BWGAN-GP: An EEG Data Generation Method for Class Imbalance Problem in RSVP Tasks

Meng Xu, Yuanfang Chen, Yijun Wang, *Member, IEEE*, Dan Wang, Zehua Liu, and Lijian Zhang

Abstract—In the rapid serial visual presentation (RSVP) classification task, the data from the target and non-target classes are incredibly imbalanced. These class imbalance problems (CIPs) can hinder the classifier from achieving better performance, especially in deep learning. This paper proposed a novel data augmentation method called balanced Wasserstein generative adversarial network with gradient penalty (BWGAN-GP) to generate RSVP minority class data. The model learned useful features from majority classes and used them to generate minority-class artificial EEG data. It combines generative adversarial network (GAN) with autoencoder initialization strategy enables this method to learn an accurate class-conditioning in the latent space to drive the generation process towards the minority class. We used RSVP datasets from nine subjects to evaluate the classification performance of our proposed generated model and compare them with those of other methods. The average AUC obtained with BWGAN-GP on EEGNet was 94.43%, an increase of 3.7% over the original data. We also used different amounts of original data to investigate the effect of the generated EEG data on the calibration phase. Only 60% of original data were needed to achieve acceptable classification performance. These results show that the BWGAN-GP could effectively alleviate CIPs in the RSVP task and obtain the best performance when the two classes of data are balanced. The findings suggest that data augmentation techniques could generate artificial EEG to reduce calibration time in other brain-computer interfaces (BCI) paradigms similar to RSVP.

Index Terms—Rapid serial visual presentation (RSVP), Wasserstein generative adversarial network (WGAN), data augmentation, class imbalance problem, auto-encoder.

I. INTRODUCTION

RAPID serial visual presentation (RSVP) based on electroencephalogram is a well-established brain-computer interface (BCI) paradigm for target recognition [1], [2], in which subjects must make decisions about target images from the image flow under high temporal sensitivity conditions [3], [4]. Oddball, which randomly embeds rare "target" class stimuli into a series of "non-target" classes, triggers P300, which causes a "natural" extreme class imbalance problem (CIP). Acquiring sufficient data for training is essential for deep learning. Unfortunately, electroencephalogram (EEG) data are difficult to collect over long periods, and the data do not compensate for the CIP. Lee and Wang considered class imbalance as a main bottleneck factor contributing to the poor performance of RSVP classification, and data augmentation methods can help lessen the effects of CIP [2], [5], [6]. Data augmentation mainly enlarges an existing dataset by generating and transforming original sample data, which is a promising approach in computer vision. Typical data augmentation methods are geometric transformations, such as cropping, flipping, and scaling. Unlike images, EEG signals are acquired from multiple electrodes. As a result, EEG signals are often highly spatial-temporal correlated [7]. Therefore, the geometric transformations that are useful for image data augmentation do not apply to EEG signals. Nevertheless, some methods are still available for data augmentation of EEG signals. The data augmentation methods used for EEG include adding Gaussian noise addition, segmentation, and a generation model [8]–[10]. However, the first two methods cannot meet artificial multi-channel EEG generation needs due to redundant noise or information loss. The generation model seems to be the only potential solution to this problem. Over-sampling and upsampling are general data model generation methods, such as the synthetic minority over-sampling technique (SMOTE) [11]–[13]. Generation adversarial network (GAN) techniques have recently gained widespread attention and have significantly improved EEG signals [14]–[16]. Hartmann *et al.* generated a single-channel EEG with a GAN and achieved excellent visual inspection [15]. Zhang *et al.* proposed a conditional deep convolutional generative adversarial network (cDCGAN), which network uses class label information with conditional properties [17]. However, cDCGAN experimented on only three channels (C3, Cz, and C4). Luo *et al.* introduced a conditional Wasserstein GAN (cWGAN) framework for EEG data augmentation. The model for which generated power spectral density and differential entropy of EEG signals. This framework significantly improved the accuracy of emotion recognition models [18]. However, most of the above

models generate EEG signals by extracting spectral features to generate spectrograms, then transforming them into EEG signals by the inverse transform method. RSVP only has significant differences in the time domain, with relatively little helpful information, so data generation is more complex. Only Panwar *et al.* have proposed to combine WGAN with gradient penalty (WGAN-GP) and upsampling methods that can be trained to synthesize EEG data for different cognitive events (which abbreviated as WGAN-GP-RSVP) [19], [20]. Still, this approach does not consider the CIP.

We focused on the CIP in EEG-RSVP data for this study. GAN can generate high-quality artificial data through a large amount of training data, but the minority-class data were rare in the original EEG-RSVP. The ratio of the target (minority-class) to non-target (majority-class) data was less than 1/10. Therefore, the general GAN learns toward the majority class, which aggravates the CIP. The data imbalance can cause mode collapse in the training process. The novel idea is introduced in this paper aimed at CIP, which generates minority-class EEG signals of superior quality when trained with an imbalanced RSVP dataset. This balanced WGAN model with a gradient penalty (BWGAN-GP) combines an autoencoder and a GAN to generate minority-class artificial EEG data. It learns the majority-class features and controls the ratio of minority-class data to majority-class data in the latent space. This paper provides multiple perspectives of minority-class data generation performance to explore the proposed method's effectiveness and robustness. This study also surveyed the performance of BWGAN-GP when reducing the proportion of training data to meet practical needs.

The paper is organized as follows. The background of CIP and the concept of GAN are shown in section II. Section III presents the information of subjects, RSVP paradigm, experimental data acquisition, and preprocessing stage. In section IV, the architectures for BWGAN-GP, experimental procedures, and evaluation metrics are described. Section V, the experimental results and multi-perspective analyses of the quality of generating minority-class artificial EEG by BWGAN-GP are presented. Finally, in section VI, the main results of this study are summarized.

II. RELATED WORK

A. Class Imbalance Problem

A CIP involves challenging of classification problems in machine learning, in which the distribution of data across classes is biased or skewed [21], [22]. This result poses a severe dilemma in prediction models using many machine learning classifiers, as most are designed around balanced classes. Class imbalance brings poor prediction performance, especially for minority class data rarity or under-representation. It makes the classifier more heavily weighted toward the majority class, which causes minority class errors that are more sensitive than those of the majority class. Class imbalance problems can be described by the imbalance rate (IR) [23], which is defined as the percentage of the number of minority classes compared to the number of majority classes. The IR can be calculated via Equation (1),

$$IR = \frac{N_{minority-class}}{N_{majority-class}} \tag{1}$$

where $N_{majority-class}$ is the sample size of the majority class and $N_{minority-class}$ is the sample size of the minority class. When $IR = 1$, the datasets are balanced. When $IR < 1$, the value of the IR is smaller, so the datasets have a more significant imbalanced level. In binary classification problems, the majority class dominates the whole dataset, resulting in low classification accuracy for minority classes, enhancing overall classification performance. This study IR was used to measure the degree of imbalanced classes in BCI paradigms. In our collected RSVP dataset, IR is $\frac{1}{24}$.

B. The Rationale of GAN

The GAN proposed by Goodfellow utilizes adversarial plays between two components, known as the Generator (G) and the Discriminator (D), to learn the probability distribution of the target data [24], [25]. By its mathematical definition, GAN can be expressed through Equation (2). In GANs, G generates artificial data, whereas D distinguishes which data are real or generated. The real training data and artificial data by generator G are defined as P_r and P_g. As the number of epochs increases, the discriminator forces the generator to produce improved generated data.

$$\min_{\theta_G} \max_{\theta_D} L\left(P_r, P_g\right) = E_{x \sim P_r}\left[\log\left(D\left(x\right)\right)\right]$$
$$+ E_{z \sim P_g}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \tag{2}$$

where E denotes the expectation operator, θ_G and θ_D are the training parameters of the generator and discriminator, and $D(x)$ represents the probability of x sampled from real distribution P_r. Respectively $G(z)$ is the generated data by G from a Gaussian vector z.

Wasserstein GAN (WGAN) uses Earth-mover's distance instead of Jensen-Shannon divergence of GAN to enhance the training stability [26]–[28]. The WGAN can be obtained with Equation (3)

$$W\left(P_r, P_g\right) = \inf_{\gamma \sim \prod\left(P_r, P_g\right)} E_{\left(x_r, x_g\right)}\left[x_r - x_g\right] \tag{3}$$

where $\prod(P_r, P_g)$ is the set of the joint distribution of the real and generated distribution, γ is the joint distribution, x_r and x_g are real data and generated data, respectively. The WGAN still has slow convergence throughout the whole training process. The performance improvement is not very obvious compared to the traditional GAN, so the problem mentioned above can be solved by adding a regular term, gradient penalty, which is a variant called WGAN-GP [29]. The details can be presented in Equation (4)

$$\tilde{W}\left(P_r, P_g\right) = E_{x \sim P_r}[\log(D(x))] + E_{z \sim P_g}[\log(1 - D(G(z)))]$$
$$+ \lambda E_{\hat{x} \sim P(\hat{x})}\left[\left(\nabla_{\hat{x}} D(\hat{x})_2 - 1\right)^2\right] \tag{4}$$

where λ is the value to balance between loss and regular terms, the gradient penalty ensures that $D(x)$ is under the K-Lipschitz constraint and $\hat{x}$ is the data between real and generated data.
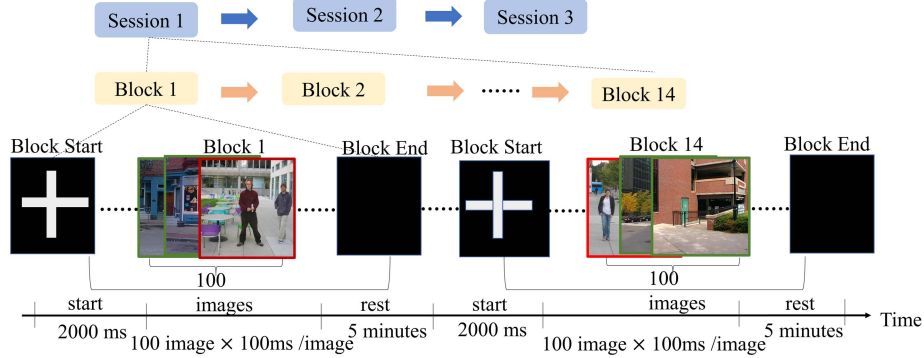
Fig. 1. The framework of the rapid serial visual presentation (RSVP) paradigm.

III. MATERIALS

A. Subjects

Nine graduate students aged 22–26 (including two female students, all right-handed) participated in the experiment. The Subject'information can be seen appendix Table. V. None of them had experienced RSVP-BCI before. In addition, no participant had any previous history of visual disorders, neurological disease or injury. All subjects had a normal or corrected-to-normal vision, signed the informed consent agreement and received monetary compensation for his or her participation. The Beijing Institute of Mechanical Equipment review committee approved this experiment.

B. RSVP Paradigm

The stimulus images for the image dataset used in the RSVP experiment were derived from street-view images collected by the Computer Science and Artificial Intelligence Library of the Massachusetts Institute of Technology. Subjects were required to identify target images from a random stream of images sequences.

The experimental design of data acquisition involved three sessions. Each session contained 14 blocks, each had 2000 ms cross-marks indicating the experimental start time, and 100 pictures (pixels) presented at a frequency of 10 Hz. The duration of each block was approximately 5 minutes, and the total number of images in the whole session was 4200. The average rest time between two sessions was about 15 min to alleviate the subjects' fatigue. The subjects were required to identify a target image from a session (the target image showing a human from non-target images containing no human count the number of the target image. The proportion of the target image to the non-target image was 1:24. Fig. 1 shows the RSVP paradigm framework.

C. EEG Acquisition and Preprocessing

The EEG was recorded through the Sysnamps2 system (Neuroscan, Inc.) at a sampling rate of 1000 Hz, and EEG data from this dataset were saved in ". cnt" format, with all 64 Ag/Agcl electrode channel positions placed according to the 10-20 system. The reference electrode was placed at the left mastoid and electrode impedance was maintained below 10kΩ. This system filtered EEG data from 0.1 Hz to 100 Hz. To remove the power-line noise, EEG data were filtered below 50 Hz. The two bad channels(M1 and M2) were removed from the 64 channels since the electrode impedances were higher than 10 kΩ.

The continuous EEG signals were bandpass filtered between 2 Hz and 30 Hz during the preprocessing phase with a fourth-order Butterworth filter. The data of each block were segmented into EEG trials (each trial corresponds to a picture). In our experiment paradigm, each subject has 14 blocks × 100 trials EEG samples per session.

IV. METHODS

A. Architecture of BWGAN-GP

The most crucial aim of this paper was to propose a novel EEG data augmentation method based on the GANs, which effectively maintain the feature of original minority class EEG signals. The general generative adversarial networks cannot generate high-quality EEG signals in imbalanced conditions, and it is challenging to generate a rare class in the RSVP dataset. Our proposed method borrowed balancing ideals [30], [31] to produce synthetic minority class EEG signals.

1) Pre-Input Autoencoder Training: An Autoencoder (AE) enables the generator and discriminator to learn common knowledge about all classes. The encoder converts the raw EEG signals into a vector in latent space, and the decoder translates the latent vector into constructed EEG signals. The L2-norm loss is applied to train AE networks. Label information is embedded in the latent vector to guide in the later training of GAN.

2) Enhanced GAN Network Structure: In the GAN initialization, the generator and discriminator inherit the same architecture and weights from the trained decoder and encoder. Unlike the general generative network structure in balancing GAN, our network structure is similar to WGAN-GP-RSVP [19], which added bicubic interpolation and bilinear weight initialization operations. The GAN network
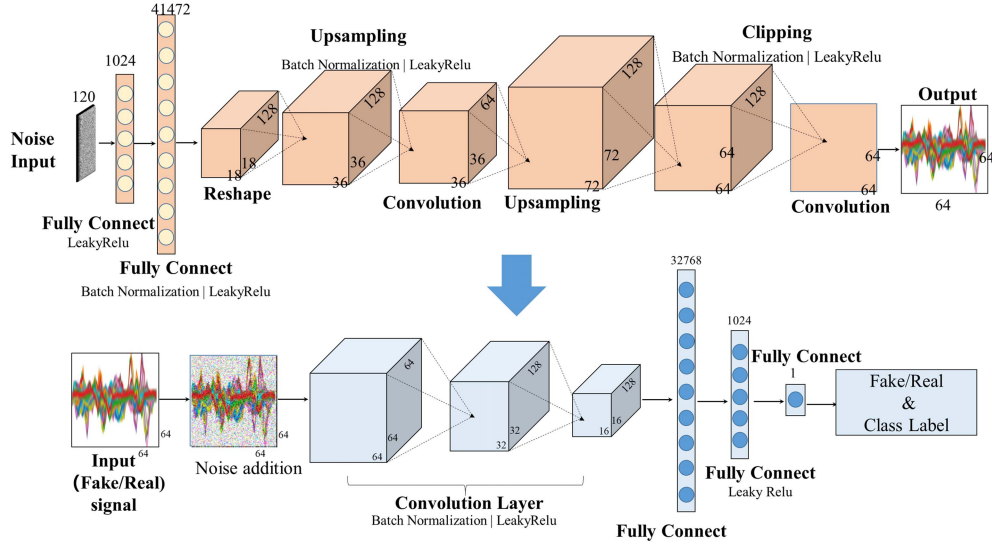
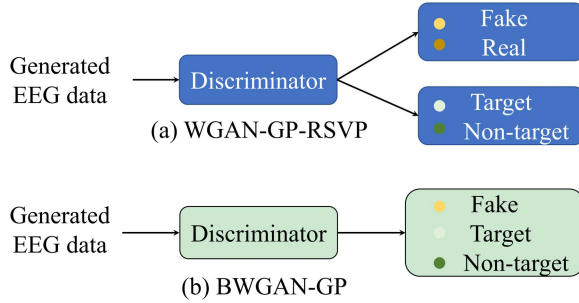Fig. 2. The BWGAN-GP network structure is similar to WGAN-GP-RSVP [19].

Fig. 3. The output of discriminator between BWGAN-GP and WGAN-GP-RSVP.

structure of BWGAN-GP is presented in Fig. 2. These upsampling methods could improve the quality of generated EEG data. The interpolation method can help the upsampling layers of the AE and GAN network avoid monumental frequency artifacts in time series signals, such as EEG signals.

3) The One Output in GAN Training: The discriminator has only one output in our proposed method. Two outputs, real/fake and categories label, conflict with each other when there is class imbalance. The discriminators tend to associate fake with minority-class real data, so the data generated is often rewarded for generating data that looks real but does not represent the minority class. The balanced design guarantees that the minority class EEG signal will not be ignored. The loss function of the discriminator is shown in Fig. 3.

4) The Loss Functions: BWGAN-GP is the idea of adding balance to WGAN-GP. Since a specific class of labels is required, it is a conditional GAN(cWGAN). The loss function of discriminator can be described as shown in Equation (5),

$$
\begin{aligned}
L^{(D)} & \left(P_r, P_g, Y_r, Y_f \right) \\
= & -E_{x, y_r \sim P_r, Y_r} \left[\log \left(D \left(x, y_r \right) \right) \right] \\
& - E_{z, y_f \sim P_g, Y_f} \left[\log \left(1 - D \left(G \left(z, y_f \right), y_f \right) \right) \right] \\
& + \lambda E_{\hat{x}, y_r \sim P(\hat{x}), Y_r} \left[\left(\nabla_{\hat{x}, y_r} D \left(\hat{x}, y_r \right)_2 - 1 \right)^2 \right]
\end{aligned}
\tag{5}
$$

where Y_f are fake labels. As the discriminators of traditional GAN, real labels are used to discriminate real or fake data. If the dataset is unbalanced, the discriminator is more biased toward the majority class, whereas BWGAN-GP will randomly sample a fake label from the balanced datasets generated by the generator. The loss function of the generator can be expressed through Equation (6),

$$
L^{(G)}(Y_f, P_g) = -E_{z, y_f \sim P_g, Y_f}[\log(D(G(z, Y_f)))]
\tag{6}
$$

As can be seen, BWGAN-GP includes two parts, a pre-input autoencoder and a GAN. The entire process of BWGAN-GP is depicted in Fig. 4. First, an unbalanced RSVP dataset and class labels information are used to train the autoencoder to learn common knowledge about two classes. Second, the GAN early process utilizes weights from encoder training to initialize the discriminator of the GAN. In contrast, results from the decoder can initialize the generator of the GAN. The reconstructed EEG data from the decoder is given as input to the GAN, which c good, stable solutions. Last, the GAN network is trained to generate high-quality minority-class artificial EEG signals.

B. Classifiers

We evaluated our method of data augmentation from different classifiers. The EEGNet is considered a primary classifier, but we used seven recently published classifiers to fairly compare the quality of the generated data of different augmentation algorithms, including HDCA [32], MDRM [33], MCNN [34], BN3 [35], DeepConvNet [36], LeNet [37] and EEGNet [38].

- HDCA: A method used linearly to discriminate the weights learned separately for the temporal and spatial of EEG signals.
- MDRM: A classifier based on the Riemannian manifold. It views each covariance matrix as a point in Riemannian space. The features are obtained by using the geodesic class at the center. https://pyriemann.readthedocs.io/en/latest/
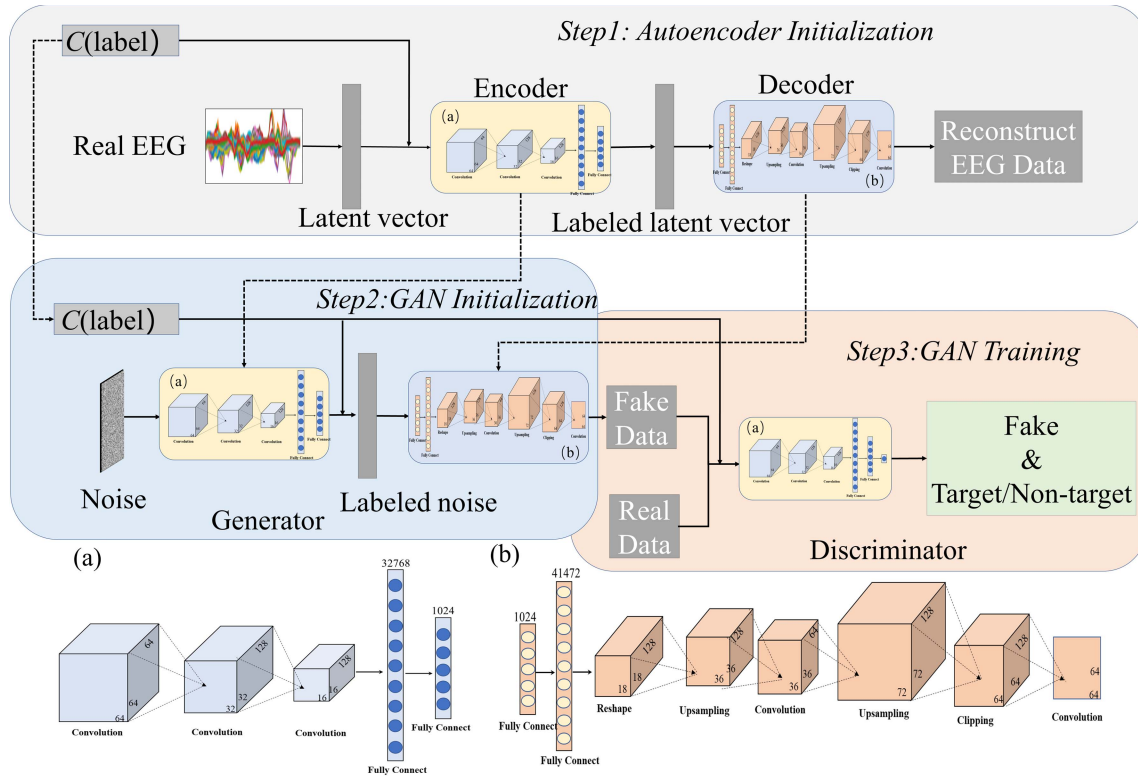
Fig. 4. The architecture of our proposed method (BWGAN-GP).

- MCNN: A CNN-based model proposed by Manor *et al.*, we re-implemented the network as described in [34]
- BN3: A batch-normalization CNN model we re-implemented the network. https://github.com/gibranfp/P300-CNNT
- DeepConvNet: A deeply CNN model consisting of five convolution layers with a softmax layer for classification. https://github.com/mochakai/DeepConvNet
- LeNet: A general CNN architecture operated in ERP applications, the network was re-implemented according to [37]
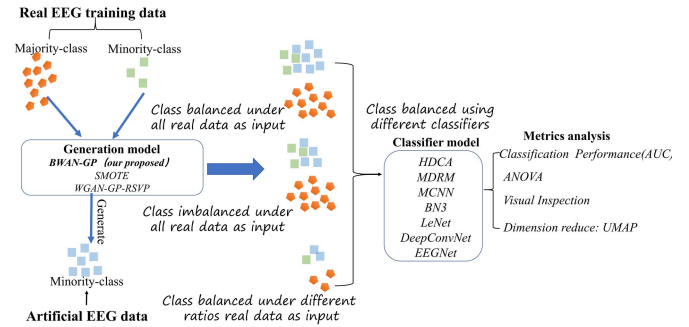- EEGNet is a CNN-based approach that achieves comparable accuracy through deep convolution and separable convolution architectures. https://github.com/vlawhern/

Fig. 5. Experimental design flow of BWGAN-GP in RSVP data augmentation.

C. Experiment Design

Our study investigates the advantages of utilizing BWGAN-GP with other augmentation methods for the RSVP classification task and metrics analysis through four approaches. Then, we validated the efficiency of the proposed algorithm. Fig. 5 shows the whole experimental procedure. First, BWGAN-GP was used to generate minority data to find the optimum proportion of data sizes for each class to produce the best classification performance. Then, with this optimal proportion, we implemented our proposed algorithm utilizing different classifiers, thus investigating the robustness of the new algorithm. Finally, the appropriate amount of artificially generated data needed to replace the original data under balanced conditions was estimated.

D. Evaluation Metrics

However, it is not as easy for multi-channel EEG signals to evaluate the quality of data generation. Some quantitative methods, such as the Fréchet inception score and the Log-likelihood distance from Gaussian mixture models [39], [40], may produce bases that are far from the truth. So, we experimented with some indicators to evaluate the quality of the generated signal.

1) The Metrics of Classification Performance: Various classification performance metrics were used in this study. We compared different methods using the area under the curve (AUC) [41], Balanced-Accuracy [42], F1-score and Cohen's kappa coefficient [43] to evaluate the methods of data augmentation.

TABLE I
HYPERPARAMETERS OF THE GENERATE NETWORK MODEL

Algorithm	Epochs	Batch Size	Input Noise	Learning Rate	Autoencoder			K Neighbors	Number of Generator	Number of Discriminator
					Epochs	Learning Rate	Batch size			
BWGAN-GP	100	128	120	0.0001	80	0.0001	256	/	1	5
WGAN-GP-RSVP [19]					/				1	1
SMOTE [11]	/							10	/	/

2) Analysis of Variance (ANOVA): Using ANOVA, differences in classification performance were compared among various algorithms. Statistical significance was defined as $p < 0.05$ [44].

3) Visual Inspection: Data were evaluated directly through visual inspection to examine the similarity of the generated data with the original data in a single channel. The quality of the data generated can be fairly evaluated.

4) Uniform Manifold Approximation and Projection (UMAP): High-dimensional samples of augmented data are mapped to a two-dimensional space by UMAP, which enables visual observation of the distribution after the data augmentation [45].

V. RESULTS AND DISCUSSION

A. Experimental Setup and Parameters

In this experiment, we compared a total of three methods of data augmentation: SMOTE [11], WGAN-GP-RSVP [19], BWGAN-GP, and EEG signals without augmentation as a baseline—original (imbalanced). The hyperparameters of the generation model are shown in Table. I. The method of SMOTE is a general method for solving CIPs in many areas. To our knowledge, WGAN-GP-RSVP is the only method of generating multi-channel RSVP EEG data. So, these methods can more convincingly evaluate the performance of our proposed generative model.

We conducted a five-fold cross-validation strategy to evaluate classification performance. All data were downsampled at 64 Hz for later experiment use in the augmentation method, and the data were normalized (mean of 0 and variance of 1). All blocks in each session were unified into one dataset during the experiments. We divided the data in the dataset according to 80% and 20% of the training and test sets. The training data were used to generate artificial EEG data using different augmentation methods. The testing set was used for classifier training.

The batch size of the training classifier was set to 128, λ_{gp} was set to 1.0, and the network was optimized with Adam optimizer. The time window number of the HDCA was set to 4. In the last experiment, the training set was fed with the augmented model in different proportions to explore how much data needs to be generated to meet real-world needs. In contrast, the test set remained the same as in all experiments. All models were tested and trained on a computer with an NVIDIA 1080Ti GPU, Xeon E5-2643 v4 CPU, and 32G RAM.

Fig.6 denotes the BWGAN-GP training process. Fig.6(a) shows that the discriminator loss of BWGAN-GP did not perform well for the first 20 epochs and increased the loss.
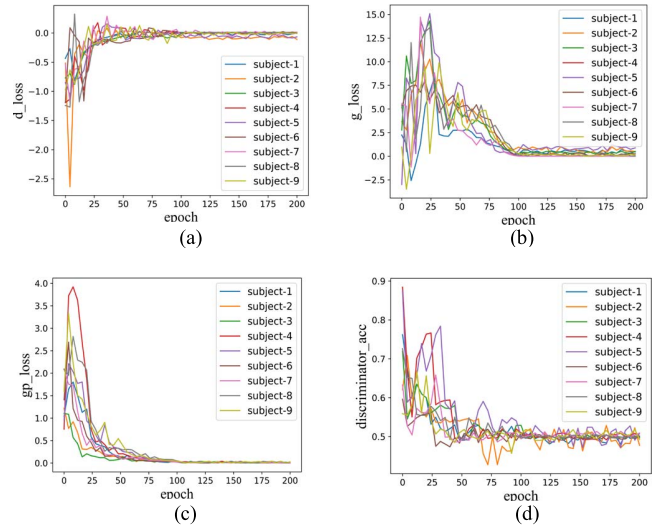

Fig. 6. BWGAN-GP model training visualization. (a) The loss of discriminator, (b) The loss of generator, (c) The loss of gradient penalty, (d) The accuracy of the discriminator.

However, the loss of all subjects decreased rapidly after 20 epochs and converged after approximately 40 epochs. Fig.6(b) shows that generator loss converged rapidly after approximately 100 epochs. Both of these losses gradually approached zero. Fig.6(c) shows the loss of the gradient penalty (GP). The GP is the part of the loss which meets the Lipschitz continuity condition and makes the descent smoother. The loss of GP close to zero indicates that the discriminator's parameters are limited in a small distribution, thus effectively speeding up the convergence rate and stabilizing the training more stable. Fig.6(d) shows the discriminator's accuracy, the value close to 0.5, which indicates that the discriminator cannot distinguish between the generated and the real data. All of the diagrams illustrate that BWGAN-GP had successfully reached Nash equilibrium.

B. Overall Classification Performance Improvement of BWGAN-GP With the Same Classifier

We used EEGNet to evaluate the classification performance using different data augmentation methods to ensure a fair comparison. The results are shown in Table. II. All augmentation methods added numerous minority-class data points to the original majority-class data, thus achieving class balance.

In Table. II, the average AUC of the BWGAN-GP algorithm was $94.43\pm2.18\%$, $\pm$ means standard deviation, which was higher than that of the WGAN-GP-RSVP

Subject	Original (Imbalanced)	SMOTE[11]	WGAN-GP-RSVP[19]	BWGAN-GP
1	0.9321	0.9423	0.9432	**0.9681**
2	0.8495	0.8694	0.8835	**0.9135**
3	0.9015	0.8712	0.9132	**0.9345**
4	0.9123	0.9417	0.9246	**0.9560**
5	0.8561	0.8651	0.8852	**0.9084**
6	0.9257	0.9407	0.9511	**0.9673**
7	0.9429	0.9097	0.9215	**0.9528**
8	0.9304	0.9259	0.9351	**0.9662**
9	0.9127	0.9174	0.9190	**0.9315**
Average	0.9070± 0.0313	0.9093± 0.0307	0.9196± 0.0220	**0.9443± 0.0218**

Fig. 8. Comparison of the real (FPz channel) and generated minority-class EEG in the single-channel in subject 1. (a) Real signal trial.452, (b) Gen signal trial.631, (c) Real signal trial.1439, (d) Gen signal trial.80, (e) Real signal trial.1979, (f) Gen signal trial.4. The blue colour represents the real signal (1st row), and the green colour represents the generated signal (2nd row).
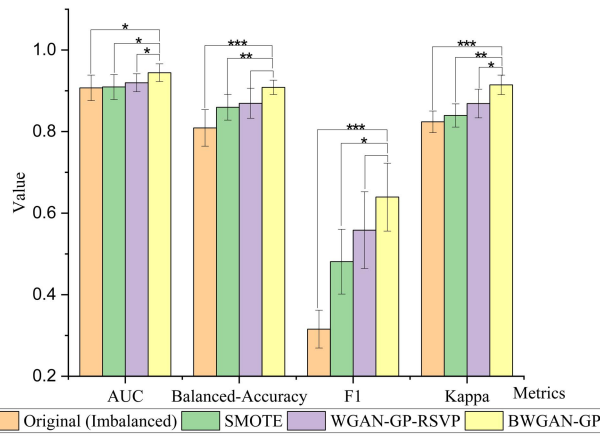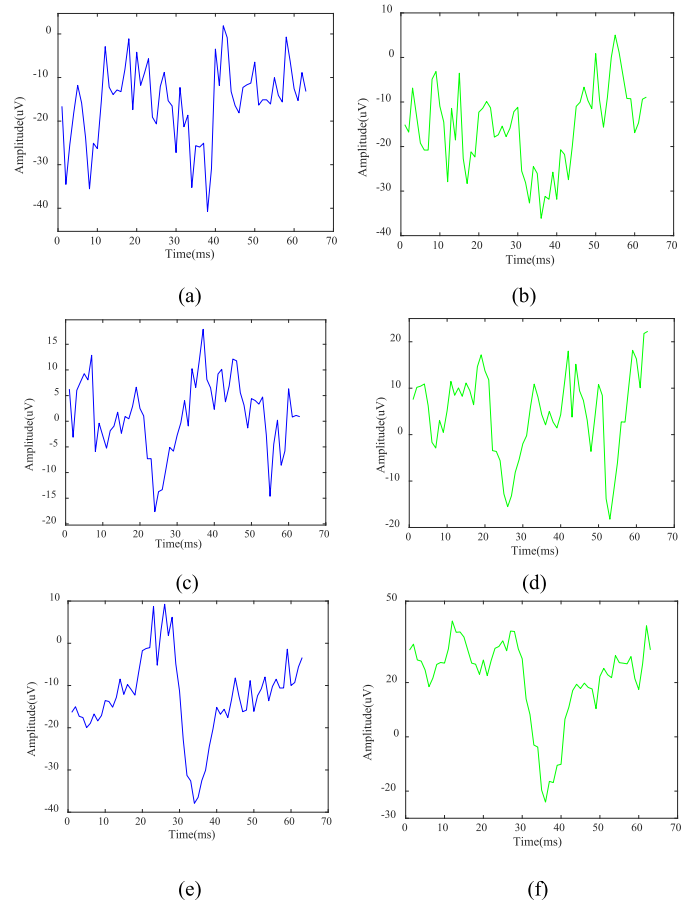

Fig. 7. Classification performance indices for different data augmentation methods, the * sign indicates statistical significance between the marked method and BWGAN-GP in each group (*, p < 0.05, **, p < 0.01, ***, p < 0.001).

(91.96±2.20%), SMOTE (90.93±3.07%) and original (imbalanced) (90.70±3.13%). It suggests that our proposed augmentation model can improve the performance of the RSVP classifier and implies that BWGAN-GP can generate higher quality minority-class data.

As seen in Table II, both WGAN-GP-RSVP and BWGAN-GP obtained better AUC values than SMOTE, indicating that implementing a GAN is a more effective method for data augmentation when performing RSVP tasks. SMOTE does not perform well as there was original (imbalanced) in Subject 3 and Subject 7, due to this interpolation method having only a change in amplitude. The generated minority-class signal is recognized as non-target due to its low amplitude. Appendix Table VI∼VIII also shows the Balanced-Accuracy, F1-score and kappa values under different enhancement methods.

The overall results are presented in Fig.7. It can be seen that the data augmentation method significantly improved the values of Balanced-Accuracy, F1-score and kappa value, despite the slight change in AUC. This indicates that the data augmentation method can alleviate the challenges posed by unbalanced data and improve the classification ability of EEGNet. The ANOVA results showed statistical differences in AUC between the different data augmentation models with EEGNet. The results of comparison of BWGAN-GP with the original(imbalanced) was F = 7.65, P < 0.05, SMOTE was F = 6.93, P < 0.05, and WGAN-GP-RSVP was F = 5.10, P < 0.05. The ANOVA result of kappa value has a significant difference, while the Balanced-Accuracy and F1-score have no significant difference between BWGAN-GP and WGAN-GP-RSVP. It indicates that there is not much difference between BWGAN and WGAN-GP-RSVP in augmenting data on some subjects under the EEGNet classifier.

This study also uses the visual inspection of RSVP under a single channel to explain the quality of the generated minority class EEG signals. Experiments were conducted to compare the real EEG and the artificial data generated by BWGAN-GP for Subject 1. Fig. 8 represents the real EEG signal (FPz channel) and the generated EEG signal corresponding to the target image.

The channel of Fpz was selected because this channel is important for the RSVP task [46], which shows that BWGAN-GP can generate a higher-quality EEG signal in Subject 1. Fig. 8 shows that the proposed model can learn the
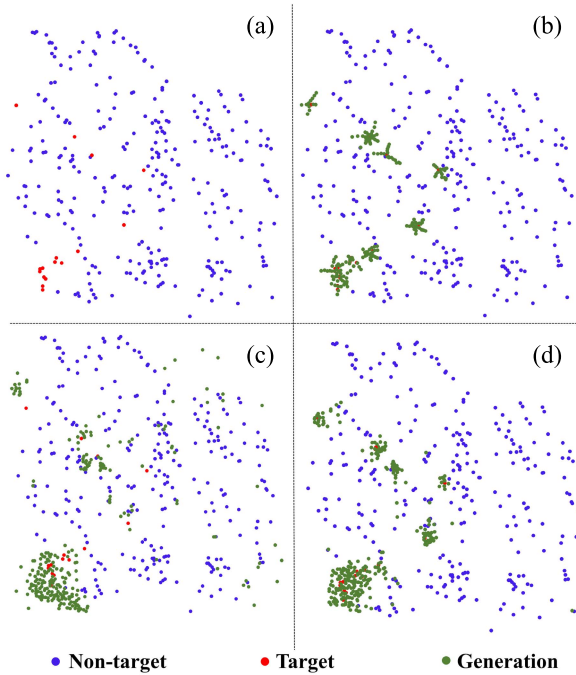
Fig. 9. An example of Real and generated data visualizations of different RSVP data augmentation methods. (a) EEG sample from Real data (no augmentation), (b) EEG samples from SMOTE generated and real, (c) EEG samples from WGAN-GP-RSVP generated and real, (d) EEG samples from BWGAN-GP generated and real. The Red colour points represent the target, blue colour points represent the non-target, and green colour points represent the generated target(minority-class).

distribution trend of real minority-class data and generate new artificial data rather than simply replicate the original data, thus overcoming the problem of the replicated operator causing mode collapse. These results offer promise in exploring directions in EEG generation.

To compare the performance of the different augmentation methods, we used the UMAP flow structure [45], which mapped the multi-channel EEG signals into two-dimensional space, as shown in Fig. 9.

In Fig. 9, 1024 random points from all data used UMAP for visualization were similar to t-SNE. However, UMAP can simultaneously compare the distribution of different algorithms and effectively preserve more details [45], [47], as shown in Fig.9. The raw EEG signal is shown in Fig. 9(a) revealed that the RSVP dataset class was unbalanced. The distribution of the target data points was very sparse, while the non-target data point was a large number. In Fig. 9 (b), in SMOTE method, most of the generated points are surrounded by the red colour points but performed poorly in the classifier, possibly due to the low diversity of the solutions. Lashgria *et al.* reported that the SMOTE augmentation method is weak for EEG generation [8]. Fig. 9(c) and Fig. 9(d) show the artificial EEG signals generated from the WGAN-GP-RSVP and BWGAN-GP. BWGAN-GP can capture real minority-class data features generate most artificial synthetic data around raw data points. In contrast, WGAN-GP-RSVP cannot generate a better initial solution because it does not discriminate between fake and minority real data, which
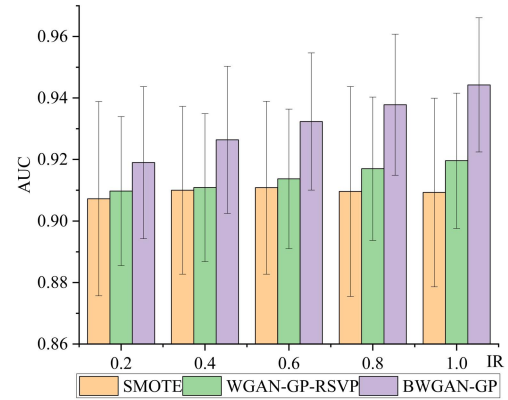

Fig. 10. The average AUC value of data augmentation methods for the EEGNet classifier at different IR values.

leads to its inability to generate higher quality minority-class EEG data.

C. Effect of the Imbalance Rate (IR) in Different Data Augmentation Methods

We explored the performance of three augmentation methods under different class proportions while still using EEGNet as the classifier. The IR (Imbalance Rate) value were 0.2, 0.4, 0.6, 0.8 and 1. In this experiment, all training samples in the real EEG were used. The original EEG and the generated minority-class data formed a new minority-class dataset to compare the majority-class real data. Fig. 10 indicates the performance of EEGNet under different proportions of two classes.

Fig. 10 shows that when the IR was 0.6, SMOTE had the best performance in itself, with the best AUC value of 91.08±1.30%. It indicates the SMOTE is already at a stopping point when the IR value reached 0.6, and the classification results achieved the best value. When IR were 0.8 and 1, the performance dropped. This result may be because SMOTE is merely a linear replication operation that cannot generate better solutions to improve the classification performance.

In contrast, WGAN-GP-RSVP and our proposed BWGAN-GP methods achieved the best classification results when the IR was 1, WGAN-GP-RSVP had the best AUC value of 91.96±2.20%, and the BWGAN-GP reached 94.43±2.19%. However, when the two data classes were extremely imbalanced, WGAN-GP-RSVP did not perform well because WGAN-GP-RSVP was less likely to produce reasonable solutions. This result also demonstrates that autoencoders in BWGAN-GP increase the possibility of generating better artificial data and improving the algorithm's classification ability. Appendix Table IX shows the classification performance of other metrics at different data class ratios.

D. The Performance of Different Augmentation Methods Under the Seven Classifiers

To further illustrate the performance of our proposed augmentation model without relying on a particular classifier, we compared the performance of the above augmentation
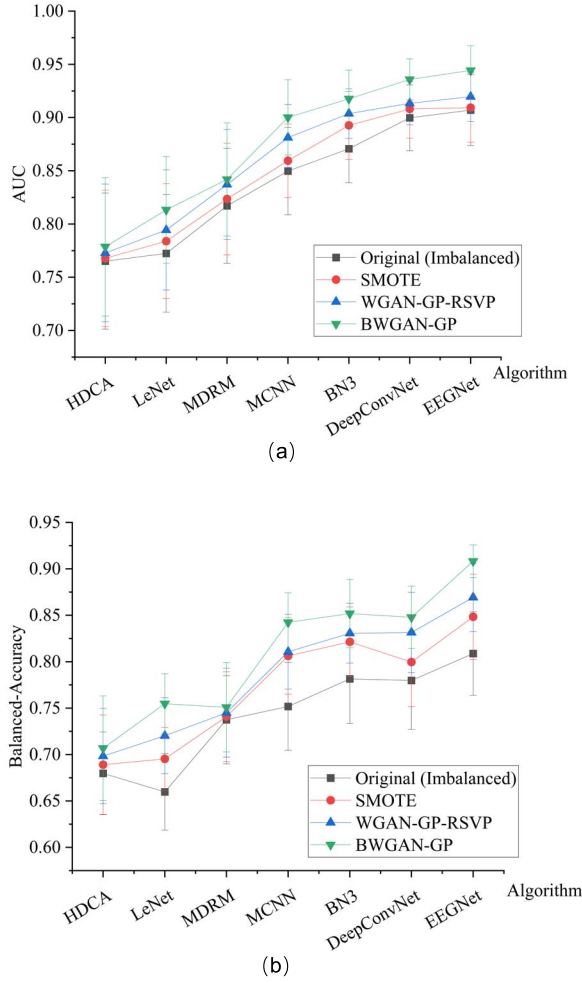
Fig. 11. F1-score and balanced-accuracy value of data augmentation methods under the seven classifiers. (a) The AUC value. (b) The balanced-accuracy value.

model on different classifiers. We evaluated the robustness of the BWGAN-GP for minority-class data augmentation using seven classifiers: HDCA, MDRM, MCNN, BN3, EEG-Net, DeepConvNet and LeNet. Fig. 11 shows the AUC and Balanced-Accuracy for different classifiers.

Fig. 11(a) shows that BWGAN-GP had the best classification performance among seven classifiers. The average of AUC value in HDCA (77.86±6.14%), LeNet (81.33±4.72%), MDRM (84.19+5.01%), MCNN (90.01±3.33%), BN3 (91.76±2.55%), DeepConvNet (93.58±1.82%) and EEGNet (94.43±2.18%). The results obtained from BWGAN-GP and EEGNet reached the best classification performance in our dataset. It may be because EEGNet was designed with a depth-wise convolution layer and a separable convolution layer, which is more specific to the EEG features extracted. It also showed that BWGAN-GP had a significant advantage over all seven classifiers. Our proposed augmentation model has good robustness.

Fig. 11(b) shows the value of Balanced-Accuracy with different classifiers. The three augmentation methods improved the value of balanced accuracy in all five deep learning

TABLE III
DIFFERENT RATIOS OF TRAINING DATA INPUT TO BWGAN-GP

Ratio	Train Data		Test Data	Generation data
	Total=3360		840	
	Target (minority class)	Nontarget (majority class)		
20%	27	645		618
40%	54	1290		1236
60%	80	2016	/	1936
80%	107	2688		2581
100%	134	3226		3092

TABLE IV
AUC VALUES OF DIFFERENT RATIOS OF TRAINING DATA WHEN USING BWGAN-GP FOR THE EEGNET CLASSIFIER

Subject	Original (Imbalanced)	20%	40%	60%	80%	100%
1	0.9321	0.5832	0.7826	0.9011	0.9355	0.9681
2	0.8495	/	0.6523	0.7935	0.8817	0.9135
3	0.9015	0.4655	0.6964	0.8628	0.9136	0.9345
4	0.9202	/	0.7048	0.8919	0.9003	0.9560
5	0.8561	0.3314	0.6491	0.8333	0.8922	0.9084
6	0.9257	0.6026	0.7274	0.9067	0.9088	0.9673
7	0.9429	/	/	0.9277	0.9307	0.9528
8	0.9304	0.6364	0.7836	0.9062	0.9277	0.9662
9	0.9127	/	0.7500	0.8793	0.9094	0.9315
Average	0.9070 ± 0.0313	0.5238 ± 0.1121	0.7183 ± 0.0491	0.8781 ± 0.0397	0.9111 ± 0.0170	0.9443 ± 0.0218

"/" represents modally collapses that cannot obtain results.

models(MCNN, BN3, DeepConvNet, LeNet and EEGNet). At the same time, machine learning methods(HDCA and MDRM) showed no significant improvement. Traditional methods require threshold adjustment to depend on subject, which does not exist in deep learning. This may be the explanation why data augmentation is significant on deep learning methods and almost fails on traditional. The F1-score and kapp values are shown in Appendix Fig. 12 and Fig. 13.

E. Effect of Different Proportions of Training Data

An attempt was made to explore whether the training data could be reduced with the generated data in a practice scenario. We designed the following experiments, in which 20%, 40%, 60% and 80% of the raw training data were taken as input in BWGAN-GP to generate artificial data, as shown in Table. III.

Table. III demonstrates the number of training data input to BWGAN-GP and the number of minority classes of data generated. For a more precise representation, an example is presented. When the training data was 20%, the number of majority class data was 645, while the minority-class training data just was 27, as input in BWGAN-GP. Therefore, the number of minority-class artificial data was 618 for balanced classes when BWGAN-GP was used.

Table. IV shows that when only approximately 60% of the training data into BWGAN-GP as input, the average AUC value reached 87.81±3.97%. In comparison, the average AUC value of the original training data was 90.7±3.13%. Only 60% of the experimental data acquisition must be

TABLE V
SUBJECT'S INFORMATION OF RSVP DATASETS

Subject	Gender	Age	Handedness	Eyesight check
sub1	Male	22	Right	Normal
sub2	Male	24	Right	Normal
sub3	Female	22	Right	Normal
sub4	Male	23	Right	Correct-to-normal
sub5	Male	26	Right	Normal
sub6	Female	23	Right	Normal
sub7	Male	25	Right	Normal
sub8	Male	22	Right	Correct-to-normal
sub9	Male	24	Right	Normal
mean	/	23.44	/	/

TABLE VI
BALANCING ACCURACY VALUES OF RSVP DATA CLASS BALANCED IN DIFFERENT AUGMENTATION METHODS ON EEGNET CLASSIFIER

Algorithm	original (imbalanced)	SMOTE[11]	WGAN-GP-RSVP[19]	BWGAN-GP
Subject 1	0.8517	0.9077	0.9159	**0.9333**
Subject 2	0.7229	0.8066	0.8265	**0.8964**
Subject 3	0.7971	0.7629	0.8514	**0.9077**
Subject 4	0.8035	0.8985	0.8726	**0.9108**
Subject 5	0.7602	0.8378	0.8253	**0.8728**
Subject 6	0.8292	0.8770	0.9280	**0.9268**
Subject 7	0.8673	0.8251	0.8540	**0.9105**
Subject 8	0.8303	0.8657	0.8953	**0.9140**
Subject 9	0.8171	0.8534	0.8542	**0.9024**
Average	0.8088±0.042	0.8483±0.043	0.8692±0.035	0.9083±0.016

obtained when allowing a slight reduction in the AUC. This result is meaningful because it can reduce subjects' training time through generating minority class artificial data from BWGAN-GP.

Starting from 80% of the training data, the average AUC was 91.11±1.70%, which exceeded that of the imbalanced original data. It indicates that the balanced new dataset was better than the original dataset regarding classification performance. The result may be because BWGAN-GP learns class labels and generates high-quality minority-class data, which alleviates CIPs. Overall, the proposed method can improve the classification performance of the RSVP task.

We also can see from Table. IV that the proposed generative model often mode collapses under the training conditions of a few minority-class data samples since these data do not provide sufficient information for the algorithm to function correctly. Hence, the availability of a certain amount of data is a precondition to ensure that the generative model can be trained effectively. Other metrics of classification performance are shown in Appendix Table. X.

F. The Limitations and Prospects of Our Work

This study demonstrates that the proposed data augmentation method is promising for improving RSVP classification task performance, yet there are still several limitations:

1) Qualitative Criteria: Qualitative criteria for generating artificial data quality in multi-channel EEG data are still lacking, and some criteria in the image domain are not suitable for EEG generation quality studies. Furthermore, research on some quantitative measurement methods would be more helpful in generating high-quality multi-channel EEG signals.

2) Single Structure: Our proposed BWGAN-GP still draws on the WGAN-GP-RSVP network structure, and we will consider combining the proposed model with a more efficient network structure in the future.

3) Mode Collapse Still Exists: Despite the improved AUC obtained for the overall classification, the above method, although having improved overall classification results, still suffers from mode collapse in the case of a tiny amount of data. In feature work, combining other techniques with GAN provides alternative solutions to CIP, such as cost-sensitive

TABLE VII
F1-SCORE VALUES OF RSVP DATA CLASS BALANCED IN DIFFERENT AUGMENTATION METHODS ON EEGNET CLASSIFIER

Algorithm	original (imbalanced)	SMOTE[11]	WGAN-GP-RSVP[19]	BWGAN-GP
Subject 1	0.3531	0.6161	0.6379	**0.7526**
Subject 2	0.2196	0.3893	0.3496	**0.5543**
Subject 3	0.2974	0.4421	0.5000	**0.6073**
Subject 4	0.3211	0.5947	0.6129	**0.6885**
Subject 5	0.2732	0.3594	0.4812	**0.4922**
Subject 6	0.3187	0.5047	0.6615	**0.7123**
Subject 7	0.3930	0.4595	0.5603	**0.6761**
Subject 8	0.3368	0.4862	0.6365	**0.7064**
Subject 9	0.3279	0.4790	0.5847	**0.5629**
Average	0.3156±0.046	0.4812±0.080	0.5583±0.094	0.6392±0.083

TABLE VIII
KAPPA VALUES OF RSVP DATA CLASS BALANCED IN DIFFERENT AUGMENTATION METHODS ON EEGNET CLASSIFIER

| Algorithm	original (imbalanced)	SMOTE[11]	WGAN-GP-RSVP[19]	BWGAN-GP
Subject 1	0.8555	0.8909	0.9178	**0.9512**
Subject 2	0.7834	0.8106	0.8147	**0.8915**
Subject 3	0.8032	0.7954	0.8517	**0.9038**
Subject 4	0.8297	0.8846	0.8677	**0.9211**
Subject 5	0.7923	0.8019	0.8375	**0.8760**
Subject 6	0.8091	0.8661	0.9333	**0.9435**
Subject 7	0.8621	0.8441	0.8562	**0.9058**
Subject 8	0.8464	0.8596	0.8792	**0.9368**
Subject 9	0.8326	0.8507	0.86	**0.8994**
Average	0.8238±0.027	0.8449±0.033	0.8686±0.035	0.9144±0.024
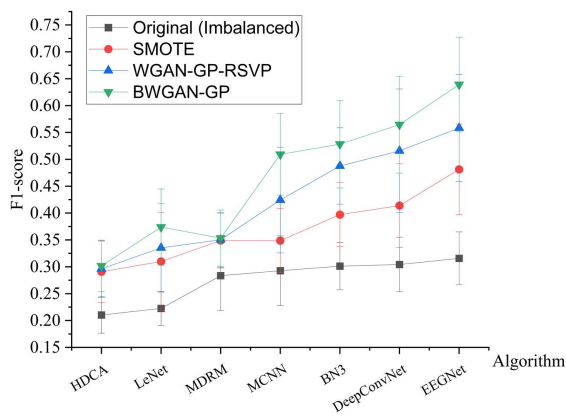
Fig. 12. F1-score value of data augmentation methods on seven classifiers.

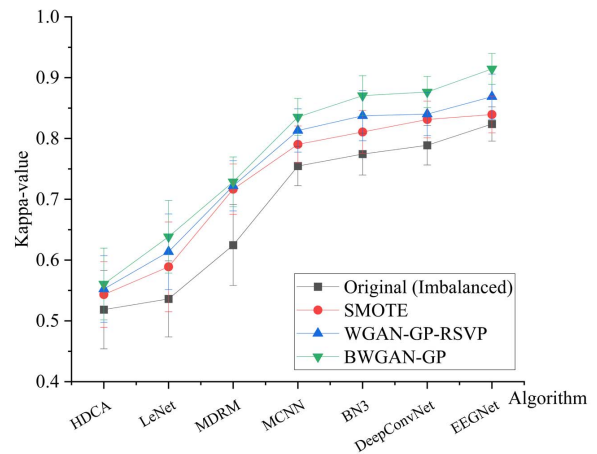Fig. 13. Kappa value of data augmentation methods on seven classifiers.

learning [48], moving the threshold [49], and downsampling methods [50] are also worth investigating. We believe that the BWGAN-GP model proposed in this study can help improve the RSVP task's classification performance and generate better quality artificial EEG data. This work can also be extended to related BCI paradigms, such as P300 signals, face

TABLE IX
METRICS VALUES OF DIFFERENT AUGMENTATION METHODS ON EEGNET CLASSIFIER AT DIFFERENT IR VALUE

IR	Metrics	SMOTE [11]	WGAN-GP-RSVP [19]	BWGAN-GP
0.2	AUC	0.9073±0.015	0.9097±0.017	0.9190±0.023
	BA	0.8257±0.027	0.8442±0.033	0.8672±0.020
	F1-score	0.4620±0.092	0.4969±0.049	0.5567±0.025
	kappa	0.8286±0.016	0.8356±0.021	0.8515±0.028
0.4	AUC	0.9100±0.010	0.9109±0.016	0.9264±0.019
	BA	0.8578±0.037	0.8500±0.046	0.8766±0.029
	F1-score	0.4826±0.080	0.5189±0.037	0.5740±0.024
	kappa	0.8335±0.025	0.8423±0.024	0.8764±0.021
0.6	AUC	0.9108±0.013	0.9137±0.015	0.9323±0.015
	BA	0.8607±0.036	0.8604±0.048	0.8893±0.023
	F1-score	0.4861±0.087	0.5311±0.032	0.6020±0.030
	kappa	0.8369±0.024	0.8526±0.031	0.9022±0.008
0.8	AUC	0.9096±0.0091	0.9170±0.016	0.9378±0.023
	BA	0.8604±0.032	0.8628±0.048	0.8971±0.014
	F1-score	0.4821±0.086	0.5465±0.035	0.6195±0.024
	kappa	0.8409±0.022	0.8597±0.031	0.8988±0.0050
1.0	AUC	0.9093±0.031	0.9196±0.022	0.9443±0.022
	BA	0.8483±0.030	0.8692±0.043	0.9083±0.016
	F1-score	0.4812±0.080	0.5583±0.037	0.6392±0.083
	kappa	0.8449±0.033	0.8686±0.035	0.9144±0.024

TABLE X
AVERAGE METRICS CLASSIFICATION PERFORMANCE VALUES OF DIFFERENT PROPORTIONS OF
TRAINING DATA TO GENERATE EEG SAMPLES BY BWGAN-GP ON EEGNET CLASSIFIER

Ratios	0.2	0.4	0.6	0.8	1.0
AUC	0.5238±0.112	0.7183±0.049	0.8781±0.040	0.9111±0.017	0.9443±0.021
Balanced-Accuracy	0.2492±0.158	0.5753±0.097	0.7224±0.062	0.8549±0.026	0.9083±0.016
F1-score	0.0732±0.105	0.2712±0.053	0.4162±0.088	0.5246±0.106	0.6392±0.083
Kappa	-0.0949±0.221	0.4096±0.064	0.5362±0.058	0.8119±0.045	0.9144±0.022

recognition and other class imbalanced applications of EEG signal classification.

VI. CONCLUSION

This study aimed to develop a novel data augmentation approach to address CIPs in the RSVP task. We first presented the data augmentation method generated for minority-class data and evaluated the RSVP task's classification performance. The results indicated that the BWGAN-GP approach contributed to better overall classification performance. In most cases, the class-wise performance improved, as reflected by a better classification metrics value. We also investigated the effect of artificial data generated on the experimental calibration stage. The results suggested that the proposed BWGAN-GP could alleviate CIPs by generating the minority-

class data of EEG signals, further improving classification performance and reducing calibration time in the RSVP tasks.

APPENDIX

See, Tables V–X and Figs. 12 and 13.

REFERENCES

[1] N. Bigdely-Shamlo, A. Vankov, R. R. Ramirez, and S. Makeig, "Brain activity-based image classification from rapid serial visual presentation," *IEEE Trans. Neural. Syst. Rehabil. Eng.*, vol. 16, no. 5, pp. 432–441, Oct. 2008.

[2] S. Lees *et al.*, "A review of rapid serial visual presentation-based brain–computer interfaces," *J. Neural. Eng.*, vol. 15, no. 2, 2018, Art. no. 021001.

[3] H. Cecotti, J. Sato-Reinhold, J. L. Sy, J. C. Elliott, M. P. Eckstein, and B. Giesbrecht, "Impact of target probability on single-trial EEG target detection in a difficult rapid serial visual presentation task," in *Proc. EMBC*, Aug. 2011, pp. 6381–6384.

[4] A. D. Gerson, L. C. Parra, and P. Sajda, "Cortical origins of response time variability during rapid discrimination of visual objects," *NeuroImage*, vol. 28, no. 2, pp. 342–353, Nov. 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811905004180

[5] Z. Wang, G. Healy, A. F. Smeaton, and T. E. Ward, "A review of feature extraction and classification algorithms for image RSVP-based BCI," in *Signal Processing and Machine Learning for Brain-Machine Interfaces* (Control, Robotics & Sensors). IET Digital Library, 2018, ch. 12, pp. 243–270. [Online]. Available: https://digital-library.theiet.org/content/books/10.1049/pbce114e_ch12, doi: 10.1049/PBCE114E_ch12.

[6] Z. Wang, G. Healy, A. F. Smeaton, and T. E. Ward, "Spatial filtering pipeline evaluation of cortically coupled computer vision system for rapid serial visual presentation," *Brain-Comput. Interfaces*, vol. 5, no. 4, pp. 132–145, Oct. 2018.

[7] Q. Wen, L. Sun, X. Song, J. Gao, X. Wang, and H. Xu, "Time series data augmentation for deep learning: A survey," *CoRR*, vol. abs/2002.12478, pp. 1–8, Feb. 2020.

[8] E. Lashgari, D. Liang, and U. Maoz, "Data augmentation for deep-learning-based electroencephalography," *J. Neurosci. Methods*, vol. 346, Dec. 2020, Art. no. 108885. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027020303083

[9] J. Pan *et al.*, "Physics-based generative adversarial models for image restoration and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2449–2462, Jul. 2021.

[10] C. Athanasiadis, E. Hortal, and S. Asteriadis, "Audio–visual domain adaptation using conditional semi-supervised generative adversarial networks," *Neurocomputing*, vol. 397, pp. 331–344, Jul. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219316170

[11] T. Lee, M. Kim, and S.-P. Kim, "Data augmentation effects using borderline-SMOTE on classification of a P300-based BCI," in *Proc. BCI*, Feb. 2020, pp. 1–4.

[12] A. Bhattacharyya and R. B. Pachori, "A multivariate approach for patient-specific EEG seizure detection using empirical wavelet transform," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2003–2015, Sep. 2017.

[13] S. K. Satapathy, S. Dehuri, and A. K. Jagadev, "SMOTE and ABC optimised RBF network for coping with imbalanced class in EEG signal classification," *Int. J. Med. Inform.*, vol. 10, no. 3, pp. 215–234, 2018.

[14] S. Haradal, H. Hayashi, and S. Uchida, "Biosignal data augmentation based on generative adversarial networks," in *Proc. EMBC*, Jul. 2018, pp. 368–371.

[15] K. G. Hartmann, R. T. Schirrmeister, and T. Ball, "EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals," 2018, *arXiv:1806.01875*.

[16] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, and M. Shah, "Generative adversarial networks conditioned by brain signals," in *Proc. ICCV*, Oct. 2017, pp. 3430–3438.

[17] Q. Zhang and Y. Liu, "Improving brain computer interface performance by data augmentation with conditional deep convolutional generative adversarial network," *CoRR*, vol. abs/1806.07108, pp. 1–4, Jun. 2018.

[18] Y. Luo, L.-Z. Zhu, Z.-Y. Wan, and B.-L. Lu, "Data augmentation for enhancing EEG-based emotion recognition with deep generative models," *J. Neural Eng.*, vol. 17, no. 5, Oct. 2020, Art. no. 056021.

[19] S. Panwar, P. Rad, T.-P. Jung, and Y. Huang, "Modeling EEG data distribution with a Wasserstein generative adversarial network to predict RSVP events," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 8, pp. 1720–1730, Aug. 2020.

[20] S. Panwar, P. Rad, J. Quarles, and Y. Huang, "Generating EEG signals of an RSVP experiment by a class conditioned Wasserstein generative adversarial network," in *Proc. SMC*, Oct. 2019, pp. 1304–1310.

[21] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608018302107

[22] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002.

[23] S. Wang, L. L. Minku, and X. Yao, "A learning framework for online class imbalance learning," in *Proc. CIEL*, Apr. 2013, pp. 36–45.

[24] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[25] L. Wang, W. Chen, W. Yang, F. Bi, and F. R. Yu, "A state-of-the-art review on image synthesis with generative adversarial networks," *IEEE Access*, vol. 8, pp. 63514–63537, 2020.

[26] T. M. Sutter, I. Daunhawer, and J. E. Vogt, "Multimodal generative learning utilizing Jensen–Shannon-divergence," *CoRR*, vol. abs/2006.08242, pp. 1–26, Jun. 2020.

[27] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.

[28] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. ICML*, vol. 70, D. Precup and Y. W. Teh, Eds., Aug. 2017, pp. 214–223. [Online]. Available: http://proceedings.mlr.press/v70/arjovsky17a.html

[29] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," *CoRR*, vol. abs/1704.00028, pp. 1–20, Mar. 2017.

[30] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and A. C. I. Malossi, "BAGAN: Data augmentation with balancing GAN," *CoRR*, vol. abs/1803.09655, pp. 1–9, Mar. 2018.

[31] G. Huang and A. H. Jafari, "Enhanced balancing GAN: Minority-class image generation," *Neural Comput. Appl.*, pp. 1–10, Jun. 2021, doi: 10.1007/s00521-021-06163-8.

[32] L. C. Parra *et al.*, "Spatiotemporal linear decoding of brain state," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 107–115, Jan. 2008.

[33] A. Barachant and M. Congedo, "A plug & play P300 BCI using information geometry," 2014, *arXiv:1409.0107*.

[34] R. Manor and A. B. Geva, "Convolutional neural network for multi-category rapid serial visual presentation BCI," *Frontiers Comput. Neurosci.*, vol. 9, p. 146, Dec. 2015.

[35] M. Liu, W. Wu, Z. Gu, Z. Yu, F. F. Qi, and Y. Li, "Deep learning based on batch normalization for P300 signal detection," *Neurocomputing*, vol. 275, pp. 288–297, Jan. 2018.

[36] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Aug. 2017.

[37] T. Wang, C. Lu, G. Shen, and F. Hong, "Sleep apnea detection from a single-lead ECG signal with automatic feature-extraction through a modified LeNet-5 convolutional neural network," *PeerJ*, vol. 7, p. e7731, Sep. 2019.

[38] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013.

[39] A. Obukhov and M. Krasnyanskiy, "Quality assessment method for GAN based on modified metrics inception score and Fréchet inception distance," in *Software Engineering Perspectives in Intelligent Systems*, R. Silhavy, P. Silhavy, and Z. Prokopova, Eds. Cham: Springer, 2020.

[40] A. Grover, M. Dhar, and S. Ermon, "Flow-GAN: Bridging implicit and prescribed learning in generative models," *CoRR*, vol. abs/1705.08868, pp. 1–12, May 2017.

[41] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016786550500303X

[42] W. Wei, S. Qiu, X. Ma, D. Li, B. Wang, and H. He, "Reducing calibration efforts in RSVP tasks with multi-source adversarial domain adaptation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 11, pp. 2344–2355, Nov. 2020.

[43] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[44] H. M. Park, "Comparing group means: T-tests and one-way ANOVA using STATA, SAS, R, and SPSS," Univ. Inf. Technol. Services, Center Stat. Math. Comput., Indiana Univ., Bloomington, IN, USA, Work. Paper. [Online]. Available: http://www.indiana.edu/~statmath/stat/all/ttest

[45] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.

[46] M. Xu, Y. Chen, D. Wang, Y. Wang, L. Zhang, and X. Wei, "Multi-objective optimization approach for channel selection and cross-subject generalization in RSVP-based BCIs," *J. Neural Eng.*, vol. 18, no. 4, Jun. 2021, Art. no. 046076, doi: 10.1088/1741-2552/ac0489.

[47] S. Arora, W. Hu, and P. K. Kothari, "An analysis of the t-SNE algorithm for data visualization," in *Proc. COLT*, 2018, pp. 1455–1462.

[48] M. K. Siddiqui, X. Huang, R. Morales-Menendez, N. Hussain, and K. Khatoon, "Machine learning based novel cost-sensitive seizure detection classifier for imbalanced EEG data sets," *Int. J. Interact. Des. Manuf.*, vol. 14, no. 4, pp. 1491–1509, Dec. 2020.

[49] D.-K. Tran, T.-H. Nguyen, and T.-N. Nguyen, "Detection of EEG-based eye-blinks using a thresholding algorithm," *Eur. J. Eng. Technol. Res.*, vol. 6, no. 4, pp. 6–12, May 2021.

[50] N. Islah, J. Koerner, R. Genov, T. A. Valiante, and G. O'Leary, "Machine learning with imbalanced EEG datasets using outlier-based sampling," in *Proc. EMBC*, Jul. 2020, pp. 112–115.