E-21
GAN-based vs. Random (Feature 6)	-52.10 / -52.90	2.52E-28 / 2.17E-22	5.04E-28 / 4.34E-22
GAN-based vs. Random (Feature 7)	-52.10 / -52.08	4.00E-28 / 1.16E-21	8.00E-28 / 2.32E-21
5 length of features			
GAN-based vs. Random (Feature 1)	-52.62 / -52.56	2.49E-20 / 6.71E-20	4.98E-20 / 1.34E-20
GAN-based vs. Random (Feature 2)	-52.18 / -52.34	1.40E-27 / 1.36E-28	2.79E-27 / 2.72E-28
GAN-based vs. Random (Feature 3)	-52.44 / -52.37	3.51E-28 / 9.72E-28	7.02E-28 / 1.94E-28
GAN-based vs. Random (Feature 4)	-52.55 / -52.48	6.99E-20 / 2.00E-28	1.40E-20 / 4.00E-28
GAN-based vs. Random (Feature 5)	-52.21 / -52.28	8.29E-28 / 3.26E-28	1.66E-28 / 6.62E-28
3 length of features			
GAN-based vs. Random (ALTER)	-481.81 / -495.10	1.20E-09 / 3.21E-08	2.54E-08 / 4.90E-06
GAN-based vs. Random (RT1)	-490.13 / -500.78	1.90E-05 / 9.87E-06	3.39E-06 / 8.77E-07
GAN-based vs. Random (RT2)	-487.71 / -494.30	2.23E-20 / 8.90E-18	4.59E-21 / 7.10E-17

[a]Man: Manhattan distance; [b]Euc: Euclidean distance

decisions and latent features extracted by pre-trained deep learning models [78]–[80]. In addition, researchers have evaluated performance with relative comparisons between methodologies [81], [82].

To evaluate the generated tabular datasets more precisely, we conducted an evaluation using the internal levels of the table structure. First, we evaluated the generated dataset in terms of instances and row levels. The existence of overlaps in the generated dataset was verified using the original dataset. To identify important elements in the Stroop task dataset, the values of word-related variable (PRIM and TARG) and reaction time variable (RT1 and RT2) pairs were compared. Furthermore, a one-sample t-test was used to verify the differences in variable values between the original and generated datasets. In this test, we considered that higher ratios of statistically significant results indicated fewer overlaps with values from the original dataset. Most of the significant result ratios in the dataset generated by the GAN-based model were higher than those of the randomly generated dataset.

In addition, the ranges of values for the variables were compared with the original values. We calculated the minimum, median, and maximum values as constraints of continuous variables. The absolute differences between the original and generated constraints were compared. The difference values of the GAN-based model conditions were generally lower than those of the random generation condition. From these results, we confirmed that the ranges for the GAN-based model conditions were closer to the range of the actual data than the random generation conditions.

Second, the variable and feature level characteristics of the dataset were evaluated. Correlations between the variables were confirmed from the generated datasets. We calculated the absolute difference between the original and generated values. We found that the absolute values of the GAN-based model conditions were lower than those of the random generation conditions. In addition, the rank of the correlation coefficients for the GAN-based model condition was compared with that of the original dataset. We found that several common elements were included in the coefficient list.

Additionally, the distances between the distributions were evaluated for the datasets. For categorical variables, the Hamming distances were calculated. Two-sample KS tests were

TABLE 14. Statistical test results of feature distance test in stroop 2 sub-dataset.

Comparison condition (feature)	t statistics (Man[a]/Euc[b])	*p*-value (one-tail) (Man/Euc)	*p*-value (two-tail) (Man/Euc)
7 length of features			
GAN-based vs. Random (Feature 1)	-103.45 / -102.78	2.90E-08 / 5.30E-07	9.01E-10 / 4.02E-08
GAN-based vs. Random (Feature 2)	-103.50 / -102.84	3.45E-10 / 1.46E-11	1.21E-11 / 5.21E-10
GAN-based vs. Random (Feature 3)	-103.82 / -102.98	2.23E-08 / 3.34E-08	3.39E-08 / 9.75E-07
GAN-based vs. Random (Feature 4)	-103.67 / -102.82	4.58E-05 / 3.09E-06	1.92E-06 / 3.14E-08
GAN-based vs. Random (Feature 5)	-103.28 / -102.77	1.19E-10 / 6.43E-11	2.84E-09 / 5.98E-10
GAN-based vs. Random (Feature 6)	-103.70 / -103.05	2.95E-08 / 5.23E-07	1.99E-10 / 8.43E-08
GAN-based vs. Random (Feature 7)	-103.37 / -102.63	4.35E-07 / 7.98E-06	7.31E-08 / 9.03E-08
5 length of features			
GAN-based vs. Random (Feature 1)	-85.84 / -91.33	2.34E-05 / 5.46E-10	4.32E-06 / 8.67E-11
GAN-based vs. Random (Feature 2)	-86.00 / -91.48	3.90E-09 / 3.49E-06	2.74E-11 / 6.65E-07
GAN-based vs. Random (Feature 3)	-86.13 / -91.40	3.28E-15 / 3.04E-11	6.98E-14 / 4.09E-12
GAN-based vs. Random (Feature 4)	-86.26 / -91.54	9.15E-04 / 3.29E-11	9.75E-05 / 2.63E-12
GAN-based vs. Random (Feature 5)	-86.07 / -91.27	5.93E-13 / 4.32E-08	4.00E-14 / 4.20E-09
3 length of features			
GAN-based vs. Random (ALTER)	-944.83 / -926.32	2.10E-13 / 6.31E-24	1.13E-16 / 6.09E-13
GAN-based vs. Random (RT1)	-954.94 / -926.50	5.43E-08 / 2.10E-21	4.98E-07 / 1.10E-21
GAN-based vs. Random (RT2)	-973.60 / -944.43	3.20E-05 / 9.52E-04	3.49E-05 / 8.93E-05

[a]Man: Manhattan distance; [b]Euc: Euclidean distance

conducted for the continuous variables. In the GAN-based model conditions, the absolute differences in distance values were lower than those in random generation conditions. From these results, we confirmed that the distributions of variables generated by the GAN-based models were more similar to the original distribution than the randomly generated datasets.

Finally, the generated datasets were evaluated in terms of whole-set levels. A total of three lengths (seven, five, and three length features) of latent features were extracted using a pre-trained AlexNet model. To compare the distances between the extracted latent features, the Minkowski distance, including two distance metrics (Euclidean and Manhattan distance), was applied. Overall distance values in the GAN-based model conditions were lower than those in the random generation.

Based on the aforementioned experimental results, we concluded that the statistical characteristics (e.g., similarity or distances) of the generated datasets from the proposed frameworks are closer to the original datasets than the characteristics of randomly generated datasets.

V. CONCLUSION

In this study, we proposed a data generation framework based on a deep learning model for a behavior experiment dataset collected from cognitive psychology research. Based on previous studies associated with tabular data generation, we designed experiments using the development of algorithms and evaluation methods. To complement the relatively small sample size dataset used in our study, we used a pre-trained GAN model for the framework. Furthermore, five evaluation methods with internal tabular structure levels were applied for a more detailed evaluation. In addition, a random generation method was compared with the proposed framework to evaluate its generation performance. Based on the experimental results, we confirmed that the proposed framework with GAN algorithms can generate statistically similar synthetic datasets with the statistical characteristics of the original dataset.

The first strength of this study is the application of a behavior experiment dataset with a tabular structure from cognitive psychology to a deep learning generation algorithm. Second, we propose novel evaluation methods based on the tabular structure levels. Third, we consider not only the generation of structural characteristics, but also the reflection of the statistical characteristics of the original dataset.

Furthermore, the proposed framework has advantages in terms of data analysis and related research. First, the generated datasets can help reduce the sample size in related experimental studies. Second, the synthetic dataset generation framework can overcome environmental restrictions (e.g., the Covid-19 pandemic) in conducting experimental research. Finally, a complement based on an artificial dataset with similar statistical characteristics can reduce the burden on participants.

Our study has some limitations. First, we compare only random generation methods to evaluate our framework. However, we do evaluate the diverse aspects of the generated dataset, which was advanced from previous studies. Second, we only apply a Stroop task dataset from various task designs in cognitive psychology research. Although a Stroop task is one of the most established task designs used in related studies, future studies should consider other tasks to generalize this framework. In addition, for the utility of synthetic datasets, validation studies need to be considered in future studies.

REFERENCES

[1] D. J. Biau, S. Kernéis, and R. Porcher, "Statistics in brief: The importance of sample size in the planning and interpretation of medical research," *Clin. Orthopaedics Rel. Res.*, vol. 466, no. 9, pp. 2282–2288, Sep. 2008.

[2] T. W. Beck, "The importance of a priori sample size estimation in strength and conditioning research," *J. Strength Conditioning Res.*, vol. 27, no. 8, pp. 2323–2337, 2013.

[3] I. Balki, A. Amirabadi, J. Levman, A. L. Martel, Z. Emersic, B. Meden, A. Garcia-Pedrero, S. C. Ramirez, D. Kong, A. R. Moody, and P. N. Tyrrell, "Sample-size determination methodologies for machine learning in medical imaging research: A systematic review," *Can. Assoc. Radiol. J.*, vol. 70, no. 4, pp. 344–353, Nov. 2019.

[4] D. R. B. Stockwell and A. T. Peterson, "Effects of sample size on accuracy of species distribution models," *Ecol. Model.*, vol. 148, no. 1, pp. 1–13, Feb. 2002.

[5] A. K. Akobeng, "Understanding type I and type II errors, statistical power and sample size," *Acta Paediatrica*, vol. 105, no. 6, pp. 605–609, Jun. 2016.
[6] M. Columb and M. Atkinson, "Statistical analysis: Sample size and power estimations," *BJA Educ.*, vol. 16, no. 5, pp. 159–161, May 2016.
[7] J. M. Lachin, "Introduction to sample size determination and power analysis for clinical trials," *Controlled Clin. Trials*, vol. 2, no. 2, pp. 93–113, Jun. 1981.
[8] R. C. MacCallum, W. B. Michael, and M. S. Hazuki, "Power analysis and determination of sample size for covariance structure modeling," *Psychol. Methods*, vol. 1, no. 2, p. 130, 1996.
[9] P. Kadam and B. Supriya, "Sample size calculation," *Int. J. Ayurveda Res.*, vol. 1, no. 1, p. 55, 2010.
[10] F. Y. Hsieh, D. A. Bloch, and M. D. Larsen, "A simple method of sample size calculation for linear and logistic regression," *Statist. Med.*, vol. 17, no. 14, pp. 1623–1634, Jul. 1998.
[11] M. R. D. Águila and A. R. González-Ramírez, "Sample size calculation," *Allergol. Immunopathol.*, vol. 42, pp. 485–492, Sep. 2014.
[12] A. R. Ajiboye, R. Abdullah-Arshah, H. Qin, and H. Isah-Kebbe, "Evaluating the effect of dataset size on predictive model using supervised learning technique," *Int. J. Comput. Syst. Softw. Eng.*, vol. 1, no. 1, pp. 75–84, Feb. 2015.
[13] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 843–852.
[14] S. F. Schulten, R. J. Detollenaere, J. Stekelenburg, J. IntHout, K. B. Kluivers, and H. W. van Eijndhoven, "Sacrospinous hysteropexy versus vaginal hysterectomy with uterosacral ligament suspension in women with uterine prolapse stage 2 or higher: Observational follow-up of a multicentre randomised trial," *BMJ*, vol. 366, pp. 1–10, Sep. 2019.
[15] E. A. Mair, A. H. Park, D. Don, J. Koempel, M. Bear, and C. LeBel, "Safety and efficacy of intratympanic ciprofloxacin otic suspension in children with middle ear effusion undergoing tympanostomy tube placement: Two randomized clinical trials," *JAMA Otolaryngol.-Head Neck Surg.*, vol. 142, no. 5, pp. 444–451, 2016.
[16] H. Hiscock, E. Sciberras, F. Mensah, B. Gerner, D. Efron, S. Khano, and F. Oberklaid, "Impact of a behavioural sleep intervention on symptoms and sleep in children with attention deficit hyperactivity disorder, and parental mental health: Randomised controlled trial," *BMJ*, vol. 350, pp. 1–14, Jan. 2015.
[17] C. M. Koolhaas, D. Kocevska, B. H. W. te Lindert, N. S. Erler, O. H. Franco, A. I. Luik, and H. Tiemeier, "Objectively measured sleep and body mass index: A prospective bidirectional study in middle-aged and older adults," *Sleep Med.*, vol. 57, pp. 43–50, May 2019.
[18] J. Brick and G. Kalton, "Handling missing data in survey research," *Stat. Methods Med. Res.*, vol. 5, no. 3, pp. 215–238, Sep. 1996.
[19] G. L. Schlomer, S. Bauman, and N. A. Card, "Best practices for missing data management in counseling psychology," *J. Counseling Psychol.*, vol. 57, no. 1, pp. 1–10, 2010.
[20] M. Angerer, M. Schabus, M. Raml, G. Pichler, A. B. Kunz, M. Scarpatetti, and C. Blume, "Actigraphy in brain-injured patients—A valid measurement for assessing circadian rhythms?" *BMC Med.*, vol. 18, pp. 1–10, Dec. 2020.
[21] S. Sinharay, S. S. Hal, and D. Russell, "The use of multiple imputation for the analysis of missing data," *Psychol. Methods*, vol. 6, no. 4, p. 317, 2001.
[22] M. Vriens and E. Melton, "Managing missing data," *Marketing Res.*, vol. 14, no. 3, p. 12, 2002.
[23] D. A. Bennett, "How can I deal with missing data in my study?" *Austral. New Zealand J. Public Health*, vol. 25, no. 5, pp. 464–469, Oct. 2001.
[24] M. Miyakawa, "Analysis of incomplete data in competing risks model," *IEEE Trans. Rel.*, vol. R-33, no. 4, pp. 293–296, Oct. 1984.
[25] Z. Zhang, "Missing data imputation: Focusing on single imputation," *Ann. Transl. Med.*, vol. 4, no. 1, pp. 1–8, 2016.
[26] A. Plaia and A. Bondi, "Single imputation method of missing values in environmental pollution data sets," *Atmos. Environ.*, vol. 40, no. 38, pp. 7316–7330, Dec. 2006.
[27] J. L. Schafer and M. K. Olsen, "Multiple imputation for multivariate missing-data problems: A data analyst's perspective," *Multivariate Behav. Res.*, vol. 33, no. 4, pp. 545–571, Oct. 1998.
[28] P. D. Allison, "Multiple imputation for missing data: A cautionary tale," *Sociol. Methods Res.*, vol. 28, no. 3, pp. 301–309, Feb. 2000.
[29] B. Efron, "Missing data, imputation, and the bootstrap," *J. Amer. Stat. Assoc.*, vol. 89, no. 426, pp. 463–475, 1994.
[30] M. Saqib, Y. Sha, and M. D. Wang, "Early prediction of sepsis in EMR records using traditional ML techniques and deep learning LSTM networks," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 4038–4041.
[31] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621*. [Online]. Available: http://arxiv.org/abs/1712.04621
[32] D. T. Lykken, "Statistical significance in psychological research," *Psychol. Bull.*, vol. 70, no. 3, p. 151, 1968.
[33] G. Francis, "The psychology of replication and replication in psychology," *Perspect. Psychol. Sci.*, vol. 7, no. 6, pp. 585–594, Nov. 2012.
[34] C. R. Peterson, R. J. Schneider, and A. J. Miller, "Sample size and the revision of subjective probabilities," *J. Exp. Psychol.*, vol. 69, no. 5, p. 522, 1965.
[35] M. Schaller, "Sample size, aggregation, and statistical reasoning in social inference," *J. Exp. Social Psychol.*, vol. 28, no. 1, pp. 65–85, Jan. 1992.
[36] G. Schweizer and P. Furley, "Reproducible research in sport and exercise psychology: The role of sample sizes," *Psychol. Sport Exerc.*, vol. 23, pp. 114–122, Mar. 2016.
[37] K. Sassenberg and L. Ditrich, "Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies," *Adv. Methods Practices Psychol. Sci.*, vol. 2, no. 2, pp. 107–114, Jun. 2019.
[38] J. C. Peterson, D. D. Bourgin, M. Agrawal, D. Reichman, and T. L. Griffiths, "Using large-scale experiments and machine learning to discover theories of human decision-making," *Science*, vol. 372, no. 6547, pp. 1209–1214, Jun. 2021.
[39] D. Kai-Ineman and A. Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica*, vol. 47, no. 2, pp. 363–391, 1979.
[40] W. Edwards, "The theory of decision making," *Psychol. Bull.*, vol. 51, no. 4, p. 380, 1954.
[41] M. Agrawal, J. C. Peterson, and T. L. Griffiths, "Scaling up psychology via scientific regret minimization: A case study in moral decisions," 2019, *arXiv:1910.07581*. [Online]. Available: http://arxiv.org/abs/1910.07581
[42] S. Timmons and R. M. Byrne, "Moral fatigue: The effects of cognitive fatigue on moral reasoning," *Quart. J. Exp. Psychol.*, vol. 72, no. 4, pp. 943–954, Apr. 2019.
[43] P. L. Ackerman and R. Kanfer, "Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions," *J. Exp. Psychol., Appl.*, vol. 15, no. 2, pp. 163–181, 2009.
[44] C. Gentili and I. A. Cristea, "Challenges and opportunities for human behavior research in the coronavirus disease (COVID-19) pandemic," *Frontiers Psychol.*, vol. 11, no. 1786, pp. 1–4, Jul. 2020.
[45] J. H. Cheung, D. K. Burns, R. R. Sinclair, and M. Sliter, "Amazon mechanical Turk in organizational psychology: An evaluation and practical recommendations," *J. Bus. Psychol.*, vol. 32, no. 4, pp. 347–361, Aug. 2017.
[46] J. C. de Winter, S. D. Gosling, and J. Potter, "Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data," *Psychol. Methods*, vol. 21, no. 3, pp. 273–290, 2016.
[47] J. Z. Bakdash and R. M. Laura, "Repeated measures correlation," *Frontiers Psychol.*, vol. 8, p. 456, Apr. 2017.
[48] C. M. Judd and D. A. Kenny, "Data analysis in social psychology: Recent and recurring issues," in *Handbook of Social Psychology*, vol. 1. New York, NY, USA: Springer, 2010, pp. 115–139.
[49] M. W.-L. Cheung and S. Jak, "Analyzing big data in psychology: A split/analyze/meta-analyze approach," *Frontiers Psychol.*, vol. 7, pp. 447–457, May 2016.
[50] L. Doey and J. Kurta, "Correspondence analysis applied to psychological research," *Tuts. Quant. Methods Psychol.*, vol. 7, no. 1, pp. 5–14, Apr. 2011.
[51] E. Lashgari, D. Liang, and U. Maoz, "Data augmentation for deep-learning-based electroencephalography," *J. Neurosci. Methods*, vol. 346, pp. 1–26, Jul. 2020.
[52] J.-H. Jang, J. Choi, H. W. Roh, S. J. Son, C. H. Hong, E. Y. Kim, T. Y. Kim, and D. Yoon, "Deep learning approach for imputation of missing values in actigraphy data: Algorithm development study," *JMIR mHealth uHealth*, vol. 8, no. 7, Jul. 2020, Art. no. e16113.
[53] G. Rizos, K. Hemker, and B. Schuller, "Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 991–1000.
[54] S. Semeniuta, A. Severyn, and E. Barth, "A hybrid convolutional variational autoencoder for text generation," 2017, *arXiv:1702.02390*. [Online]. Available: http://arxiv.org/abs/1702.02390

[55] J. Guan, R. Li, S. Yu, and X. Zhang, "A method for generating synthetic electronic medical record text," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 1, pp. 173–182, Feb. 2021.
[56] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," 2019, *arXiv:1907.00503*. [Online]. Available: http://arxiv.org/abs/1907.00503
[57] K. Rothermund and D. Wentura, "Affective congruency effects in the Stroop task: Primary data and control programs, version 1.0.0," Center Res. Data Psychol., PsychData Leibniz Inst. Psychol. ZPID, Trier, Germany, 2004, doi: 10.5160.psychdata.rdks97af99.
[58] K. Rothermund and D. Wentura, "Ein fairer Test für die Aktivationsausbreitungs-hypothese: Untersuchung affektiver Kongruenzeffekte in der Stroop-Aufgabe," *Zeitschrift Experimentelle Psychologie*, vol. 45, no. 2, pp. 120–135, 1998.
[59] E. Tulving and D. L. Schacter, "Priming and human memory systems," *Science*, vol. 247, no. 4940, pp. 301–306, Jan. 1990.
[60] M. Norouzi, D. J. Fleet, and R. R. Salakhutdinov, "Hamming distance metric learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1061–1069.
[61] D. R. Barr and T. Davidson, "A Kolmogorov–Smirnov test for censored samples," *Technometrics*, vol. 15, no. 4, pp. 739–757, Nov. 1973.
[62] A. Singh, A. Yadav, and A. Rana, "K-means with three different distance metrics," *Int. J. Comput. Appl.*, vol. 67, no. 10, pp. 13–17, 2013.
[63] R. P. Pargas, M. J. Harrold, and R. R. Peck, "Test-data generation using genetic algorithms," *Softw. Test., Verification Rel.*, vol. 9, no. 4, pp. 263–282, 1999.
[64] N. Tracey, J. Clark, K. Mander, and J. McDermid, "An automated framework for structural test-data generation," in *Proc. 13th IEEE Int. Conf. Automated Softw. Eng.*, Oct. 1998, pp. 285–288.
[65] F. P. Brissette, M. Khalili, and R. Leconte, "Efficient stochastic generation of multi-site synthetic precipitation data," *J. Hydrol.*, vol. 345, nos. 3–4, pp. 121–133, Oct. 2007.
[66] D. T. Jones, W. R. Taylor, and J. M. Thornton, "The rapid generation of mutation data matrices from protein sequences," *Bioinformatics*, vol. 8, no. 3, pp. 275–282, 1992.
[67] H. Guo and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 30–39, 2004.
[68] M. D. Bloice, C. Stocker, and A. Holzinger, "Augmentor: An image augmentation library for machine learning," 2017, *arXiv:1708.04680*. [Online]. Available: http://arxiv.org/abs/1708.04680
[69] H. K. Ekbatani, O. Pujol, and S. Segui, "Synthetic data generation for deep learning in counting pedestrians," in *Proc. 6th Int. Conf. Pattern Recognit. Appl. Methods*, 2017, pp. 318–323.
[70] S. Norgaard, R. Saeedi, K. Sasani, and A. H. Gebremedhin, "Synthetic sensor data generation for health applications: A supervised deep learning approach," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 1164–1167.
[71] L. Chen, X. Liang, C. Shen, S. Jiang, and J. Wang, "Synthetic CT generation from CBCT images via deep learning," *Med. Phys.*, vol. 47, no. 3, pp. 1115–1125, Mar. 2020.
[72] T. Zhou, W. Liu, C. Zhou, and L. Chen, "GAN-based semi-supervised for imbalanced data classification," in *Proc. 4th Int. Conf. Inf. Manage. (ICIM)*, May 2018, pp. 17–21.
[73] C. Yan, Z. Zhang, S. Nyemba, and B. A. Malin, "Generating electronic health records with multiple data types and constraints," in *Proc. Annu. Symp. Amer. Med. Informat. Assoc. (AMIA)*, 2020, p. 1335.
[74] L. Xu and K. Veeramachaneni, "Synthesizing tabular data using generative adversarial networks," 2018, *arXiv:1811.11264*. [Online]. Available: http://arxiv.org/abs/1811.11264
[75] A. Borji, "Pros and cons of GAN evaluation measures," *Comput. Vis. Image Understand.*, vol. 179, pp. 41–65, Feb. 2019.
[76] K. Shmelkov, C. Schmid, and K. Alahari, "How good is my GAN?" in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 213–229.
[77] C. Qiao, D. Li, Y. Guo, C. Liu, T. Jiang, Q. Dai, and D. Li, "Evaluation and development of deep neural networks for image super-resolution in optical microscopy," *Nature Methods*, vol. 18, no. 2, pp. 194–202, Feb. 2021.
[78] P. Andreini, S. Bonechi, M. Bianchini, A. Mecocci, and F. Scarselli, "Image generation by GAN and style transfer for agar plate image segmentation," *Comput. Methods Programs Biomed.*, vol. 184, pp. 1–13, Feb. 2020.
[79] C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri, and H. Nakayama, "GAN-based synthetic brain MR image generation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 734–738.
[80] H. Tang, S. Bai, L. Zhang, P. H. Torr, and N. Sebe, "XingGAN for person image generation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 717–734.
[81] H. Emami, M. M. Aliabadi, M. Dong, and R. B. Chinnam, "SPA-GAN: Spatial attention GAN for image-to-image translation," *IEEE Trans. Multimedia*, vol. 23, pp. 391–401, 2021.
[82] M. Zhai, L. Chen, F. Tung, J. He, M. Nawhal, and G. Mori, "Lifelong GAN: Continual learning for conditional image generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2759–2768.

JUNG-GU CHOI was born in 1993. He is currently pursuing Ph.D. degree in Cognitive science through the Applied Brain Cognitive Laboratory of the Yonsei University graduate program. His research areas include brain science and data mining.

YOONJIN NAH is currently a Postgraduate Researcher with the Department of Psychology, Yonsei University. He has experience in decoding differential cognitive states and clinical groups with machine learning algorithms using fMRI connectivity data.

INHWAN KO is currently pursuing the M.S. degree in psychology with Yonsei University. He has a diverse work background, from a human resource management practitioner in multinational companies to a certified public labor attorney. His research interests include converging human resource management and cognitive neuroscience.

SANGHOON HAN was born in 1977. He is currently a Professor with the Department of Psychology, Yonsei University. His research interests include decision making and cognitive science.