applied sciences

MDPI

Article

Applying a Method for Augmenting Data Mixed from Two Different Sources Using Deep Generative Neural Networks to Management Science

Shinya Watanuki [1,*], Yumiko Nomura [2], Yuki Kiyota [3], Minami Kubo [4], Kenji Fujimoto [5], Junko Okada [6] and Katsue Edo [7]

[1] Department of Marketing, Faculty of Commerce, University of Marketing and Distribution Sciences, 3-1 Gakuen-Nishimachi, Nishi-ku, Kobe 651-2188, Hyogo, Japan
[2] Department of Nursing, Japanese Red Cross Hiroshima College of Nursing, 1-2 Ajinadai-Higashi, Hatsukaichi 738-0052, Hiroshima, Japan; yn11199@jrchcn.ac.jp
[3] Graduate School of Comprehensive Scientific Research Program in Health and Welfare, Prefectural University of Hiroshima, 1-1 Gakuencho, Mihara 723-0053, Hiroshima, Japan; ta.n.k.yu.05@gmail.com
[4] Hiroshima Red Cross Hospital Atomic-Bomb Survivors Hospital, 1-9-6 Sendamachi, Naka-ku, Hiroshima 730-8619, Hiroshima, Japan; kmnm.3159@gmail.com
[5] Hiroshima Office, Survey Research Center Co., Ltd., 2-29 Tatemachi, Naka-ku, Hiroshima 730-0032, Hiroshima, Japan; fujimo_k@surece.co.jp
[6] Faculty of Health and Welfare, Prefectural University of Hiroshima, 1-1 Gakuencho, Mihara 723-0053, Hiroshima, Japan; ojunko@pu-hiroshima.ac.jp
[7] Hiroshima Business and Management School, Prefectural University of Hiroshima, 1-1-71 Ujina-Higashi, Minami-ku, Hiroshima 734-8558, Hiroshima, Japan; edocats@gmail.com
* Correspondence: shinya_watanuki@red.umds.ac.jp

Abstract: Although a multimodal data analysis, comprising physiological and questionnaire survey data, provides better insights into addressing management science concerns, such as challenging the predictions of consumer choice behavior, studies in this field are scarce because of two obstacles: limited sample size and information privacy. This study addresses these challenges by synthesizing multimodal data using deep generative models. We obtained multimodal data by conducting an electroencephalography (EEG) experiment and a questionnaire survey on the prediction of skilled nurses. Subsequently, we validated the effectiveness of the synthesized data compared with real data regarding the similarities between these data and the predictive performance. We confirmed that the synthesized big data were almost equal to the real data using the trained models through sufficient epochs. Conclusively, we demonstrated that synthesizing data using deep generative models might overcome two significant concerns regarding multimodal data utilization, including physiological data. Our approach can contribute to the prevailing combined big data from different modalities, such as physiological and questionnaire survey data, when solving management issues.

Keywords: questionnaire survey; marketing research; EEG; deep learning; generative adversarial networks (GANs)

Citation: Watanuki, S.; Nomura, Y.; Kiyota, Y.; Kubo, M.; Fujimoto, K.; Okada, J.; Edo, K. Applying a Method for Augmenting Data Mixed from Two Different Sources Using Deep Generative Neural Networks to Management Science. *Appl. Sci.* **2024**, *14*, 378. https://doi.org/10.3390/app14010378

Academic Editor: Chihhsuan Wang

Received: 28 October 2023
Revised: 27 December 2023
Accepted: 29 December 2023
Published: 31 December 2023

1. Introduction

In multiple cases [1,2], prediction performance may be enhanced using multimodal data when predicting a diagnosis and decisions. The effectiveness of combining physiological and questionnaire survey data has been reported. For example, Hakim et al. (2020) [3] stated that the accuracy of predicting purchase behaviors was enhanced by combining physiological data from unconscious mind sources with asking-based data from a questionnaire survey of conscious mind sources. Thus, analyzing different modality-sourced data could improve predictive performance and determine participants' insights. Despite their effectiveness, management science databases incorporating physiological and questionnaire

survey data are yet to be developed. The cause of this condition may be due to two significant variables. The subject of a limited sample is among the primary reasons. Furthermore, the small sample problem results from three aspects. First, developing a massive database is challenging because performing multiple experiments is expensive. It also requires time and effort to perform the experiments. For example, an electroencephalography (EEG) experiment requires 30 min to 1 h per participant. Second, few management science and human resource professionals have sufficient experience and knowledge to perform physiological experiments. Because of the scarcity of human resource professionals, there are fewer opportunities for performing physiological experiments in this field. Third, recruiting experimental participants from a small population is challenging. Because data sampled from rare groups provide better insights for businesses, methods of recruiting from rare groups have proven controversial in management science [4,5]. The other primary reason is permission to apply the data. In return for disguising their personal information, participants agree to participate in an experiment and utilize data for the initial research objectives and additional purposes; however, in particular, the additional purposes are challenging. Notably, regarding commercial applications, most participants might decline to permit the use of experimental data despite masking their personal information. Furthermore, most participants are anxious about the maltreatment of their physiological data because they include essential personal details such as disease and emotional disposition [6].

Considering these issues are underlying conditions intertwined by multiple factors, developing and using combined databases based on physiological and questionnaire survey data are challenging. Eventually, a database based on these multimodal data in the field of management science will be limited to participants. This aspect suggests that it is demanding to improve predictive performance by conducting additional experiments. Concretely, when conducting experiments on luxurious real estate and cars for persons with high income, as there are innately few of those persons, it is challenging to conduct another experiment based on the same conditions because recruiting other participants with the same background is challenging. Actually, as it costs too much to conduct experiments on those rare populations, conducting the experiments is hard because of budget limitations. Thus, addressing limited samples and privacy issues may be necessary when using combined physiological and questionnaire survey data in management science. Specifically, it may be possible to address these issues by developing a secure database with a sufficient sample size that combines physiological and questionnaire survey data.

Regarding the limited sample size problem, various data augmentation approaches have been developed [7]. Data augmentation methods based on the synthesized data using deep generative models such as generative adversarial networks (GANs) [8,9] and variational autoencoders (VAEs) [10] have been proposed. Other methods, such as mix–up [11], have also been proposed. In the former data augmentation methods, augmented data are used as synthesized data based on appropriate probability distributions related to actual data. However, regardless of the probability distributions, the latter approach involves trial-and-error based on actual data. Recently, prospective data augmentation approaches, including GAN-based synthesizers and combined approaches of generative models and mix-ups, have been presented [12,13]. Notably, because these data augmentation approaches can augment larger-sized data, they can solve limited sample issues. Moreover, to overcome concerns regarding privacy security, synthetic data-generation methods are presented using this deep generative model [14]. Generating synthetic data has an essential advantage over masking personal details. This technology can create anonymized data that inherit the characteristics of the original data while increasing the data size [14,15]. Personal information may not be revealed even if someone accesses or invades the data. However, the applications of these synthetic data-generation technologies primarily focus on a single modality [14]. It is necessary to address the application of synthetic data-generation technologies for multimodal data, such as combined physiological and asking-based data. This study uses deep generative models to address the data augmentation approach based on synthesized data.

Additionally, regardless of these necessaries and effectiveness, few studies have applied the synthetic data-generation approach to multimodal data in management science contexts. Although a few studies were reported in management science study areas, these studies deal with unimodal data, as well as the other studies on synthesizing data [16,17]. We attempt to validate the data augmentation approach by synthesizing small-combined data consisting of physiological and questionnaire survey data in a management science issue. Concretely, we address the difficulties, such as the limited sample size and privacy problems, by applying deep generative models to the combined multimodality data. In this study, we considered the effectiveness of solving these problems in identifying skilled nurses, an issue in human resource management. Details of this issue are described in the Methods section.

2. Related Work

After the experiment, the physiological and questionnaire survey data were tabulated. Accordingly, this study describes deep generative models for tabular data. Several GAN-based data-synthesis methods have been proposed. Medical GAN (medGAN) [18] is an approach focused on synthesizing discrete and continuous patient record data. The medGAN architecture is composed of a combination of an autoencoder and a GAN. Table GAN (TGAN) [19], which applies a Deep Convolutional GAN (DCGAN) [9] to tabular data, can be used to synthesize various types of data for general purposes (i.e., continuous, discrete-like, categorical-type data, and time). Conditional Tabular GAN (CTGAN) [20] is a method for modifying the TGAN to address imbalanced data derived from non-Gaussian and multimodal distributions better than the TGAN. CopulaGAN is a deep generative model modifying CTGAN to generate data containing information on dependencies between features [20,21]. The tabular variational autoencoder is a model that adjusts the VAE [10] to tabular data. Xu et al. [20] demonstrated that CTGAN was advanced, and TVAE was a contender among these representative GAN-based synthesizing data methods. Cheon et al. [22] demonstrated that CTGAN outperformed TGAN in synthesizing physiological data (EEG).

Therefore, we attempt to identify the most suitable deep generative algorithm to solve our issues, among the latest approaches to synthesizing data for tabularly formed data, CTGAN, CopulaGAN, and TVAE, and find out conditions to improve predictive performances. These deep generative models are described in detail below.

2.1. CTGAN

The CTGAN architecture was developed based on the GAN (Figure 1) and modified the TGAN regarding the loss function and normalization method for continuous data. CTGAN introduces the Wasserstein GAN (WGAN) [23] into the loss function (Equation (1)). During learning in the discriminator, the WGAN measures the convergence of the distance between distributions using the Wasserstein distance rather than the Jensen–Shannon divergence in the vanilla GAN. The loss function was optimized using adaptive moment estimation (Adam). The Wasserstein loss function is expressed as

$$\mathcal{L}(D, G) = E_{G(z)\sim p_g}[D(G(z))] - E_{x\sim P_r}[D(x)] \quad (1)$$

The loss function is represented as $\mathcal{L}(\cdot)$. $E[\cdot]$ represents the expected loss. D is the discriminator function. G is the generator function. Neural networks implement both functions. z is the noise derived from the standard normal distribution ($N(0,1)$). $G(z)$ represents the synthetic data generated by the generator function. p_r and p_g represent the real and generated data distribution. x represents real data. This modification decreased the mode-dropping phenomenon observed in vanilla GAN [8]. Moreover, CTGAN implements the normalization method with the variational Gaussian mixture model (VGMM) [24] rather than min–max normalization using the Gaussian mixture model (GMM) [24] used in TGAN. Unlike the min–max normalization, VGMM can generate continuous data from normalized complex distributions. This approach is referred to as mode-specific normalization.

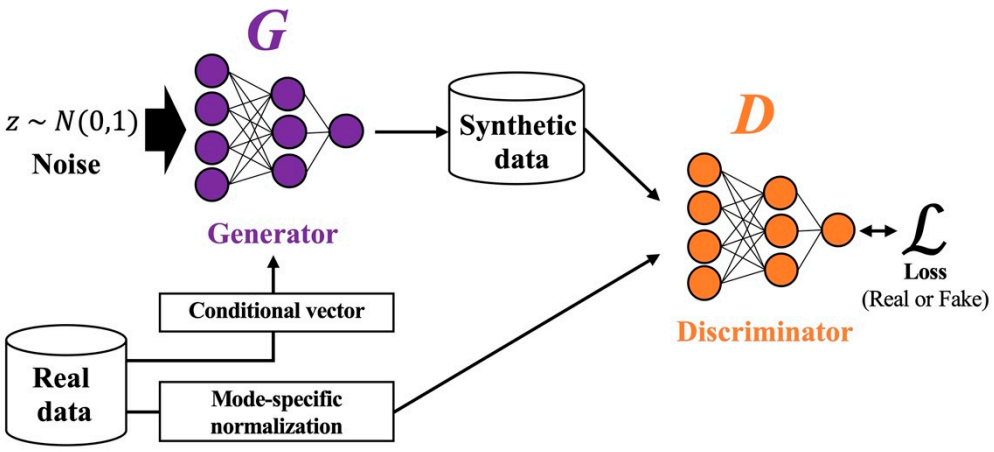

Figure 1. CTGAN architecture. The figure is based on Xu et al. [20].

2.2. CopulaGAN

Copula is a function that transforms multiple marginal cumulative distribution functions (CDFs) into a univariate joint CDF. Thus, a copula includes information on the dependencies between each CDF of variables. CopulaGAN is an algorithm applying the copula theory to the learning of the CTGAN. Accordingly, CopulaGAN has the ability to learn correlations between columns in tabular data.

2.3. TVAE

The architecture of the TVAE is similar to VAE (Figure 2a). The loss function is represented as $\mathcal{L}(\cdot)$. $E[\cdot]$ represents the expected loss. $q_\phi(z|x)$ is the probabilistic encoder; $p_\theta(x|z)$ is the probabilistic decoder. These modules were both implemented using neural networks. $p_\theta(z)$ is the prior distribution. Z is the latent variables sampled from the latent distribution ($N(\mu, \sigma)$). Figure 2b depicts a graphical model of the VAE. ϕ is variational parameters; θ is generative model parameters. Both are parameters of neural networks. Similar to the vanilla VAE [10], the evidence lower bound (ELBO) is used as a loss function; however, it was modified as in the CTGAN framework [20]. The ELBO loss can be expressed as follows:

$$\mathcal{L}_{\theta,\phi} = E_{q_{\phi(z|x)}}[log p_\theta(x|z)] - KL(q_\phi(z|x)p_\theta(z)) \quad (2)$$

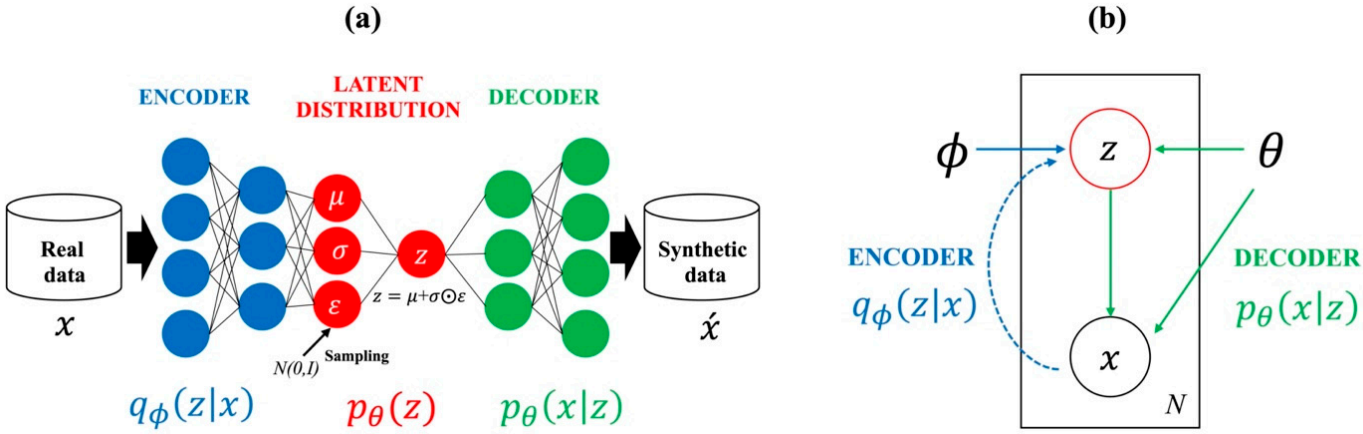

Figure 2. VAE. (**a**) Architecture of VAE. (**b**) Graphical model of VAE.

The first term of Equation (2) is the reconstruction loss. The second term is the Kullback–Leibler loss. Because the ELBO loss function is trained using the Adam, Equation (2) is transformed as follows:

$$\mathcal{L}_{\theta,\phi} = -E_{q_{\phi(z|x)}}[log p_\theta(x|z)] + KL(q_\phi(z|x)p_\theta(z)) \quad (3)$$

To handle tabular data in the VAE, each row r is pre-processed as follows [25]:

$$r = cat(\alpha_{Nc}, \beta_{Nc}, d_{Nd}) \quad (4)$$

Accordingly, $p_\theta(x|z)$ is expressed as $p_\theta(r|z)$. Nc is a continuous column. Nd is a discrete column. α_{Nc} are normalized values, and β_{Nc} is one hot vector coming from Gaussian distributions. d_{Nd} are discrete variables. Thus, regarding the TVAE, the optimized probabilistic decoder is eventually expressed as the following joint distribution:

$$log p_{\theta(r|z)} = \sum^{Nc} log \frac{1}{\sqrt{2\pi\delta}} exp \frac{(\alpha - \hat{\alpha})}{2\delta^2} + \sum^{Nc} CE(\hat{\beta}, \beta) + \sum^{Nd} CE\left(\hat{d}, d\right) + constant \quad (5)$$

Here, $\hat{\alpha}$, $\hat{\beta}$, and $\hat{d}$ are random variables. $CE(\cdot)$ is the conditional expected loss. δ is the parameter in this network and it is trained using the Adam optimizer.

3. Methods

This study was conducted in accordance with the procedure depicted in Figure 3. First, we conducted an experiment to collect the seed data for augmentation. Second, we augmented the seed data using synthesized data generated using deep generative algorithms. Third, we built a predictive model. Because the built model in this phase is derived from the seed data, it could be considered a normative promising model. Accordingly, in the validation block, whether synthesized data become close to real data can be confirmed by comparing predictive results based on both data. Accordingly, in the validation phase, whether synthesized data become close to real data can be confirmed by comparing predictive results based on both data. In the last step, we validated the usefulness of the synthesized data using a machine learning algorithm in terms of predictive performance. The detailed explanations are presented below.

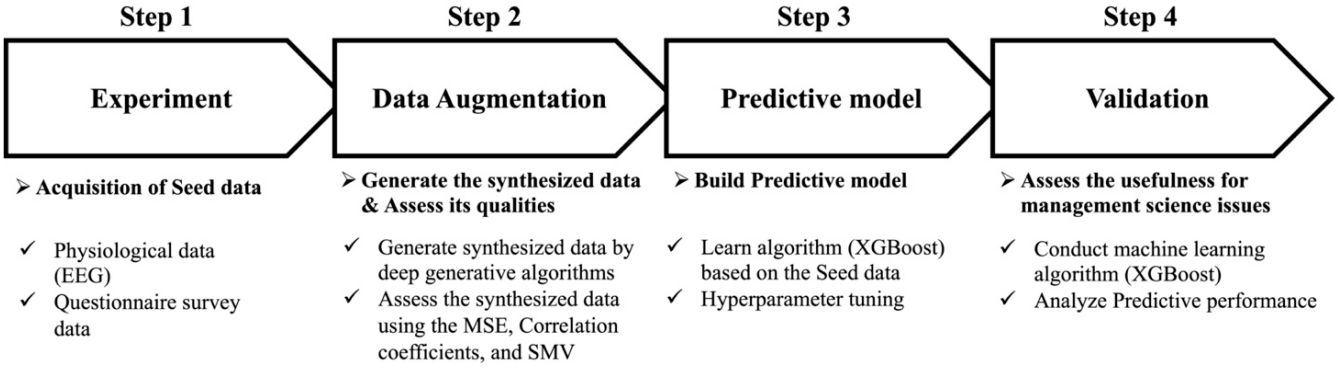

Figure 3. Study procedure. Abbreviations: EEG, electroencephalography; MSE, mean squared error; SMV, structure matching value.

3.1. Acquisition of Seed Data for Generative Synthetic Data

We conducted an EEG physiological experiment and questionnaire survey to acquire seed data and generate synthetic data for discriminating and predicting skilled nurses. The procedure is explained below.

3.1.1. Experimental Background: Building Multimodal Database for Predicting Skillful Nurses

Educators must assess the potential of their students to develop skills comparable to those of experienced nurses when teaching nursing students. Similarly, managers may prefer to identify candidates likely to fit well into their team of veteran nurses when hiring new nurses. Previous studies have demonstrated that empathy is essential in building relationships between patients and nurses [26,27]. Given that empathy is among the nursing

skills [28], this aspect suggests that empathetic abilities and dispositions may influence the professional attitudes of nurses [29]. We used EEG signals and questionnaire surveys to evaluate the empathetic dispositions of nurses at both the conscious and non-conscious levels in this study. Developing a database based on a combination of these data sources may be an effective way to manage nursing staff resources. Analyzing the combined multimodal data in the database could potentially enhance the predictive performance of skilled nurses. However, owing to their busy schedules, recruiting sufficient nurses to participate in the experiment proved challenging. Subsequently, the sample size for evaluating nurses' skills in the experiment was anticipated to be limited, and the prediction model had to achieve high accuracy with limited data. In addition, when nurses leave a hospital, privacy concerns prevent their data from being used for educational or hiring purposes, reducing the amount of available data. Thus, nurses can be considered a rare population [5], and experiments with nurses also include privacy concerns. Developing a reliable database to assess the skills of nurses is a complex challenge due to issues related to recruitment, privacy protection, and sample size limitations. These challenges are intertwined with those described in the Introduction, making the problem even more challenging. A secure database of multimodal data with a sufficient sample size is required to manage the resources of nurses.

3.1.2. Participants

The participants were 11 healthy right-handed Japanese women (mean age: 24.5 years old, range: 21–29 years old). As there were concerns about yielding different EEG data depending on dominant hands [30], we collected right-handed participants because right-handed persons might generally be major in handedness. They were divided into two groups. Group 1 was the low-nursing-skill group, which consisted of five nursing school students (mean age: 21.2 years old, range: 21–22 years old). Group 2 was the high-nursing-skill group, which consisted of six veteran nurses with five years of experience in nursing jobs (mean age: 27.5 years old, range: 26–29 years old, mean job duration: 6 years, range: 5–7 years). They agreed to participate in this experiment after receiving explanations of their informed consent and signing to acknowledge the disclaimers and their rights. The Ethics Committee at Prefectural University of Hiroshima approved this study. The present study was conducted according to the Declaration of Helsinki. One Group 2 participant was excluded from the data analysis because their EEG data could not be obtained due to machine troubles. The Bluetooth connection between the EEG device and the personal computer was disrupted for some reason during the experiment for that participant. Finally, the number of participants was 10 (Group 1 = 5; Group 2 = 5).

3.1.3. Stimuli and Procedure

Visual stimuli were prepared according to previous studies concerning the prevailing pain–empathy experimental procedure [31]. The visual stimuli consisted of twelve color photos: pain pictures (six photos) and no-pain pictures (six photos). The pain picture was depicted in gruesome scenes in which a knife or scissors pierced the left hand. The no-pain picture was similarly depicted in no-pain scenarios to the pain picture. All images were captured from a first-person perspective. The stimuli were randomly presented using Psychopy [32] implemented in Python, a program package for controlling psychophysics experiments. Participants wore an EEG headset and stared at a computer screen. After the EEG experiment, the participants were instructed to answer 20 questions concerning empathy disposition.

3.1.4. Data Collection and Analysis

EEG data were collected using an EMOTIV Epoc X headset composed of 14 channel electrodes set according to the International 10–20 system at a sampling rate of 128 Hz. The signals were analyzed using EEGLAB [33] implemented in MATLAB to remove artifacts from the raw EEG signals. First, high-pass (1 Hz) and low-pass (40 Hz) filters were

applied. After filtering, cleaned EEG signals were obtained for each participant using appropriate artifact-removing techniques such as an independent component analysis (ICA) and various epoch rejection methods, according to standard procedures regarding the artifacts and epoch rejection of the EEG signal [34,35]. Regarding the questionnaire survey for measuring empathetic dispositions, the participants evaluated 20 items on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree).

3.2. Data Augmentation: Generating Synthetic Data and Checking the Quality

We developed deep generative models using the SDV package [36] in Python and generated synthetic data using the default parameters of the deep generative models (CTGAN, CopulaGAN, and TVAE), excluding the number of epochs. Furthermore, we set the number of epochs from 5 to 1000 in eight grades (5, 10, 30, 50, 100, 300, 500, and 1000) and generated synthetic data from two deep generative models (CTGAN, CopulaGAN, and TVAE) according to multiple numbers of real data from 100% to 3000%. As well as the number of those epochs, the other parameters set the default parameters. The default parameters are listed in Supplementary Table S1. We assessed the quality of the synthetic data from two perspectives: the reproducibility of each variable and the entire structure. Two measurements were obtained from the former perspective.

First, the mean squared errors (MSEs) and correlation coefficients (CCs) of the real and generated synthetic data for each augmented sample in each epoch were calculated. These indices were calculated for the same variables in each dataset. Lower MSE values and higher CC values indicate a more satisfactory quality. The degree of constructive matching between real and generated synthetic data is the measurement in the latter view. This measurement calculates the total average of the squared error for each corresponding pair of elements of the CC matrices between the real and generated synthetic data. The small values of this measurement indicate that the synthetic data can be accurately generated by considering the relationships between the variables. We refer to this measurement as the structure matching value (SMV). The SMV is defined as follows:

$$SMV = \frac{1}{N}\sum^{N}\left(r^{G} - r^{R}\right)^{2} \quad (6)$$

where r^G is an element of the correlation coefficient matrix in the generated synthetic data, r^R is an element of the correlation coefficient matrix in the real data, and N denotes the number of variables. Consequently, if the MSE is small, the CC is high, and the SMV is low, the generated data can be sufficiently close to the real data.

3.3. Predictive Modeling

We used the XGBoost algorithm [37] as the predictive model. The feature variables were EEG signals from 14 electrode channels and Likert scale data from 20 empathic disposition questionnaires. The dependent variables were binary data (veterans = one, not veterans <students> = zero). The hyperparameters of the XGBoost algorithm were optimized using Optuna in the Python package. The optimized hyperparameters are listed in Supplementary Table S2.

3.4. Validation Procedures for Synthetic Data

Whether the generated synthetic data can replace real data must be assessed. We evaluated the effectiveness of synthetic data using an evaluation framework (Figure 4). The pipeline is classified into four blocks. The Experiment block is the data collection phase. The Deep Generative Models block is the synthesized data generation phase. Deep generative models learn the seed data in this phase. The Building Predictive Model block is the parameter optimization phase to build the trained model based on the seed data. In the validation block, evaluation metrics (accuracy, precision, recall, AUC, and F1 score)

are calculated on every input piece of data (generated data in each epoch). The main code (deep generative model and XGBoost) is described in Supplementary Codes S1 and S2.

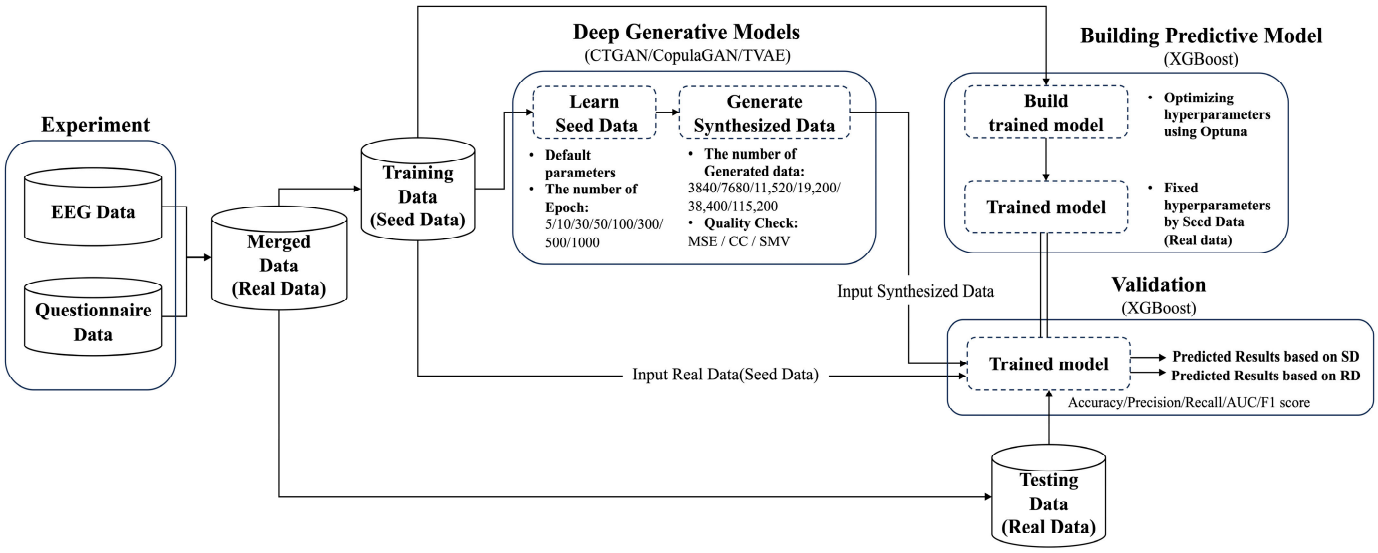

Figure 4. Analysis pipeline for evaluation. Abbreviations: EEG, electroencephalography; MSE, mean squared error; CC, correlation coefficient; SMV, structure matching value; SD, seed data; RD, real data; AUC, the area under the curve.

The data were divided into training and testing sets in a ratio of 7:3. The training dataset was used to model and generate synthetic data. In the modeling phase, the training dataset was used in the XGBoost algorithm. The synthetic data were generated using the same training dataset and deep generative models. The trained XGBoost model validated both the synthetic and test data. Evaluation metrics (accuracy, area under the curve (AUC), precision, recall, and F1 score) were used to assess the predicted results produced through these two datasets. Here, as for evaluation metrics, the relationships between predicted and actual values are represented in Table 1. The table organizing the relationship is referred to as a confusion matrix. True positive (TP) are data classified as positive by the predictive model that actually are positive. False positive (FP) are data classified as positive by the predictive model that actually are negative. False negative (FN) are data classified as negative by the predictive model that actually are positive. True Negative (TN) are data classified as negative by the predictive model that actually are negative.

Table 1. Confusion matrix.

		Actual Values	
		Positive (1)	**Negative (0)**
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

The row represents predicted values, and the column represents actual values.

Evaluation metrics are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN\ (= All\ samples)} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (10)$$

Accuracy is the overall collectiveness of predicted results (Equation (7)). Precision is the ratio of true positive values to the total number of positive values (Equation (8)). Recall is the ratio of true positive values to the total actual positive values (Equation (9)). The F1 score is the harmonic mean of precision and recall (Equation (10)). The F1 score measures the balance of precision and recall. If one or the other values are high or low, the F1 score is low. However, when both indices are high, the F1 score becomes high. As for AUC, we explain the index, referring to Figure 5. To calculate the AUC, the receiver operating characteristic (ROC) curve needs to be created. The ROC curve is composed of the false positive rate (FPR) and the true positive rate (TPR). The FPR and TPR are calculated as follows:

$$FPR = \frac{FP}{FP + TN} \quad (11)$$

$$TPR = \frac{TP}{TP + FN} \quad (12)$$

The FPR is the ratio of false positive values to the total actual negative values (Equation (11)). The TPR is the ratio of true positive values to the total actual positive values (Equation (12)). The TPR is ultimately the same as the recall. The ROC curve is created on the horizontal axis as the FPR and the vertical axis as the TPR (Figure 5). The ROC curve is created by plotting FPR and TPR depending on different thresholded values for classification. As depicted in Figure 5a, the diagonal blue dot line is lined by randomly predicting outcomes. The red line is placed over the blue dot line. This indicates that the predictive model of the red line achieved better than the chance level performances. The AUC is the area under the ROC curve, and the values of the AUC fall between 0 and 1 (Figure 5b). A value close to 1 signals good predictive performance.

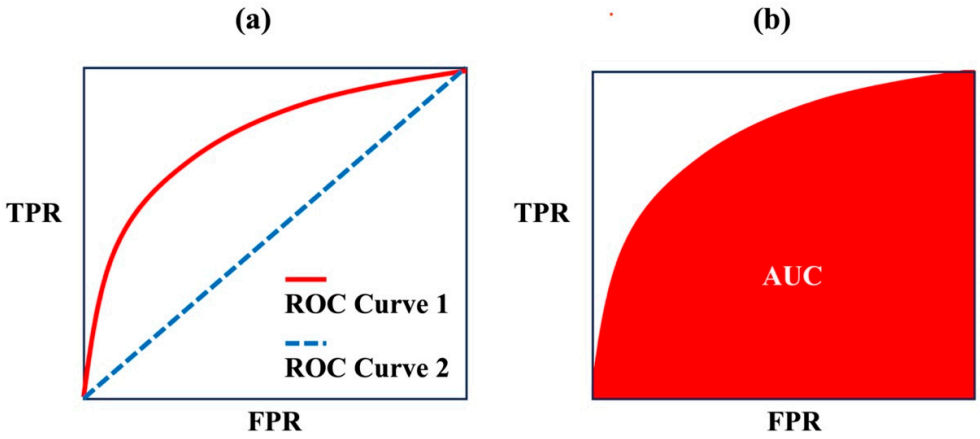

Figure 5. Explanation of AUC. (**a**) ROC curve. (**b**) AUC. Abbreviations: TPR, True Positive Rate; FPR, False Positive Rate; ROC, Receiver Operating Characteristic; AUC, Area Under the Curve.

3.5. Experimental Environment

The EEG signal was collected using a Windows 10 computer equipped with 64 GB memory, a Quadro RTX3000 GPU with Max-Q Design 6 GB, and an Intel Core i7 9750H 6 core/12 thread 4.5 GHz CPU. Synthetic data were generated using Mac OS Ventura on a machine with 32 GB memory, 2.3 GHz, QuadCore Intel Core i7.

4. Results

4.1. EEG and Questionnaire Survey Data

EEG data were obtained from each participant. The time sequence data of the EEG has 384 rows for each participant; thus, ten participants totaled 3840 rows. The columns of the EEG data are the EEG channels (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6,

F4, F8, and AF4). Accordingly, the EEG data contained 3840 rows and 14 columns. The questionnaire survey data were organized into datasets with 10 rows and 20 columns. Descriptive statistics are shown in Supplementary Table S3.

4.2. Combined Data from the EEG and Questionnaire Survey Data

EEG and questionnaire survey data were merged into a single dataset (Figure 6). However, because the two datasets had different numbers of rows, the empty fields in the questionnaire survey data were filled with the same values in the first-row data for each participant. Eventually, the dataset for the analysis contained 3840 rows and 34 columns, excluding the column of respondents.

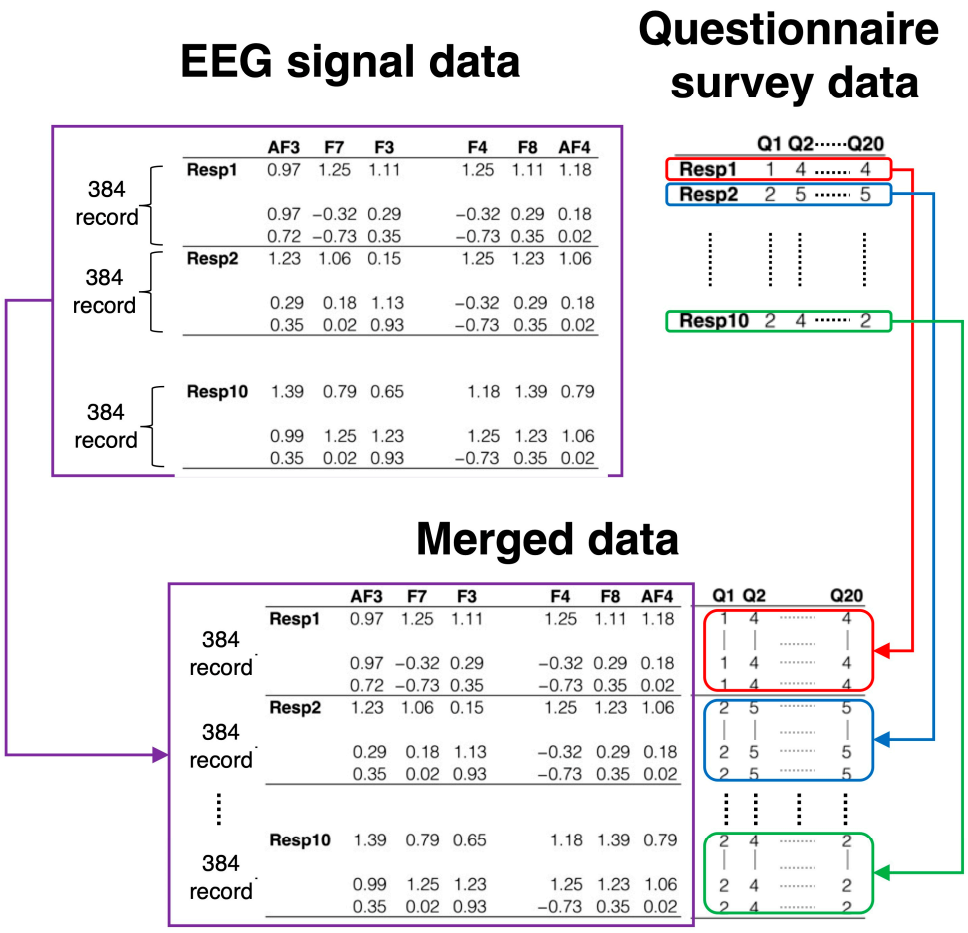

Figure 6. Data merge schema. Abbreviation: Resp represents the respondent. The black dot lines represent different numbers; the black straight lines represent the same numbers. The purple straight lines represent the EEG signal data. The red, blue, and green straight lines represent respondents 1, 2, and 10's answered data, respectively.

4.3. Qualities of the Synthetic Data

We checked the quality of the synthetic data generated using three indices: MSE, CC, and SMV. A detailed explanation of these indices is provided in the Methods section. The results of these indices are shown in Figure 7 and Supplementary Table S4. In all the deep generative models, all data quality indices improved with an increase in the number of epochs. Although the TVAE outperformed both GAN-based synthesizers (CTGAN and CopulaGAN) in overall quality indices, the data generated by the TVAE had no variance until 30 epochs owing to the generation of the same numbers. Accordingly, the CCs could not be calculated from 5 to 30 epochs in the TVAE. In particular, the CopulaGAN

outperformed CTGAN in terms of CCs during early epochs and yielded better SMV performances than CTGAN during late epochs.

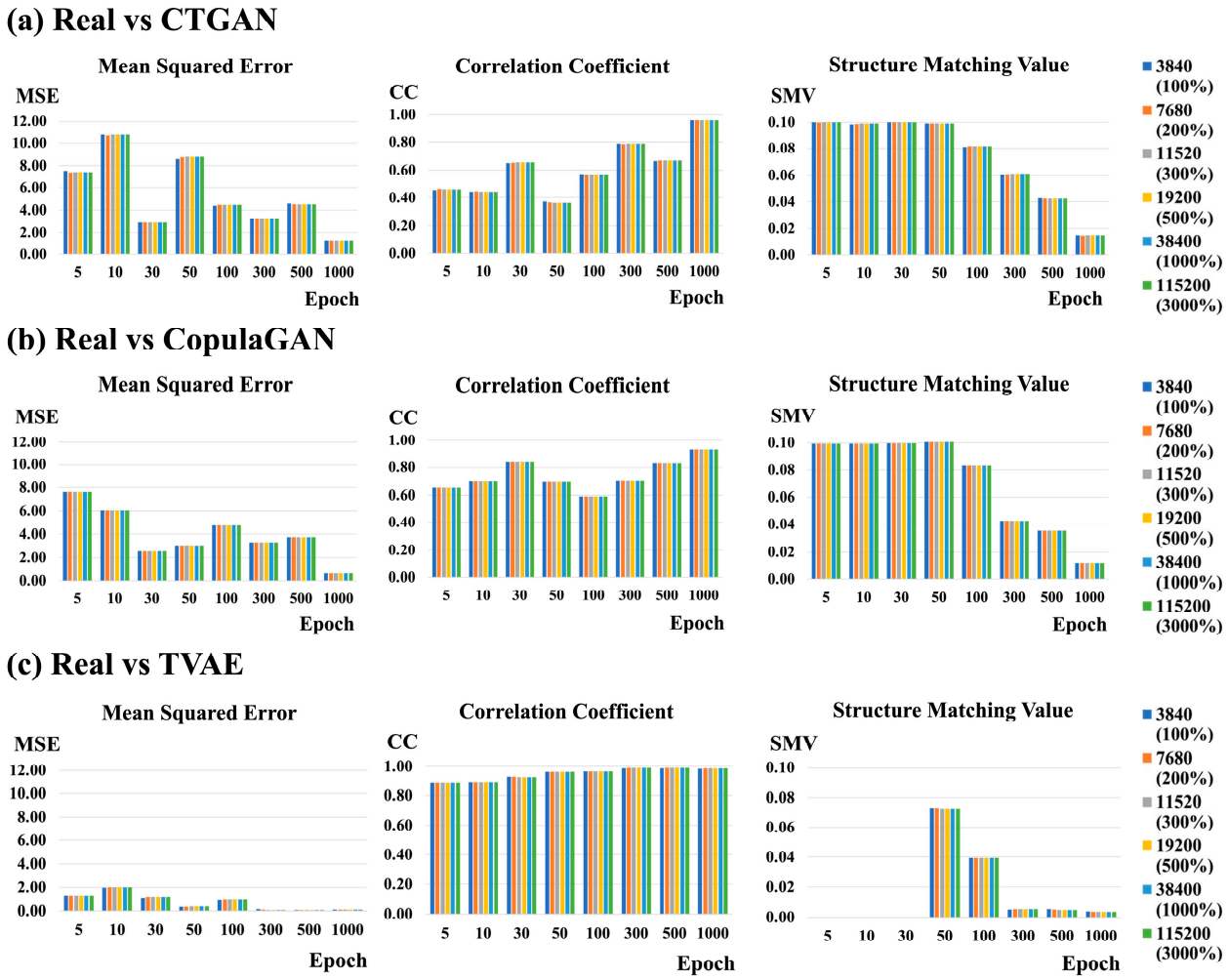

Figure 7. Results of comparison between the real and generated synthetic data. Abbreviations: MSE: mean squared error, SMV: structure matching value. (**a**) Real vs. CTGAN. Legend is at the top of the right end. Legend represents the sample sizes generated by the CTGAN. (**b**) Real vs. CopulaGAN. Legend is at the top of the right end. Legend represents the sample sizes generated by CTGAN. (**c**) Real vs. TVAE. Legend is at the bottom of the right end. Legend represents the sample sizes generated by TVAE.

4.4. Predictive Results

The predictive performances are shown in Table 2 and Supplementary Table S5. The evaluation metrics for both deep generative models improved as the number of epochs increased, regardless of the amount of synthetic data.

Table 2. Comparison of predictive results.

Epoch				**10**			**100**			**1000**		
Algorithms				**CT GAN**	**Cop-GAN**	**TVAE**	**CT GAN**	**Cop-GAN**	**TVAE**	**CT GAN**	**Cop-GAN**	**TVAE**
Calculation time				15.3 s	16.3 s	**12.4 s**	1 min 25 s	1 min 25 s	**39.1 s**	14 min 29 s	14 min 20 s	**4 min 18 s**
Evaluation metrics	**Accuracy**	**Syn-sample sizes**	**3840 (100%)**	0.484	0.465	**0.548**	0.971	0.833	**1.000**	**1.000**	**1.000**	**1.000**
			115,200 (3000%)	**0.540**	0.465	0.480	0.992	0.833	**0.994**	**1.000**	**1.000**	**1.000**
	AUC	**Syn-sample sizes**	**3840 (100%)**	0.478	0.487	**0.538**	0.892	0.762	**0.993**	**1.000**	0.999	**1.000**
			115,200 (3000%)	**0.530**	0.487	0.487	0.895	0.762	**0.961**	**1.000**	0.999	0.997
	Precision	**Syn-sample sizes**	**3840 (100%)**	0.476	0.489	**0.533**	**1.000**	0.802	0.995	**1.000**	**1.000**	**1.000**
			115,200 (3000%)	**0.536**	0.489	0.488	**1.000**	0.802	0.973	**1.000**	**1.000**	**1.000**
	Recall	**Syn-sample sizes**	**3840 (100%)**	0.432	0.568	**0.602**	0.784	0.695	**0.992**	**1.000**	0.997	**1.000**
			115,200 (3000%)	0.445	**0.568**	0.536	0.789	0.695	**0.948**	**1.000**	0.997	0.995
	F1	**Syn-sample sizes**	**3840 (100%)**	0.453	0.525	**0.565**	0.879	0.745	**0.993**	**1.000**	0.999	1.000
			115,200 (3000%)	0.486	**0.525**	0.511	0.882	0.745	**0.960**	**1.000**	0.999	0.997

Parentheses in the synthesized sample size row represent the multiple of the seed data. Bolded numbers represent the best performance in each synthesized sample size row within each epoch. Abbreviations: AUC, Area Under Curve; Cop-GAN, CopulaGAN; Syn-sample sizes, Synthesized Sample Sizes.

According to the increasing number of epochs, both GAN-based synthesizers (CTGAN and CopulaGAN) required much more computation time than the TVAE; however, there were minor differences in the calculation time between those GAN-based synthesizers and TVAE during shorter epochs (Figure 8, Table 2 and Table S6). When conducting a predictive analysis concerning real data, all predictive performances (accuracy, AUC, precision, recall, and F1 score) were 1.0. Therefore, regardless of the generated sample size, the predictive performances of the synthetic data generated using the trained models over 500 epochs were similar to those using real data. However, high volatilities in predictive performance were observed during shorter epochs (<100 epochs) in all generative models.

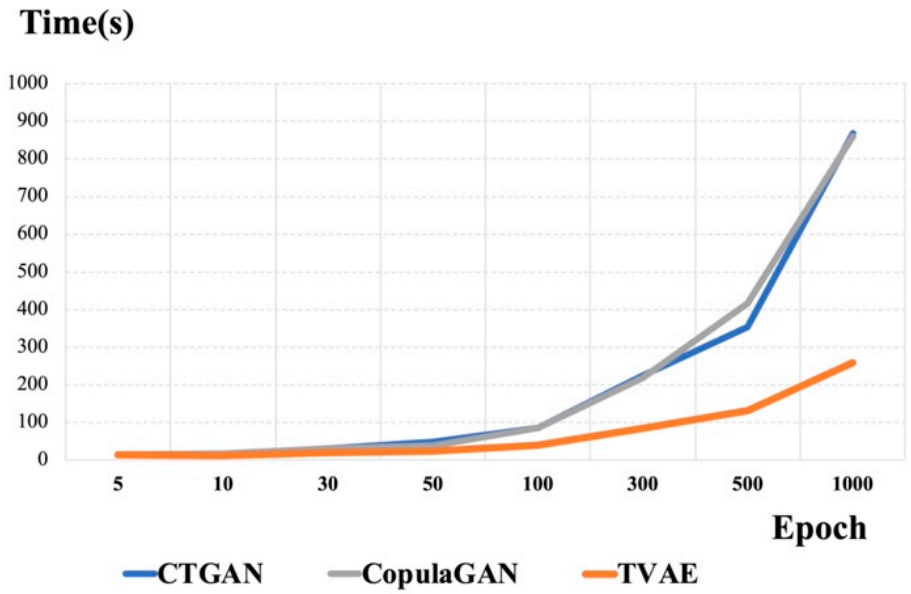

Figure 8. Computation times.

5. Discussion and Conclusions

To the best of our knowledge, this is the first study to apply deep generative models to data consisting of two different modalities in a management science study. We demonstrated that data augmentation by synthesizing different modalities of data using deep generative models might contribute to yielding good predictive performances in detecting whether candidates are veteran nurses. Moreover, we found out that the number of epochs plays a crucial role in improving predictive performance rather than the number of generated sample sizes. This suggests that our approach could apply synthesized data to business fields by anonymizing individual information because the real data might be replaceable with synthesized data, even if the synthesized data were small and limited in size. As for selecting appropriate deep generative models, because all deep generative models learned through sufficient epochs might yield good performance, an approach to synthesizing data derived from different sourced modalities might be effective for management science issues. In particular, we confirmed that TVAE outperformed the GAN-based synthesizers regarding the stability of the generated data quality through all epochs. Thus, the TVAE is the better synthesizer in validated deep generative models in our study, regarding predictive performances, computation time, and data quality. However, there are some concerns regarding the application of such models. Considering that our results are consistent with previous studies [38,39], the GAN-based synthesizers are expected to require longer computation durations than the TVAE across almost all epochs. Meanwhile, the TVAE generated no variance data for several variables during the shorter epochs (5–30 epochs). This outcome implies that the TVAE has a computational time advantage over the CTGAN; however, it requires sufficient epochs to attain satisfactory data quality. Moreover, although CopulaGAN is the GAN-based synthesizer modifying the CTGAN

to grasp dependencies between features, there were a few differentiations among deep generative models. This suggests that the usage of the other GAN-based synthesizers, as well as CopulaGAN, might also be appropriate to replicate the correlation structures of data. Regardless of the model type, the predictive performance improved proportionally to the epochs and not the sample sizes generated. This outcome indicates that sufficient epochs play a crucial role in improving predictive performance. This suggests that a priory setting of sufficient epochs must be required to yield good performance using synthesized data generated using deep generative models.

In this way, our study contributes to applying different modal-based-synthesized data generated using deep generative models to management science issues. However, our study has several limitations and concerns. Most brain science studies are conducted with a small number of participants. However, Button et al. [40] pointed out that experiments with a small sample size have low statistical power and reproducibility of an experiment. Executing the experiment with a larger sample size might have been desirable in our study. Although our results may be effective under certain conditions, they were derived from only one case of predicting skilled nurses. There are a lot of issues in management science contexts to solve: social networks [41] and multiple source data fusion for management decision-making systems in the data transformation era [42]. Therefore, our results should be applied with caution. This study may have attained better predictive performance because we generated and predicted data in a well-designed experiment. Given the few cases in which data based on well-designed experiments can be obtained in the management science field, the extent to which our results can be applied is still being determined. Summarily, trials using various types of multimodal seed data are required. In addition, when solving a more complex problem, an approach combining multiple synthesizing methods might be required [12,13]. Moreover, given that the newly GAN-based approach of introducing a roundtrip method to the conditional GAN is proposed [43], applying it to tabular-formed data might be promising. Although the present study calculated the deep generative models using default parameters, the approach of automatically optimizing hyperparameters of the GAN-based model has been proposed [44]. The evolutionary architectural search GAN (EAS-GAN) enables the optimization of hyperparameters, including network architecture [45]. These automated methods do not require searching for optimized parameters through trial-and-error. Although those methods are not applied to the tabular data format, another finding could be brought to our study if the optimized parameters were adopted in this study. Thus, further research is required to facilitate the use of generative multimodal data for management science challenges. Concretely, first, although the experiment in the present study was conducted with a small participant sample, we need to perform a comparable experiment between the results of the augmented data based on a small participant sample and a large participant sample's results under the condition of sampling these participants from the same population in the next step. This experiment might enable us to confirm the reproducibility of the results of experiments with a large participant sample by using augmented data. Second, to confirm the robustness of our results concerning the other new multimodality data besides both EEG and questionnaire survey data, we need to experiment with adding other modalities, such as photo images, movies, texts, music, and other physiological data. Third, to validate generalizing our results, we should apply the approaches of the present study to a wide variety of management science fields, including digital marketing and decision-making systems for managing businesses.

In conclusion, the present study provides a beneficial approach for applying multimodal data to management science issues as a first step. We demonstrated that with sufficient training in deep generative models, we could generate big data almost similar to real data, using small multimodal data as seed data, which included EEG signals and questionnaire survey data. Furthermore, our results could be extended to other issues in management science, such as marketing management and marketing research, since

they overcome the restrictions of the limited sample size and privacy concerns when using multimodal data, including physiological data such as EEG signals.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app14010378/s1, Table S1: The default parameters in each deep generative model. Table S2: Results of optimizing parameters. Table S3: Basic statistics. Table S4: Data qualities in each deep generative model. Table S5: Predictive performance in each deep generative model. Table S6: Computation time. Code S1(A): CTGAN, Code S1(B): CopulaGAN, Code S1(C): TVAE, Code S2(A): XGBoost(Validation: Setting the trained model), Code S2(B): XGBoost(Validation: Predicted results).

Author Contributions: Conceptualization, S.W. and K.E.; methodology, S.W.; software, S.W.; validation, S.W.; formal analysis, S.W.; investigation, S.W.; resources, K.F., J.O. and K.E.; data curation, S.W., Y.N., Y.K. and M.K.; writing—original draft preparation, S.W.; writing—review and editing, S.W.; visualization, S.W.; supervision, S.W., J.O. and K.E.; project administration, S.W., J.O. and K.E; funding acquisition, K.E. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by JSPS KAKENHI (Grant no. JP17K03992).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of the Prefectural University of Hiroshima.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets in this study are available via request to the corresponding author.

Conflicts of Interest: Author Kenji Fujimoto was employed by the company Hiroshima Office, Survey Research Center Co., Ltd., and declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Liem, F.; Varoquaux, G.; Kynast, J.; Beyer, F.; Masouleh, S.K.; Huntenburg, J.M.; Lampe, L.; Rahim, M.; Abraham, A.; Craddock, R.C. Predicting Brain-Age from Multimodal Imaging Data Captures Cognitive Impairment. *Neuroimage* **2017**, *148*, 179–188. [CrossRef] [PubMed]
2. Lahat, D.; Adali, T.; Jutten, C. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proc. IEEE* **2015**, *103*, 1449–1477. [CrossRef]
3. Hakim, A.; Klorfeld, S.; Sela, T.; Friedman, D.; Shabat-Simon, M.; Levy, D.J. Machines Learn Neuromarketing: Improving Preference Prediction from Self-Reports Using Multiple EEG Measures and Machine Learning. *Int. J. Res. Mark.* **2021**, *38*, 770–791. [CrossRef]
4. Malhotra, N.K.; Agarwal, J.; Peterson, M. Methodological Issues in Cross-Cultural Marketing Research: A State-of-the-Art Review. *Int. Mark. Rev.* **1996**, *13*, 7–43. [CrossRef]
5. Thompson, W. *Sampling Rare or Elusive Species: Concepts, Designs, and Techniques for Estimating Population Parameters*; Island Press: Washington, DC, USA, 2013.
6. Ko, J.; Lu, C.; Srivastava, M.B.; Stankovic, J.A.; Terzis, A.; Welsh, M. Wireless Sensor Networks for Healthcare. *Proc. IEEE* **2010**, *98*, 1947–1960. [CrossRef]
7. Alomar, K.; Aysel, H.I.; Cai, X. Data Augmentation in Classification and Segmentation: A Survey and New Strategies. *J. Imaging* **2023**, *9*, 46. [CrossRef] [PubMed]
8. Courville, A.; Bengio, Y. Generative Adversarial Nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680. [CrossRef]
9. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv* **2015**, arXiv:1511.06434. [CrossRef]
10. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2013**, arXiv:1312.6114. [CrossRef]
11. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond Empirical Risk Minimization. *arXiv* **2017**, arXiv:1710.09412.
12. Feng, H.-Z.; Kong, K.; Chen, M.; Zhang, T.; Zhu, M.; Chen, W. Shot-VAE: Semi-Supervised Deep Generative Models with Label-Aware ELBO Approximations. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 7413–7421. [CrossRef]
13. Si, C.; Zhang, Z.; Qi, F.; Liu, Z.; Wang, Y.; Liu, Q.; Sun, M. Better Robustness by More Coverage: Adversarial Training with Mixup Augmentation for Robust Fine-Tuning. *arXiv* **2020**, arXiv:2012.15699. [CrossRef]

14. Lu, Y.; Wang, H.; Wei, W. Machine Learning for Synthetic Data Generation: A Review. *arXiv* **2023**, arXiv:2302.04062. [CrossRef]
15. Assefa, S.A.; Dervovic, D.; Mahfouz, M.; Tillman, R.E.; Reddy, P.; Veloso, M. Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls. In Proceedings of the First ACM International Conference on AI in Finance, New York, NY, USA, 14–16 October 2020; pp. 1–8.
16. Anand, P.; Lee, C. Using Deep Learning to Overcome Privacy and Scalability Issues in Customer Data Transfer. *Mark. Sci.* **2023**, *42*, 189–207. [CrossRef]
17. Burnap, A.; Hauser, J.R.; Timoshenko, A. Product Aesthetic Design: A Machine Learning Augmentation. *Mark. Sci.* **2023**, *42*, 1029–1056. [CrossRef]
18. Choi, E.; Biswal, S.; Malin, B.; Duke, J.; Stewart, W.F.; Sun, J. Generating Multi-Label Discrete Patient Records Using Generative Adversarial Networks. In Proceedings of the Machine Learning for Healthcare Conference, PMLR, Boston, MA, USA, 18–19 August 2017; pp. 286–305.
19. Park, N.; Mohammadi, M.; Gorde, K.; Jajodia, S.; Park, H.; Kim, Y. Data Synthesis Based on Generative Adversarial Networks. *arXiv* **2018**, arXiv:1806.03384. [CrossRef]
20. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling Tabular Data Using Conditional Gan. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 7335–7345.
21. Kamthe, S.; Assefa, S.; Deisenroth, M. Copula Flows for Synthetic Data Generation. *arXiv* **2021**, arXiv:2101.00598.
22. Cheon, M.J.; Lee, D.H.; Park, J.W.; Choi, H.J.; Lee, J.S.; Lee, O. CTGAN VS TGAN? Which One Is More Suitable for Generating Synthetic EEG Data. *J. Theor. Appl. Inf. Technol.* **2021**, *99*, 2359–2372.
23. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved Training of Wasserstein Gans. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5767–5777.
24. Bishop, C.M.; Nasrabadi, N.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 4.
25. Xu, L. Synthesizing Tabular Data Using Conditional GAN. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2020.
26. Olson, J.; Hanchett, E. Nurse-Expressed Empathy, Patient Outcomes, and Development of a Middle-Range Theory. *Image J. Nurs. Scholarsh.* **1997**, *29*, 71–76. [CrossRef]
27. Reynolds, W.J.; Scott, B. Empathy: A Crucial Component of the Helping Relationship. *J. Psychiatr. Ment. Health Nurs.* **1999**, *6*, 363–370. [CrossRef] [PubMed]
28. Williams, J.; Stickley, T. Empathy and Nurse Education. *Nurse Educ. Today* **2010**, *30*, 752–755. [CrossRef]
29. Pérez-Fuentes, M.d.C.; Herera-Peco, I.; Jurado, M.d.M.M.; Ruiz, N.F.O.; Ayuso-Murillo, D.; Linares, J.J.G. The Development and Validation of the Healthcare Professional Humanization Scale (HUMAS) for Nursing. *Int. J. Environ. Res. Public Health* **2019**, *16*, 3999. [CrossRef] [PubMed]
30. Provins, K.A.; Cunliffe, P. The Relationship between EEG Activity and Handedness. *Cortex* **1972**, *8*, 136–146. [CrossRef] [PubMed]
31. Gu, X.; Han, S. Attention and Reality Constraints on the Neural Processes of Empathy for Pain. *Neuroimage* **2007**, *36*, 256–267. [CrossRef] [PubMed]
32. Peirce, J.W. PsychoPy—Psychophysics Software in Python. *J. Neurosci. Methods* **2007**, *162*, 8–13. [CrossRef] [PubMed]
33. Delorme, A.; Makeig, S. EEGLAB: An Open Source Toolbox for Analysis of Single-Trial EEG Dynamics Including Independent Component Analysis. *J. Neurosci. Methods* **2004**, *134*, 9–21. [CrossRef]
34. Urigüen, J.A.; Garcia-Zapirain, B. EEG artifact Removal—State-of-the-Art and Guidelines. *J. Neural Eng.* **2015**, *12*, 031001. [CrossRef]
35. Jiang, X.; Bian, G.-B.; Tian, Z. Removal of Artifacts from EEG Signals: A Review. *Sensors* **2019**, *19*, 987. [CrossRef]
36. Patki, N.; Wedge, R.; Veeramachaneni, K. The Synthetic Data Vault. In Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 17–19 October 2016; pp. 399–410. [CrossRef]
37. Chen, T.; Guestrin, C. Xgboost: A Scalable Tree Boosting System. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
38. Watson, D.S.; Blesch, K.; Kapar, J.; Wright, M.N. Smooth Densities and Generative Modeling with Unsupervised Random Forests. *arXiv* **2022**, arXiv:2205.09435. [CrossRef]
39. Muñoz-Cancino, R.; Bravo, C.; Ríos, S.A.; Graña, M. Assessment of Creditworthiness Models Privacy-Preserving Training with Synthetic Data. In Proceedings of the Hybrid Artificial Intelligent Systems: 17th International Conference, HAIS 2022, Salamanca, Spain, 5–7 September 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 375–384.
40. Button, K.S.; Ioannidis, J.P.; Mokrysz, C.; Nosek, B.A.; Flint, J.; Robinson, E.S.; Munafò, M.R. Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Nat. Rev. Neurosci.* **2013**, *14*, 365–376. [CrossRef] [PubMed]
41. Anastasiei, B.; Dospinescu, N.; Dospinescu, O. Word-of-Mouth Engagement in Online Social Networks: Influence of Network Centrality and Density. *Electronics* **2023**, *12*, 2857. [CrossRef]
42. Liu, L.; Wan, X.; Li, J.; Wang, W.; Gao, Z. An Improved Entropy-Weighted Topsis Method for Decision-Level Fusion Evaluation System of Multi-Source Data. *Sensors* **2022**, *22*, 6391. [CrossRef]
43. Liu, Q.; Xu, J.; Jiang, R.; Wong, W.H. Density Estimation Using Deep Generative Neural Networks. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2101344118. [CrossRef]

44. Ganepola, V.V.V.; Wirasingha, T. Automating Generative Adversarial Networks Using Neural Architecture Search: A Review. In Proceedings of the 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 5–7 March 2021; pp. 577–582. [CrossRef]
45. Lin, Q.; Fang, Z.; Chen, Y.; Tan, K.C.; Li, Y. Evolutionary Architectural Search for Generative Adversarial Networks. *IEEE Trans. Emerg. Top. Comput. Intell.* **2022**, *6*, 783–794. [CrossRef]