frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 13 June 2025
DOI 10.3389/fmars.2025.1586237

OPEN ACCESS

EDITED BY
Chao Chen,
Suzhou University of Science and Technology, China

REVIEWED BY
Jianmin Yang,
Sun Yat-sen University, China
Tingt Lyu,
Ocean University of China, China

*CORRESPONDENCE
Shengwen Gong
gsw780604@126.com

RECEIVED 02 March 2025
ACCEPTED 21 May 2025
PUBLISHED 13 June 2025

CITATION
Jiang J, Cheng W, Gong S and Wang J (2025) A deep learning-based data augmentation method for marine mammal call signals. *Front. Mar. Sci.* 12:1586237. doi: 10.3389/fmars.2025.1586237

A deep learning-based data augmentation method for marine mammal call signals

Jiaming Jiang[1,2], Wanlu Cheng[1,2], Shengwen Gong[1,2]* and Jingjing Wang[1,2]

[1]School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, China, [2]Shandong Key Laboratory of Deep Sea Equipment Intelligent Networking, Qingdao, Shandong, China

In marine ecology research, it is crucial to accurately identify the marine mammal species active in the target area during the current season, which helps researchers understand the behavioral patterns of different species and their ecological environment. However, the difficulty and high cost of collecting marine mammal calls, coupled with limited publicly available datasets, result in insufficient data for support, making it difficult to obtain accurate and reliable identification results. To address this problem, we propose MarGEN, a deep learning-based augmentation method for marine mammal call signal data. This method processed the call data into Mel spectrograms, then designed a self-attention conditional generative adversarial network to generate new samples of Mel spectrograms that were highly similar to the real data, and finally reconstructed them into call signals using WaveGlow. The classification experiments on the calls of four Marine mammals show that MarGEN significantly enriches the diversity and volume of the data, increasing the classification accuracy of the model by an average of 4.7%. The method proposed in this paper greatly promotes marine ecological protection and sustainable development, while effectively advancing research progress in bionic covert underwater acoustic communication technology.

KEYWORDS

marine ecology, marine mammal call signals, MarGEN, deep learning, data augmentation, self-attention conditional generative adversarial network

1 Introduction

Marine mammal calls serve as important ecological signals, carrying a wealth of behavioral and environmental information. Accurately recognizing marine mammal calls not only contributes to species monitoring and conservation but also facilitates the assessment of the health of the marine environment. At the same time, accurate recognition of marine mammal calls also has important military application value, bionic covert underwater acoustic communication technology embeds secret information into marine mammal calls to improve the security of underwater communication Qiao et al.

(2018); Ma et al. (2024), the working principle diagram of this technology is shown in Figure 1. The prerequisite for realizing this technique is to accurately identify the active marine mammals in the target sea area in the current season, so as to select the appropriate calls for bionic communication. Currently, deep learning-based recognition classification offers the most effective results Shi et al. (2023); Dong et al. (2020), but its training demands a large number of data samples as support Li et al. (2021). However, the current limited availability of marine mammal call data significantly reduces the performance of deep learning-based recognition and classification models. Buda et al. (2018). Therefore, increasing the number and diversity of Marine mammal call data has become the key to improving the recognition accuracy.

Data augmentation is a method to expand the size of datasets Khan et al. (2024), which not only enhances the predictive ability of classification models but also provides diversity-rich call signals for bionic communication. Currently, data augmentation methods have performed well in the field of computer vision, which has attracted researchers to focus on its application in the field of audio Sun et al. (2024); Xu et al. (2024).

The cropping method Garcea et al. (2023) obtains multiple cropped sub-data by sliding the audio sequence over a sliding window. Scaling methods Lie and Chang (2006) are implemented by adjusting the audio amplitude or frequency, amplitude scaling is achieved by multiplying all the elements of the time series by some constant, and frequency scaling is achieved by changing the sampling rate of the audio signal. Adding some random noise to the original data can also increase data diversity Kishk and Dhillon (2017), but inappropriate noise may mask important signal features and lead to degradation of model performance. The random oversampling technique Wei et al. (2022) achieves data augmentation by randomly selecting samples for replication. The Synthetic Minority Oversampling Technique (SMOTE) Azhar et al. (2023) generates new samples by interpolating the minority class sample, which improves the problem of unbalanced data distribution. SpecAugment Kim et al. (2024) is a data augmentation method that operates on the audio spectrum. By distorting or masking the spectrogram of the speech signal, the data diversity during model training is increased. Experiments have demonstrated that this method can significantly reduce the word error rate and improve the robustness of the model in speech recognition tasks. This method performs data augmentation on individual sequences, utilizing only the nature of the sequence itself and not taking the overall distribution of the dataset into account.

In the wake of rapid advancements in artificial intelligence, researchers have started to apply deep learning techniques to data augmentation. Yan employed a convolutional neural network model for data augmentation of music in a rhythm game. He took the first 30 seconds of 16 piano arrangements as input, generated additional material that mimicked the original styles through Jukebox and extended them to 60 seconds for data enhancement. However, this method is time-consuming because it generates only one sample at a time Yan (2024). The adversarial training model of Generative Adversarial Networks (GANs) Goodfellow et al. (2020); Wu et al. (2020) gives them excellent generative results. Significant advancements and outcomes have been achieved in the generation of high-resolution and realistic images, which has a wide range of potentials in the field of computer vision and image generation, which also encourages researchers to apply GANs In the field of audio generation. Some researchers have applied GANs to environmental sounds and footstep signal generation with better results Bahmei et al. (2022); Chakraborty and Kar (2023). At present, among the published methods, no researcher has applied the data augmentation method based on GANs to marine mammal call signals.

We proposed MarGEN, a data augmentation method for marine mammal call signals based on audio transformation and a Self-Attention Conditional Generative Adversarial Network (SACGAN). It can effectively enrich the number and diversity of marine mammal call signals and greatly improve the recognition accuracy of the model. The main contributions of this paper are as follows.

1. We proposed a novel method for generating marine mammal call signals, marking the first application of generative adversarial networks in the field of marine mammal call signal data augmentation.
2. We designed a self-attention conditional generation adversarial network for generating new samples that are highly similar to the Mel spectrograms Hong and Suh (2023); Ustubioglu et al. (2023) of real marine mammal calls. The network innovatively added conditional variables representing marine mammal species and self-attention modules and replaced some of the convolutional layers with improved Inception blocks, which significantly improved the model performance and the quality of the generated samples.
3. In order to analyze the performance of our generated call signals, we performed classification experiments and compared them with baseline datasets, which demonstrated the superiority of our method in terms of prediction accuracy.
4. The proposed method can effectively extend the existing marine mammal sound database. It greatly advances the research progress in marine mammal conservation and bionic covert underwater acoustic communication technology. It also provides a reference method for the generation of other types of sound.

2 Data preprocessing

The dataset used in this study comes from the Watkins Marine Mammal Sound Database Sayigh et al. (2016), which provides a variety of call clips of marine mammals recorded in real marine environments. In this study, four marine mammal calls, which are widely distributed in China's sea area and have a relatively large

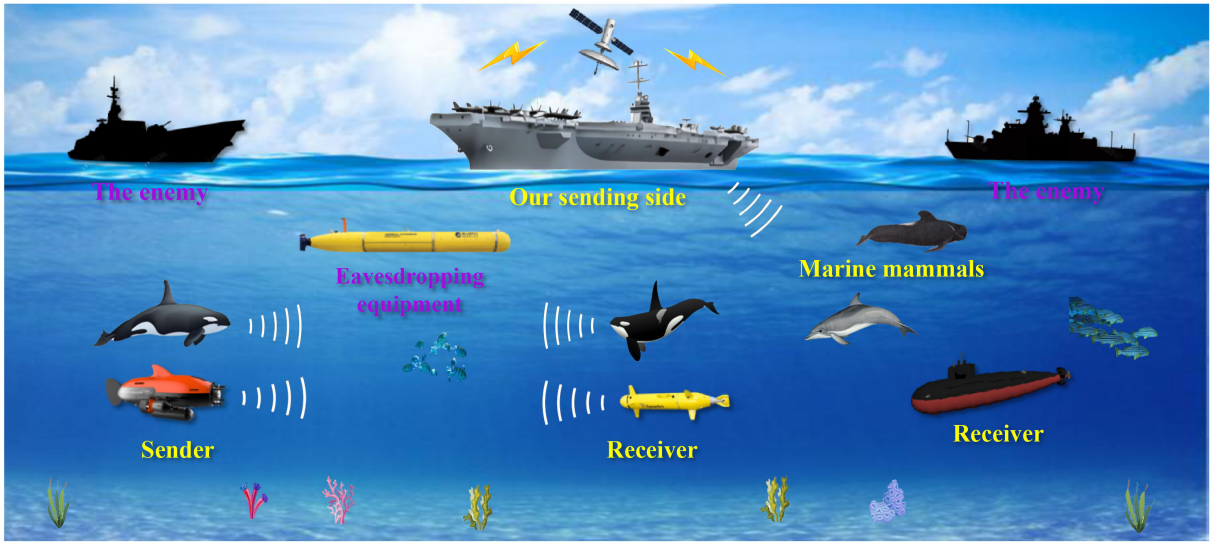

FIGURE 1
Working principle diagram of bionic covert underwater acoustic communication technology.

amount of data, were selected for downloading. After clipping, denoising, resampling, and other operations, 4190 samples with a duration of 1 second are finally obtained and labeled. The distribution and characteristics of the samples are shown in Table 1.

3 MarGEN method

The overall flowchart of the MarGEN method is illustrated in Figure 2, consisting of three main steps. In the first step, due to the large number of audio sampling points of marine mammal calls, resulting in many network parameters and training difficulties, and given that generative adversarial networks are more mature in the image generation domain, we converted the marine mammal call audio files into the form of spectrograms that are more suitable for machine learning to understand the characterization. In the second step, we innovatively designed the SACGAN, whose generator and discriminator engaged in continuous adversarial training until Nash equilibrium Lv et al. (2024) was reached, thereby generating new samples that closely resembled the original images. In the final step, the generated spectrogram was converted into audio signals using WaveGlow Prenger et al. (2019).

TABLE 1 Distribution and characteristics of samples.

Species Name	Abbreviation	Sample Size	Sampling Rate
Killer Whale	KW	1394	48000Hz
Humpback Whale	HW	908	48000Hz
Pilot Whale	PW	1165	48000Hz
Bottlenose Dolphin	BND	723	48000Hz

3.1 Feature extraction

The features of different call samples behave similarly in the time domain but differ significantly in the frequency domain. Therefore, the feature representations chosen in this study were mainly frequency domain features. First, the original signal is analyzed in time-frequency by Short-Time Fourier Transform (STFT) to extract its local frequency domain information. The STFT can effectively capture the spectral changes of the signal in the time dimension. On this basis, Mel frequency cepstrum coefficient (MFCC) based analysis can be further frequency transformed according to the auditory perception of the human ear, thus preserving the key features of the signal. Its accuracy and computational efficiency are better than other representations in the speech recognition task. Therefore, the Mel spectrogram was chosen as the feature representation in this study. The expression for the MEL frequency is shown in Equation 1:

$$M = 2595\log_{10}(1 + \frac{f}{700}) \quad (1)$$

Where M is the frequency in Mel and f is the frequency in Hz, 128 Meier filters are used in this study.

3.2 Self-attention conditional generative adversarial network

GANs consist of a generator and a discriminator. The generator receives random noise and outputs newly generated data samples, while the discriminator is responsible for determining whether the received data is real or generated by the generator. The generator and discriminator engage in adversarial training, which ultimately generates new data that closely resembles real data.

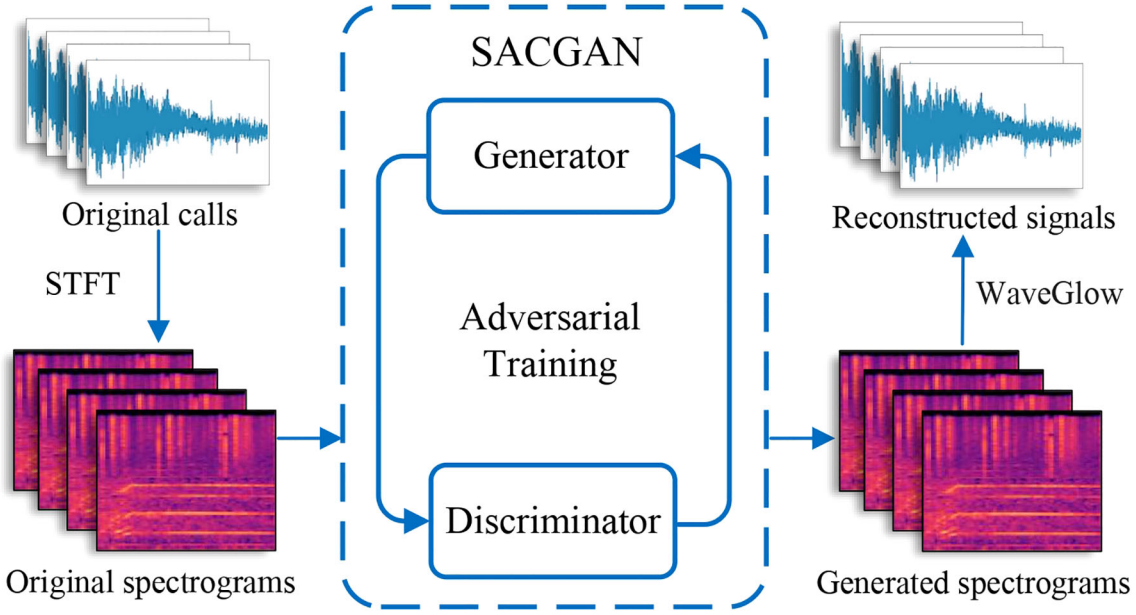

FIGURE 2
Flowchart of MarGEN Method.

We innovatively designed SACGAN, which introduced conditional variables representing marine mammal species and self-attention modules based on generative adversarial networks. Additionally, traditional convolutional neural networks consist of multiple layers of convolutional layers stacked on top of each other, which tends to lead to overfitting as well as difficulty in updating the gradient. The network we designed utilized improved Inception blocks, a structure that combines convolutional kernels of various sizes within the same layer to capture multi-scale information, thereby enhancing the capability of feature extraction.

The specific network structure of SACGAN is shown in Figure 3A. In the generator network structure, the discrete labeled variables were converted to continuous vectors through the Embedding layer, which were spliced with random noise to help the model better understand the input data. The network structure of the Inception block is shown in Figure 3B. We improved its second branch by decomposing a 3x3 convolution into a 1x3 convolution and a 3x1 convolution, further reducing the number of parameters and computational complexity. The residual block consisted of the deconvolution layer, the batch normalization layer, and the activation layer. In the residual block, the gradient information was propagated by means of skip connections to help the generator better recover the image details. A self-attention module was added between two residual blocks to enhance the generator's ability to produce specific content under given conditions, thereby improving generation precision. In the discriminator network structure, the residual block consisted of the convolution layer, the batch normalization layer, and the activation layer. The pooling layer was responsible for reducing the feature dimensions and extracting the main information of the features. We added a self-attention module after the pooling operation to help the model compensate for information loss, ensuring that the model retained some detailed information while capturing the main features. The formula expression of the self-attention mechanism is shown as Equation 2:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

Where Q denotes the query matrix, K denotes the key matrix, V denotes the value matrix, K^T is the transpose matrix of K, and d_k denotes the dimension length.

In addition, the model used the loss function of WGAN-GP Pu et al. (2022); Zhu et al. (2023) to prevent the pattern collapse problem during training. A gradient penalty term was added to the discriminator loss function to ensure that the discriminator function satisfied the Lipschitz continuity constraint, avoiding the problem of gradient explosion or gradient disappearance during the training process and enhancing the convergence speed of the model. The generator loss function is shown as Equation 3:

$$L(G) = -E_{z\sim Pz}[D(G(z|y))] \tag{3}$$

Where Pz denotes the data distribution of samples generated by the generator, z is the randomly sampled noise vector in Pz, and y is the condition variable.

The discriminator loss function is shown as Equation 4:

$$\begin{aligned} L(D) = {} & E_{x\sim pr}[D(x|y)] - E_{z\sim Pz}[D(G(z|y))] \\ & + \lambda E_{\hat{x}\sim P\hat{x}}\left[(\| \nabla_{\hat{x}} D(\hat{x}) \|_2 - 1)^2\right] \end{aligned} \tag{4}$$

Where p_r denotes the data distribution of the real sample, x is the sample in p_r, λ is the gradient penalty term weight, $\lambda E_{\hat{x}\sim P\hat{x}}[(\| \nabla_{\hat{x}} D(\hat{x}) \|_2 - 1)^2]$ is the gradient penalty term, $\hat{x}$ is the stochastic

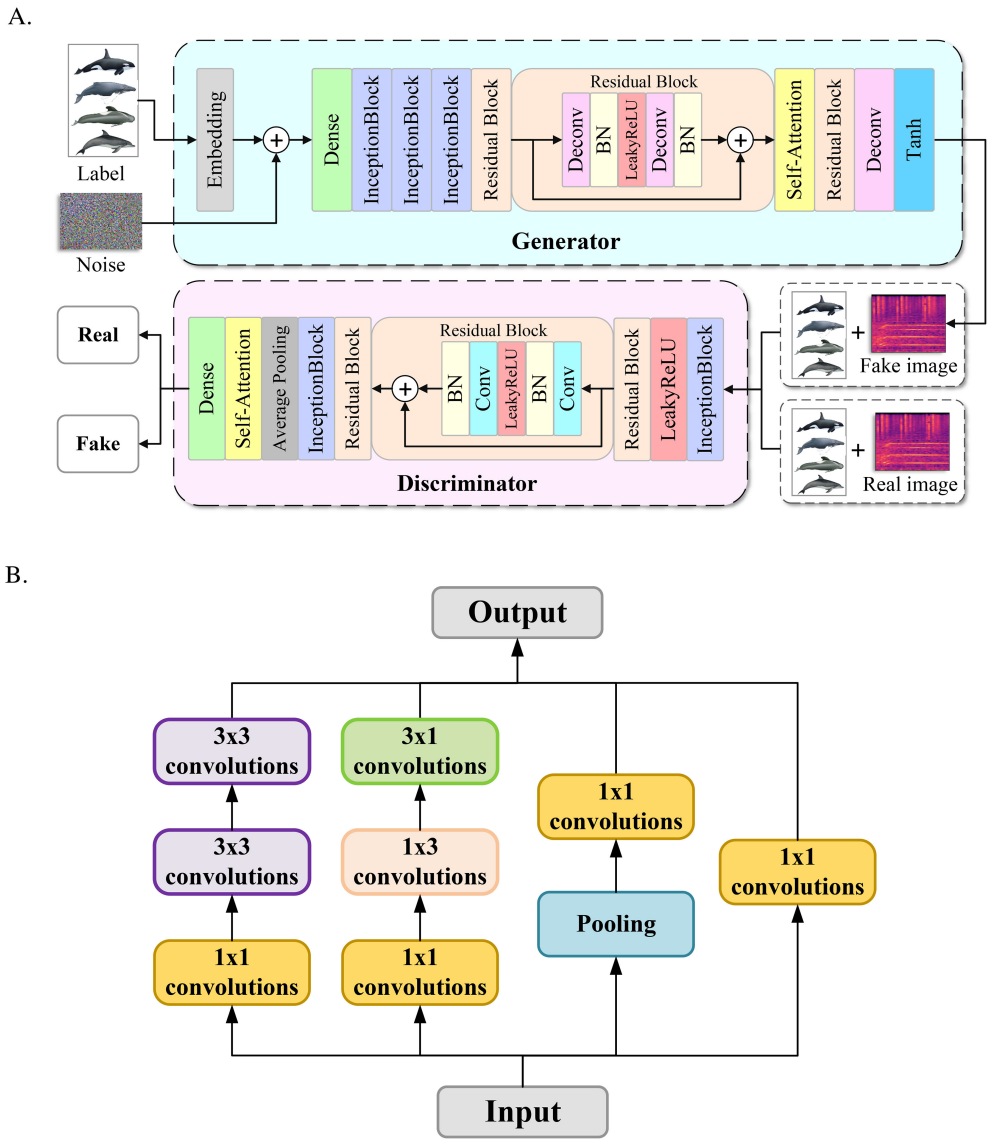

FIGURE 3
(A) Network structure of the self-attention conditional generative adversarial network; **(B)** structure of the inception block.

interpolation between the real sample and the generated sample, $P\hat{x}$ denotes the sampling distribution of the gradient penalty term, and $\| \nabla_{\hat{x}} D(\hat{x}) \|_2$ denotes the gradient parameter of $\hat{x}$, which ensures that the gradient paradigm of the discriminator function is close to 1 and satisfies the Lipschitz constraint.

3.3 Audio signal reconstruction

We used WaveGlow to reconstruct the Mel spectrogram samples generated by SACGAN into audio signals. The model can accurately learn the probability distribution of the audio data and acquire longrange information, resulting in better generation quality and generalization ability. In addition, WaveGlow supports GPU parallel operation, significantly accelerating the audio synthesis speed.

4 Experiment

We designed generation experiments and classification experiments. The generation experiments were used to increase the number and diversity of existing datasets. The classification experiments were used to validate the effectiveness of the MarGEN method.

4.1 Generation experiment

The experimental programming language was Python 3.9, and the network construction was built using Pytorch 1.10 deep learning framework. We trained SACGAN with 4,190 Mel spectrograms of marine mammal calls, setting the labels for killer whale calls to 0, humpback whale calls to 1, pilot whale calls to 2, and bottlenose

dolphin calls to 3. The experimental dataset was divided into a training set and a test set in an 8:2 ratio, with a learning rate set at 1e-4; the batch size was 64; the number of training epochs was 2000. An alternating training strategy was adopted, in which the discriminator was trained six times corresponding to the training of the generator once.

Figure 4 shows an example of Mel spectrograms generated using the SACGAN. As shown, SACGAN can generate high-quality Mel spectrograms. In this experiment, a total of 1755 samples of Mel spectrograms of marine mammal calls were generated using the SACGAN.

4.2 Classification experiment

To verify the effectiveness and superiority of the MarGEN method, this experiment trained the same ResNet classification model on two datasets separately for performance evaluation. Table 2 presents the number of samples in the two datasets and their specific distribution. Among them, OD is a dataset consisting of the original marine mammal call signals. MD is a mixed dataset consisting of the original marine mammal call signals and the call signals obtained using the MarGEN method. The 'Factor' column indicates the ratio between the total number of samples after data enhancement (original samples plus generated samples) and the number of original samples. For example, for bottlenose dolphin, a factor of 2.0 indicates that after augmentation, the dataset contains twice as many samples as the original dataset (original: 723 samples, augmented: 1,446 samples). The dataset was divided using 5-fold cross-validation, in which the entire sample was randomly divided into five non-overlapping subsets, each of which accounted for approximately 20% of the entire dataset. In each round of cross-validation, four of them were selected as the training set. The remaining one as the validation set, and a total of five rounds were executed, with a different validation subset being used in each round. The final results are aggregated by the average of the metrics obtained from the 5 rounds of experiments to ensure the stability and generalization ability of the model. At the same time, it is necessary to make sure that the ratio of original data and generated data in the training and validation sets is consistent. The learning rate for the experiments was set to 1e-4; the batch size was 32; and the training epochs were 150.

Figure 5A illustrates the confusion matrix of the classification model trained using OD, while Figure 5B illustrates the confusion matrix of the classification model trained using MD. In these matrices, the diagonal elements represent the correct classification rate for each category, while the off-diagonal elements reflect the misclassifications between species. Through comparison, it can be found that killer whales showed high classification accuracy in both confusion matrices, probably due to the even spacing between fundamental and harmonic frequencies in their calls, regular frequency bands, often accompanied by high-energy dominant frequency components, and clear transverse stripe structure on Mel spectrograms, which had good discriminability, and thus were easy to be accurately recognized by the model. The classification effect of the bottlenose dolphin was significantly improved after the data enhancement. However, the classification accuracy was still at the lowest level, which may be attributed to the following reasons: on the one hand, broad-snouted dolphin calls are complex and diverse, with a large frequency span, which increases the difficulty of identification; On the other hand, broad-snouted dolphins have the smallest number of original samples among the four categories, and the

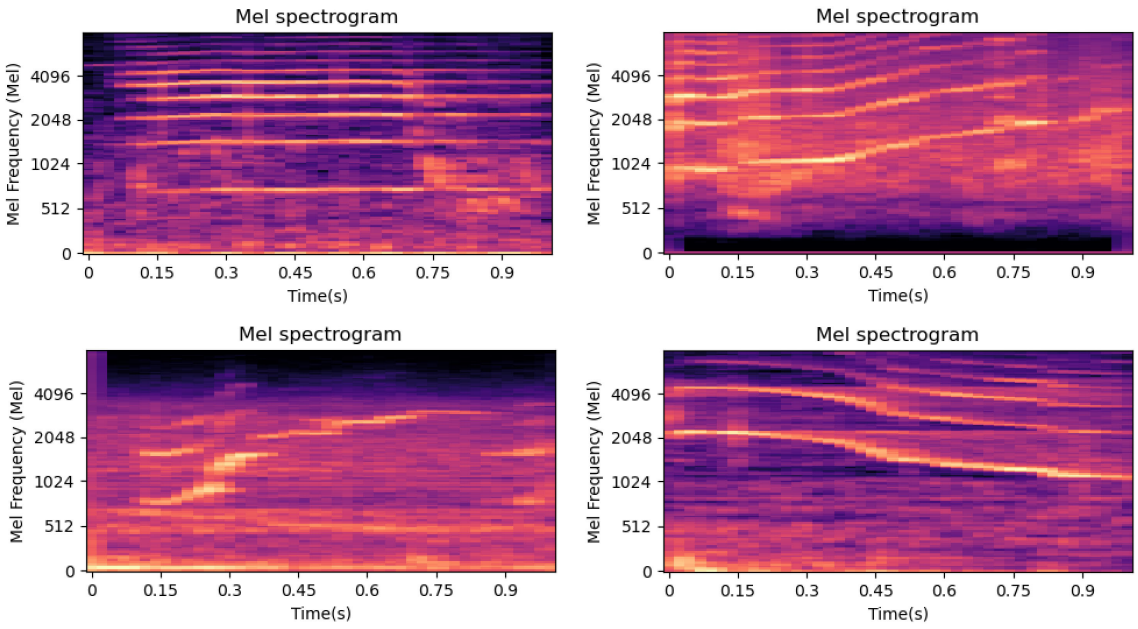

FIGURE 4
Mel spectrograms generated using SACGAN.

TABLE 2 The number of samples and their specific distribution for the two datasets.

Species Name	Abbreviation	OD	Factor	MD
Killer Whale	KW	1394	1.1	1533
Humpback Whale	HW	908	1.6	1452
Pilot Whale	PW	1165	1.3	1514
Bottlenose Dolphin	BND	723	2.0	1446

model does not learn enough of its features at the early stage of training. Although data augmentation greatly mitigates the training bias caused by the uneven samples, there are still some recognition challenges.

In general, the model trained using the MD dataset achieves a higher recognition accuracy for marine mammal calls and exhibits a significantly reduced gap in classification performance between species. These results demonstrate that the proposed MarGEN data augmentation method effectively enhances the model's generalization ability and mitigates the problem of class imbalance.

We selected four classical deep learning models for classification experiments to demonstrate that the MarGEN method can optimize the performance of multiple models. In the experiments, we calculated the Accuracy, Precision, Recall, and F1 Score of the models to comprehensively evaluate their classification performance. We calculated the accuracy, precision, recall, and F1 score of these models in the experiments. The corresponding formulas are as shown in Equations 5–8:

$$Accuracy = \frac{True\ \ Positives + True\ \ Negatives}{True\ \ Positives + False\ \ Positives + True\ \ Negatives + False\ \ Negatives} \tag{5}$$

$$Precision = \frac{True\ \ Positives}{True\ \ Positives + False\ \ Positives} \tag{6}$$

$$Recall = \frac{True\ \ Positives}{True\ \ Positives + FalseNegatives} \tag{7}$$

$$F1\ \ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

Where the F1 score is the reconciled average of precision and recall, which can comprehensively evaluate the classification performance.

Table 3 shows that the accuracy of the classification models trained using MD increased by an average of 4.7%, in which the accuracy of the ResNetSE model increased by 5.7% from 90.93% to 96.63%; the F1 score increased by an average of 5.75%, proving that

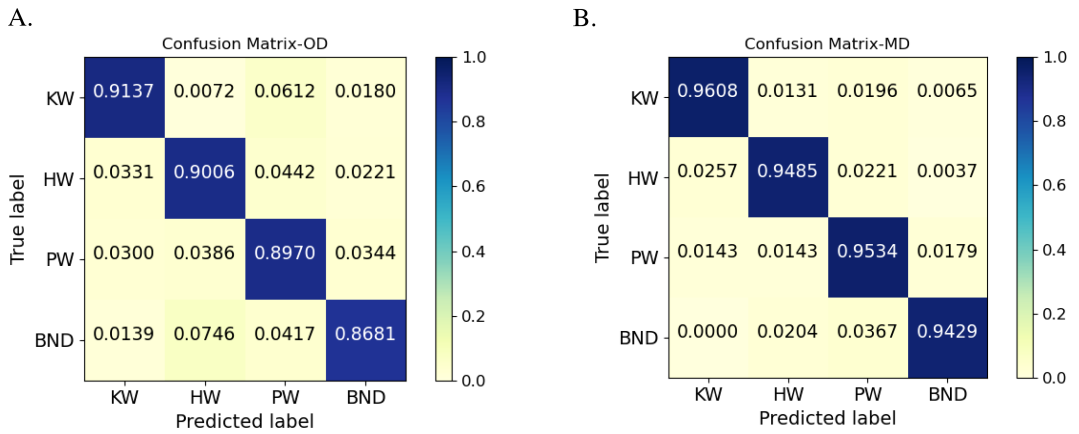

FIGURE 5
(A) Confusion matrix of the classification model trained with OD; **(B)** Confusion matrix of the classification model trained with MD.

TABLE 3 Comparison of performance evaluation indexes for two datasets applied to different classification models.

Model	Accuracy (OD/MD) (%)	Precision (OD/MD) (%)	Recall (OD/MD) (%)	F1 Score (OD/MD) (%)
CNN	89.98/94.37	88.11/95.20	90.06/94.85	89.07/95.02
Res2Net	91.77/95.37	94.42/96.39	91.37/96.08	92.87/96.23
ResNetSE	90.93/96.63	88.03/96.65	86.81/94.29	87.42/95.46
RNN	88.90/94.03	87.08/92.68	89.70/95.34	88.37/93.99

the MarGEN method can significantly improve the performance of multiple deep learning models on the marine mammal call signal recognition task.

5 Conclusion

We have innovatively presented MarGEN, which can effectively realize the high similarity generation of marine mammal call signals and improve their recognition accuracy. First, we designed SACGAN, which can generate Mel spectrograms that are highly similar to the original data, and then we converted the Mel spectrograms into call signals using WaveGlow. The experimental results demonstrated that after using the MarGEN method, the recognition accuracy of different classification models is improved by 4.7% on average, and the F1 score is improved by 5.75% on average. The proposed method in this paper greatly promotes marine ecological protection and sustainable development, and at the same time, it also greatly promotes the research progress of bionic covert hydroacoustic communication technology, which is of great strategic significance. In the future, we will further extend the applicability of the study. On the one hand, we will extend the MarGEN method to more marine species to verify its generalizability in multi-species identification tasks; on the other hand, we will also explore the migration ability of the model under fewer samples to enable the identification and study of data-scarce species.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

Author contributions

JJ: Software, Writing – original draft, Writing – review & editing, Methodology. WC: Data curation, Investigation, Writing – review & editing. SG: Validation, Visualization, Writing – review & editing. JW: Funding acquisition, Methodology, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported in part by the National Natural Science Foundation of China under Grants 62171246 , 62101298 and U24A20215.

Acknowledgments

The authors are grateful for the Watkins Marine Mammal Sound Database website for providing us with the audio of marine mammal calls needed for the experiment, and the support of Python.

Conflict of interest

The authors 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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

Azhar, N. A., Pozi, M. S. M., Din, A. M., and Jatowt, A. (2023). An investigation of smote based methods for imbalanced datasets with data complexity analysis. *IEEE Trans. Knowledge Data Eng.* 35, 6651–6672. doi: 10.1109/TKDE.2022.3179381

Bahmei, B., Birmingham, E., and Arzanpour, S. (2022). Cnn-rnn and data augmentation using deep convolutional generative adversarial network for environmental sound classification. *IEEE Signal Process. Lett.* 29, 682–686. doi: 10.1109/LSP.2022.3150258

Buda, M., Maki, A., and Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* 106, 249–259. doi: 10.1016/j.neunet.2018.07.011

Chakraborty, M., and Kar, S. (2023). Enhancing person identification through data augmentation of footstep-based seismic signals. *IEEE Signal Process. Lett.* 30, 1642–1646. doi: 10.1109/LSP.2023.3327650

Dong, S., Zhuang, Y., Yang, Z., Pang, L., Chen, H., and Long, T. (2020). Land cover classification from vhr optical remote sensing images by feature ensemble deep learning network. *IEEE Geosci. Remote Sens. Lett.* 17, 1396–1400. doi: 10.1109/LGRS.2019.2947022

Garcea, F., Serra, A., Lamberti, F., and Morra, L. (2023). Data augmentation for medical imaging: A systematic literature review. *Comput. Biol. Med.* 152, 106391. doi: 10.1016/j.compbiomed.2022.106391

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Commun. ACM* 63, 139–144. doi: 10.1145/3422622

Hong, G., and Suh, D. (2023). Mel spectrogram-based advanced deep temporal clustering model with unsupervised data for fault diagnosis. *Expert Syst. Appl.* 217, 119551. doi: 10.1016/j.eswa.2023.119551

Khan, A. A., Chaudhari, O., and Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Syst. Appl.* 244, 122778. doi: 10.1016/j.eswa.2023.122778

Kim, K.-H., Oh, K.-H., Ahn, H.-S., and Choi, H.-D. (2024). Time–frequency domain deep convolutional neural network for li-ion battery soc estimation. *IEEE Trans. Power Electron.* 39, 125–134. doi: 10.1109/TPEL.2023.3309934

Kishk, M. A., and Dhillon, H. S. (2017). Stochastic geometry-based comparison of secrecy enhancement techniques in d2d networks. *IEEE Wireless Commun. Lett.* 6, 394–397. doi: 10.1109/LWC.2017.2696537

Li, L., Qiao, G., Liu, S., Qing, X., Zhang, H., Mazhar, S., et al. (2021). Automated classification of tursiops aduncus whistles based on a depth-wise separable convolutional neural network and data augmentation. *J. Acoustical Soc. America* 150, 3861–3873. doi: 10.1121/10.0007291

Lie, W.-N., and Chang, L.-C. (2006). Robust and high-quality time-domain audio watermarking based on low-frequency amplitude modification. *IEEE Trans. Multimedia* 8, 46–59. doi: 10.1109/TMM.2005.861292

Lv, Y., Liu, Y.-J., Liu, L., Yu, D., and Chen, Y. (2024). Distributed nash equilibrium searching for multi-agent games under false data injection attacks. *Neurocomputing* 570, 127134. doi: 10.1016/j.neucom.2023.127134

Ma, X., Wang, B., Tian, W., Ding, X., and Han, Z. (2024). Strategic game model for auv-assisted underwater acoustic covert communication in ocean internet of things. *IEEE Internet Things J.* 11, 22508–22520. doi: 10.1109/JIOT.2024.3382649

Prenger, R., Valle, R., and Catanzaro, B. (2019). "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (Brighton, UK: IEEE), 3617–3621. doi: 10.1109/ICASSP.2019.8683143

Pu, Z., Cabrera, D., Li, C., and de Oliveira, J. V. (2022). Vgan: Generalizing mse gan and wgan-gp for robot fault diagnosis. *IEEE intelligent Syst.* 37, 65–75. doi: 10.1109/MIS.2022.3168356

Qiao, G., Bilal, M., Liu, S., Babar, Z., and Ma, T. (2018). Biologically inspired covert underwater acoustic communication—a review. *Phys. Communication* 30, 107–114. doi: 10.1016/j.phycom.2018.07.007

Sayigh, L., Daher, M. A., Allen, J., Gordon, H., Joyce, K., Stuhlmann, C., et al. (2016). "The watkins marine mammal sound database: an online, freely accessible resource," in *Proceedings of meetings on acoustics*, vol. 27. (Melville, New York, United States: AIP Publishing). doi: 10.1121/2.0000358

Shi, J., Liu, W., Shan, H., Li, E., Li, X., and Zhang, L. (2023). Remote sensing scene classification based on multibranch fusion attention network. *IEEE Geosci. Remote Sens. Lett.* 20, 1–5. doi: 10.1109/LGRS.2023.3262407

Sun, Y., Xu, K., Liu, C., Dou, Y., Wang, H., Ding, B., et al. (2024). Automated data augmentation for audio classification. *IEEE/ACM Trans. Audio Speech Lang. Process.* 32, 2716–2728. doi: 10.1109/TASLP.2024.3402049

Ustubioglu, B., Tahaoglu, G., and Ulutas, G. (2023). Detection of audio copy-move-forgery with novel feature matching on mel spectrogram. *Expert Syst. Appl.* 213, 118963. doi: 10.1016/j.eswa.2022.118963

Wei, G., Mu, W., Song, Y., and Dou, J. (2022). An improved and random synthetic minority oversampling technique for imbalanced data. *Knowledge-based Syst.* 248, 108839. doi: 10.1016/j.knosys.2022.108839

Wu, Z., Li, J., Wang, Y., Hu, Z., and Molinier, M. (2020). Self-attentive generative adversarial network for cloud detection in high resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* 17, 1792–1796. doi: 10.1109/LGRS.2019.2955071

Xu, X., Xie, Z., Wu, M., and Yu, K. (2024). Beyond the status quo: A contemporary survey of advances and challenges in audio captioning. *IEEE/ACM Trans. Audio Speech Lang. Process.* 32, 95–112. doi: 10.1109/TASLP.2023.3321968

Yan, N. (2024). Generating rhythm game music with jukebox. *Front. Artif. Intell.* 7. doi: 10.3389/frai.2024.1296034

Zhu, G., Zhou, K., Lu, L., Fu, Y., Liu, Z., and Yang, X. (2023). Partial discharge data augmentation based on improved wasserstein generative adversarial network with gradient penalty. *IEEE Trans. Industrial Inf.* 19, 6565–6575. doi: 10.1109/TII.2022.3197839