IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 November 2022, accepted 2 December 2022, date of publication 8 December 2022,
date of current version 14 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3227646

RESEARCH ARTICLE

Bi-Sphere Anomaly Detection With Learnable Centroid for Active Sonar Classification

GEUNHWAN KIM[1] AND YOUNGMIN CHOO[2], (Member, IEEE)
[1]Department of Ocean Systems Engineering, Sejong University, Seoul 05006, Republic of Korea
[2]Department of Defense Systems Engineering, Sejong University, Seoul 05006, Republic of Korea

Corresponding author: Youngmin Choo (ychoo@sejong.ac.kr)

This work was supported in part by the Research Project funded by the Korea Research Institute of Ships and Ocean Engineering (KRISO) under Grant NTIS 1525012176; and in part by the National Research Foundation of Korea (NRF), South Korea, under Grant 2022R1A6A3A01087548.

ABSTRACT Machine learning (ML)-based approaches are desirable for discriminating targets from clutter signals to enhance the performance of active sonar systems. However, a small dataset and imbalanced data samples between the target and clutter hinder ML applications in active sonar classification. Anomaly detection (AD), which effectively exploits the imbalance, is adopted to enhance the generalization of ML-based active sonar classifiers for small and imbalanced datasets. Generally, deep AD focuses on learning a representation of normal data samples (clutter) and finding a sphere embracing normal data samples in latent space. However, abnormal samples from artificial objects (underwater targets) have similar physical experiences as normal clutter samples from geological and biological scattering objects. Therefore, it is difficult to discriminate between the target and the clutter using conventional deep AD. To overcome the problem of active sonar classification, we propose semi-supervised learning-based bi-sphere anomaly detection (BiSAD) to find two spheres, embracing target and clutter samples each, by modifying conventional deep AD. Simultaneously, BiSAD searches for the latent space where two sphere centroids locate distantly to promote generalization. In the generalization test, the receiver operating characteristic (ROC) curve of BiSAD indicates a detection probability of 0.8 at a false alarm rate of 0.01, and the area under the ROC curve was 0.989, which was superior to the conventional deep AD and supervised learning-based approaches.

INDEX TERMS Active sonar classification, machine learning, anomaly detection, sonar clutter suppression.

I. INTRODUCTION

Modern active sonar systems for anti-submarine warfare transmit sound waves and analyze received signals to estimate the information of the underwater target. However, in addition to the target echo reflected from the artificial objects, the received signal consists of the signal reflected from various geological and biological scattering objects, such as the sea surface, sea bottom, and fish school, which is called clutter. Because a clutter has physical experiences similar to those of the target echo (sound propagation and scattering in the ocean), it generates a false alarm when a matched filter is applied. Therefore, such false alarms make it difficult to detect the targets. Consequently, the performance of active sonar systems will be degraded. To overcome this problem, an active sonar-classification algorithm that distinguishes targets from clutter is desired in active sonar systems [1], [2], [3], [4].

In the past few decades, research on machine learning (ML)-based algorithms have been conducted for active sonar classification. Early studies related to active sonar classification can be found in literature from the late 1980s. Gorman et al. [5], [6] performed an experiment to distinguish between a metal cylinder and a cylindrical rock located on a sandy seabed using a shallow neural network with a normalized spectral envelope as the input. Shin et al. [7] and Andrea Trucco [8] conducted active sonar classification using a classify-before-detection strategy based on a pattern-recognition paradigm. Murphy et al. performed active sonar classification using a Gaussian-based classifier with aural

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

features that mimicked the human auditory system [9]. Seo et al. performed active sonar classification using a support vector machine with multilayer features from the range-bearing domain [10]. Tongjing Sun et al. proposed Fisher discriminant dictionary learning combining Fisher's discriminant criterion and a dictionary learning-based sparse representation classification algorithm [11].

Recently, with the rapid development of deep learning technology, there have been increasing attempts to adopt it to active sonar classification. Several studies have been conducted on active sonar classifiers based on convolutional neural networks (CNN) trained with a supervised learning approach [12], [13]. Yule Chen et al. conducted data augmentation using a generative adversarial network to overcome the problem of a supervised learning approach in the small number of samples [14]. Research on unsupervised approaches has also been conducted [15]. Wang et al. proposed a multidomain network comprising a shared network and attention modules using images from different signal processing as inputs [16].

However, the generalization performance of active sonar classifiers remains low and limited, primarily because the active sonar dataset suffers from a small number of data samples owing to the difficulty of sea trials and the confidentiality of data samples. A small number of data samples has an adverse effect on the performance of conventional ML-based algorithms, which are guaranteed to be performed when large amounts of data samples are used. Additionally, the active sonar dataset also suffers from severely imbalanced data samples between the target and clutter because the received signal contains an abundance of clutter and few target signals. Therefore, it is necessary to understand the characteristics of the active sonar dataset and adopt an appropriate approach to solve the active sonar classification problem.

Deep anomaly detection (AD) can effectively exploit the imbalanced dataset, making it suitable to enhance the generalization performance of the active sonar classifier [17]. Conventional deep AD focuses on learning a representation of normal data samples (here, clutter data samples in the active sonar dataset) and attempts to fit normal data samples in a compact sphere manifold in latent space [18]. After learning, deep AD distinguishes abnormal data samples (here, target data samples in the active sonar dataset) by measuring the distance between the centroid of the sphere and the data samples in the latent space. However, abnormal target data samples from artificial objects have similar physical experiences as normal clutter data samples from geological and biological objects. Therefore, the abnormal target data samples and the normal clutter data samples have similar characteristics, such that the abnormal target data sample may be included within the decision boundary of the normality. Consequently, it is difficult to discriminate between the target and clutter using only normal data samples.

A deep AD using anomaly data samples has also been proposed [19]; however, the generalization performance of active sonar classification remains low because prior knowledge of active sonar data samples are not fully considered.

To overcome the problem of active sonar classification and advance generalization performance, we propose semi-supervised learning-based bi-sphere anomaly detection (BiSAD), which finds two spheres, including target and clutter samples, respectively, by modifying the conventional deep AD. Simultaneously, BiSAD searches for the latent space, where the centroids of spheres are at a long distance, to increase the generalization performance.

The remainder of this paper is organized as follows. Sec. II describes the problem of active sonar targets and clutter classification. Sec. III summarizes the ML-based training and testing strategies and Sec. IV presents BiSAD for active sonar classification. Sec. V describes the preliminaries of ML-based active sonar classification. Sec. VI describes the results of the ML-based classifiers using sea experimental data that include scattered signals from underwater artificial objects. Finally, we conclude this paper in Sec. VII.

II. PROBLEM DESCRIPTION

A. OVERVIEW OF THE SCHEME OF THE ACTIVE SONAR DETECTION AND CLASSIFICATION

Fig. 1 shows a scheme for active sonar detection and classification. There are two strategies: classify-after-detection and classify-before-detect.

Fig. 1(a) depicts the classify-after-detect strategy [9]. The received beam signal is filtered using a matched filter with a replica of the transmitted pulse. A threshold was applied to detect the target signal. However, it is inappropriate to use a fixed threshold because the clutter level fluctuates according to changes in the oceanic environment, which causes the problem of detecting the target with a different false alarm rate. Therefore, a normalization algorithm that adapts the threshold to obtain a fixed false alarm rate should be applied to the matched filter output [20]. Since modern active sonar systems have a high range-bearing resolution, multiple signals can be detected for a single object. A clustering algorithm was employed to group multiple detected signals from the same object [21]. The output of clustering is called a contact.

Meanwhile, Fig. 1(b) depicts the classification-before-detect strategy [7], [8], [22]. Unlike the classify-after-detect strategy requiring pre-processing, classifiers are directly applied to the received beam signal in classify-before-detect. Therefore, the classify-before-detect strategy can utilize redundant information by extracting proper acoustic features from raw beam signals. Recent developments in ML-based algorithms have enabled the adoption of the classify-before-detection strategy in active sonar systems.

B. ACTIVE SONAR TARGET AND CLUTTER CLASSIFICATION

Figs. 2(a), (b), and (c) display examples of the results of the beamforming process, matched filter (MF), and contacts,

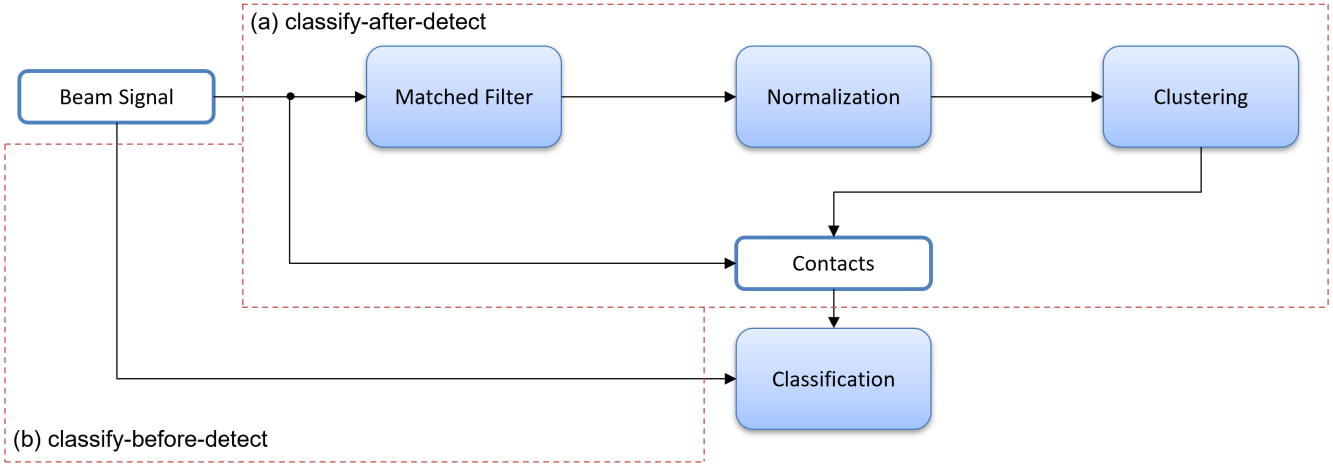


FIGURE 1. Scheme for the active sonar detection and classification. (a) Classify-after-detect strategy: matched filter, normalization, and clustering are sequentially performed on the received beam signal. The results obtained in this process are called contacts. As the contacts contain many nontargets, a classification process is required to remove them. (b) Classify-before-detect strategy: detection and classification are simultaneously performed on raw beam signal, which removes the pre-processing in the classify-after-detect.

respectively. These results are generated by sea experimental data with a single underwater target and shallow water (the specification of the sea trial data will be described in Sec.V-A in detail).

In the beamforming output of Fig. 2(a), the target signal is identified using prior information on the target location and is marked with the red arrow. Reverberation appeared in the earlier part of the time, making multiple clutter. Furthermore, strong signals (orange arrows) were observed randomly, which also resulted in clutter.

In the MF output of Fig. 2(b), it can be observed that the target signal is emphasized and the noise level is suppressed. However, multiple clutter signals remained along the target signal, primarily because many of the scattered signals have a similar experience to the target; therefore, the clutter signal is also highly correlated with the replica of MF.

In Fig. 2(c) shows the clustering results. Although MF, normalization, and clustering were applied, many contacts were observed. Classification is required because these contacts do not guarantee that they are certain targets.

Generally, classification is performed by sonar operators because they are known to be capable of distinguishing the target from a clutter [9], [23]. More specifically, the sonar operator can distinguish subtle differences between the target and clutter from a raw audio signal extracted from the selected beam signal data. This means that the human auditory system can extract aural and perceptual information from a raw audio signal. However, leaving the sonar operator solely responsible for the classification of numerous contacts is risky for human error, besides, it is slow to process and unavailable for around-the-clock surveillance. Therefore, an automatic active sonar classifier that can effectively distinguish the target from clutter using a raw audio signal is required.

III. VARIOUS ML-BASED TRAINING AND TESTING STRATEGIES

A. SUPERVISED LEARNING

ML-based approaches are commonly used to solve classification problems for various research fields [24]. Many ML-based approaches are supervised learning approaches that require a large number of labeled samples, and the class of each data sample is known. Fig. 3 illustrates the architecture of the supervised learning approach. For the labeled dataset $\chi = \{\mathbf{x}_i, y_i\}_{i=1}^{N} = \chi_0 \cup \chi_1$ where χ_0 and χ_1 are the normal and abnormal dataset, respectively, and N is the number of total data samples, the supervised learning approach can be represented as:

$$\mathbf{z} = \phi(\mathbf{x}; \mathbf{W}_\phi) \tag{1}$$

$$\mathbf{p} = a\,(f(\mathbf{z})) \tag{2}$$

where $\phi : \mathbf{x} \rightarrow \mathbf{z}$ represents an encoder function for feature learning with weight parameter $\mathbf{W}_\phi$, $\mathbf{z}$ indicates a latent vector which encodes feature of input vector, f denotes a fully-connected layer which connects latent vector and output, a represents softmax activation function, and $\mathbf{p}$ corresponds to the output that indicates the probability of each classes. Typically, the architecture of supervised learning is trained by minimizing cross-entropy loss using the gradient-descent method [24]. The anomaly score of supervised learning $s_{sup}(\mathbf{x})$ is calculated as follows:

$$s_{sup}(\mathbf{x}) = \frac{p_1}{p_0} \tag{3}$$

where p_0 and p_1 denote the element of output vector $\mathbf{p}$ which indicate normal and abnormal probabilities, respectively.

Recent networks of supervised learning approaches become deeper and more complex because the complexity of the networks has a better ability to fit dataset than shallow networks [25], [26], [27], [28]. However, the deeper and

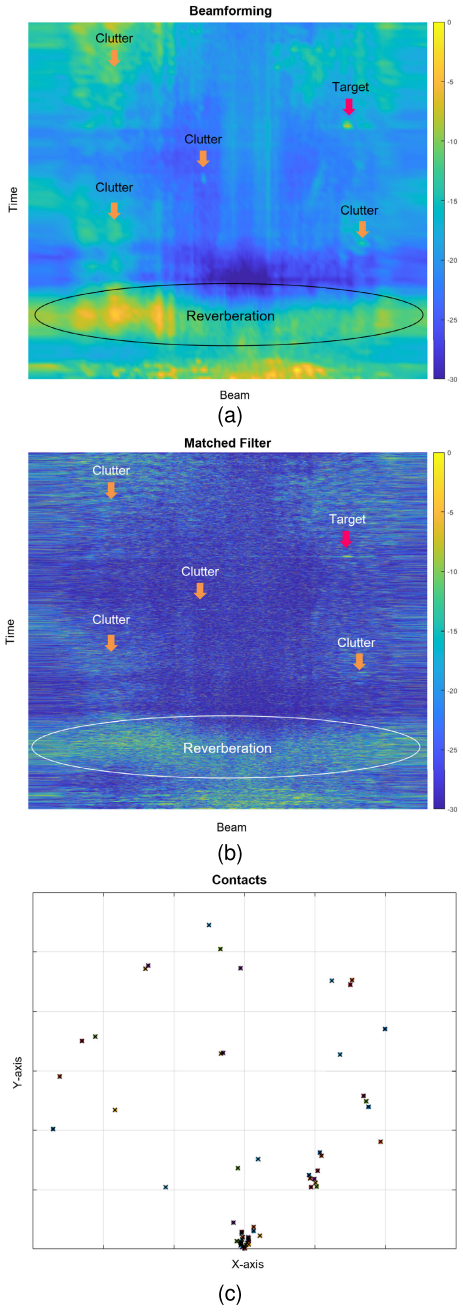


FIGURE 2. Results of the classify-after-detect strategy. (a) Beamforming output, (b) matched filter output, and (c) contacts. The beam and time domain in (a) and (b) are converted to range (X-axis) and cross-range (Y-axis) domain. In the beamforming output, it is evident that target signal (red arrow) and clutter signals (orange arrows) appeared. Clutter signals originated from reverberation and randomly located strong reflection. In the matched filter output, the target signal is emphasized owing to the effect of correlation; however, multiple clutter signals still remain. Following normalization and clustering, many contacts appear.

more complex networks require more memory size and have risk of overfitting, leading to decreased and unstable generalization performance. In general tasks of supervised learning approaches, a large size of the dataset that contains the overall data distribution is used to prevent the problem of supervised learning approaches [24]. In the active sonar

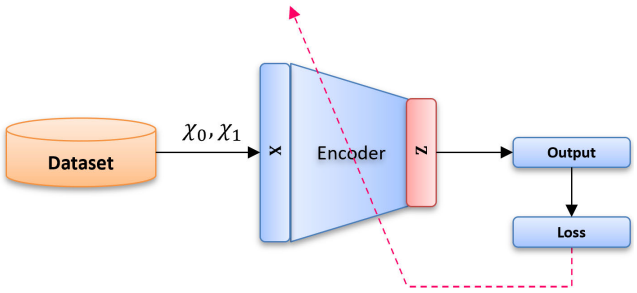


FIGURE 3. Scheme of supervised learning. Input data sample x were encoded to latent vector z using the encoder function. Output y was composed of a fully-connected layer and softmax activation function, where its index indicates the probability of each class. The cross-entropy loss function is typically used and the gradient descent method is applied to train entire networks.

classification, however, obtaining a large labeled dataset is difficult due to the cost and confidentiality of underwater defense systems. Furthermore, it is difficult to implement bulk networks in active sonar systems. Consequently, the large complexity of the networks makes it hard to apply to practical applications [29].

To overcome the limitations of supervised learning with the bulk networks, we use prior information that the sonar dataset is imbalanced as explained in the previous section. The AD approaches are suitable for the imbalanced dataset consisting of a large number of normal class and a small number of abnormal class. Therefore, we are now considering the AD approaches, which may enable shallow networks to have a similar or better performance, compared to the supervised learning approaches and we will explain it next subsection.

B. ANOMALY DETECTION: UNSUPERVISED LEARNING

Deep support vector data description (SVDD) is a form of well-known deep AD based on an unsupervised learning approach [18]. It finds a sphere embracing the normal data samples in latent space. Fig. 4(a) shows the architecture of deep SVDD. Deep SVDD was performed in two steps. First, the autoencoder (AE) is used to pre-train the weights of the encoder function to form the latent space of the normal data samples and calculate the centroid; the weights are adjusted to make outputs sames as inputs and it is well-known strategies in the deep AD approach [18]. The pre-training can be conducted to minimize the following loss:

$$\mathbf{z} = \phi\left(\mathbf{x}; \mathbf{W}_\phi\right), \tag{4}$$

$$\{\mathbf{W}_\phi^{AE}, \mathbf{W}_\psi^{AE}\} = \underset{\mathbf{W}_\phi, \mathbf{W}_\psi}{\operatorname{argmin}}\, l\left(\psi\left(\mathbf{z}; \mathbf{W}_\psi\right)\right), \tag{5}$$

where $\phi : \mathbf{x} \rightarrow \mathbf{z}$ represents an encoder function for feature learning with weight parameter $\mathbf{W}_\phi$, $\psi : \mathbf{z} \rightarrow \hat{\mathbf{x}}$ represents a decoder function for decoding the latent vector to reconstructed vector $\hat{\mathbf{x}}$ with weight parameter $\mathbf{W}_\psi$, and l represents a loss function for AE, which is typically mean square error. $\mathbf{W}_\phi^{AE}$ and $\mathbf{W}_\psi^{AE}$ represent pre-trained weights of the encoder and decoder functions, respectively. The centroid

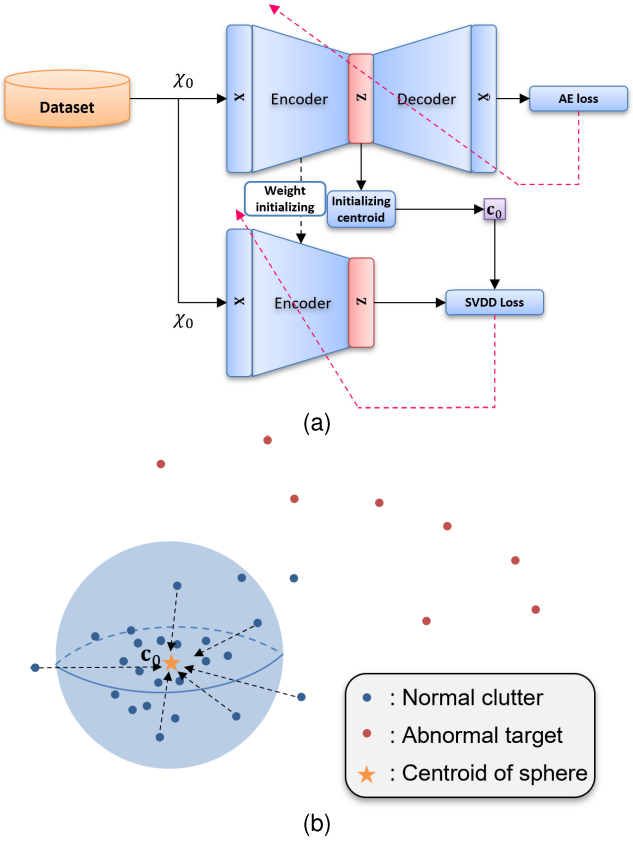


FIGURE 4. Description of the deep SVDD. (a) Network architecture of the deep SVDD. First, AE is pre-trained using only normal data samples to form the latent space. Following pre-training, centroid of the sphere of normal data samples is calculated. The weights of pre-trained encoders are used to initialize the training encoder. In the main training process, Deep SVDD attempts to concentrate the normal data samples to the centroid of the sphere using only normal data samples. After training, the anomaly score can be calculated. (b) The manifold learning concept of the deep SVDD. Normal clutter data samples (blue dots) are concentrated on the centroid $\mathbf{c_0}$ of a normal sphere (blue sphere).

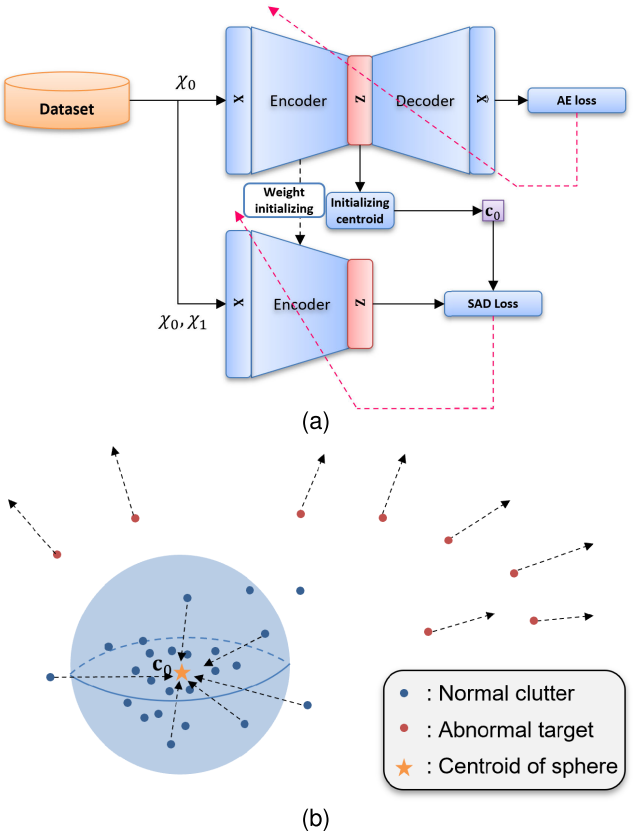


FIGURE 5. Description of the deep SAD. (a) Network architecture of the deep SAD. Pre-training of AE using normal data samples is identical to deep SVDD. However, in the main training process, deep SAD concentrates the normal clutter samples on the centroid of the sphere while penalizing the abnormal target data samples. Following the training, the anomaly score can be calculated the same as deep SVDD. (b) The concept of manifold learning of the deep SAD. Normal clutter data samples (blue dots) are concentrated on the centroid $\mathbf{c_0}$ of a normal sphere (blue sphere) while abnormal target data samples (red dots) move away from the centroid.

of the normal data samples $\mathbf{c}_0$ in the latent space can be calculated as:

$$\mathbf{c}_0 = \frac{1}{N_0}\sum_{i=1}^{N_0}\mathbf{z}_i^{AE}, \tag{6}$$

$$\mathbf{z}_i^{AE} = \phi\left(\mathbf{x}_i; \mathbf{W}_\phi^{AE}\right), \tag{7}$$

where $\mathbf{x}_i$ is i^{th} normal data sample, and N_0 denotes the number of normal data samples.

In the following step, to embrace the normal data samples using the sphere in latent space, deep SVDD minimizes the following loss:

$$\{\mathbf{W}_\phi^*\} = \underset{\mathbf{W}_\phi}{\text{argmin}}\frac{1}{N_0}\sum_{i=1}^{N_0}||\phi(\mathbf{x}_i; \mathbf{W}_\phi) - \mathbf{c}_0||^2. \tag{8}$$

The encoder ϕ whose weights are initialized from AE pre-training, $\mathbf{W}_\phi^{AE}$, is adjusted to derive the sphere. $\mathbf{W}_\phi^*$ denote the trained weights of the encoder function.

Following training, the anomaly score of deep AD approach $s_{AD}(\mathbf{x})$ can be calculated by measuring a distance of encoded latent vector for the input data sample from the sphere centroid $\mathbf{c}_0$ and it is denoted as:

$$s_{AD}(\mathbf{x}) = ||\phi(\mathbf{x}; \mathbf{W}_\phi^*) - \mathbf{c}_0||^2. \tag{9}$$

Fig. 4(b) illustrates the concept of the manifold of SVDD. A manifold is formed, such that the normal data samples are concentrated at the centroid of the sphere. However, the abnormal data samples are far from the centroid of the sphere. Therefore, distance from the centroid of the normal sphere can measure the anomaly of the data samples.

C. ANOMALY DETECTION: SEMI-SUPERVISED LEARNING

Although deep SVDD shows promising results in various fields [30], [31], [32], it has limited classification (or detection) performance because only normal data samples are used during training.

Recently, a semi-supervised approach to deep AD that utilizes labeled abnormal data samples was proposed, which is called deep semi-supervised anomaly detection (SAD) [19].

Fig. 5(a) displays the architecture of deep SAD. As in deep SVDD, deep SAD comprises two steps: pre-training the AE and learning the manifold. In the first step, pre-training was performed using normal data samples, which was the same as that in the deep SVDD. However, in the second step, deep SAD uses a loss function that is different from that in deep SVDD. The loss function searches for the sphere embracing the dominant normal samples in the latent space, while it penalizes (penalizing means pushing the samples from the centroid; as in [19]) a small number of abnormal samples from the sphere centroid. The loss function in a deep SAD is expressed as

$$\{\mathbf{W}_\phi^*\} = \underset{\mathbf{W}_\phi}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \left(||\phi(\mathbf{x}_i; \mathbf{W}_\phi) - \mathbf{c}_0||^2 \right)^{y_i}. \quad (10)$$

$$y_i = \begin{cases} 1 & \text{if } \mathbf{x}_i \in \chi_0 \\ -1 & \text{if } \mathbf{x}_i \in \chi_1 \end{cases}. \quad (11)$$

Following the training, anomaly score $s(\mathbf{x})$ can be calculated using (9) as in the deep SVDD. Fig. 5(b) illustrates the concept of the manifold of deep SAD. A manifold is formed such that the normal and abnormal samples are located near and far from the centroid of the sphere, respectively. The distances of the abnormal data samples from the sphere centroid are greater than those in the deep SVDD owing to the loss function that repels them from the sphere centroid. Therefore, a deep SAD has an enhanced generalization performance.

IV. BI-SPHERE ANOMALY DETECTION FOR ACTIVE SONAR CLASSIFICATION

In the conventional deep AD approaches, such as deep SVDD and SAD, it learns to make a single compact sphere manifold by considering the majority of the normal data samples. In particular, in deep SAD, minor abnormal data samples assist in forming the sphere, including the normal data samples, resulting in better generalization.

However, it is noteworthy that the abnormal target samples have similar experiences of propagation and scattering from the artificial objects in active sonar systems. Therefore, the corresponding latent vectors should be in close proximity. The similarity between abnormal target samples can be exploited to enhance the generalization of deep AD. Accordingly, we propose bi-sphere anomaly detection (BiSAD) in which an additional sphere embracing abnormal data samples is added to the latent space based on the properties of active sonar data. BiSAD has two spheres to embrace respectively normal and abnormal samples to improve generalization performance and finds the latent space where the distance between the centroids of the two spheres is maximized. Because the bi-sphere concept of BiSAD is motivated by the characteristics of the active sonar dataset, it can be expected that the generalization performance will be improved.

Fig. 6(a) shows the architecture of BiSAD. BiSAD comprises two steps, similar to conventional deep AD approaches. However, the details of each step were different. First, AE was

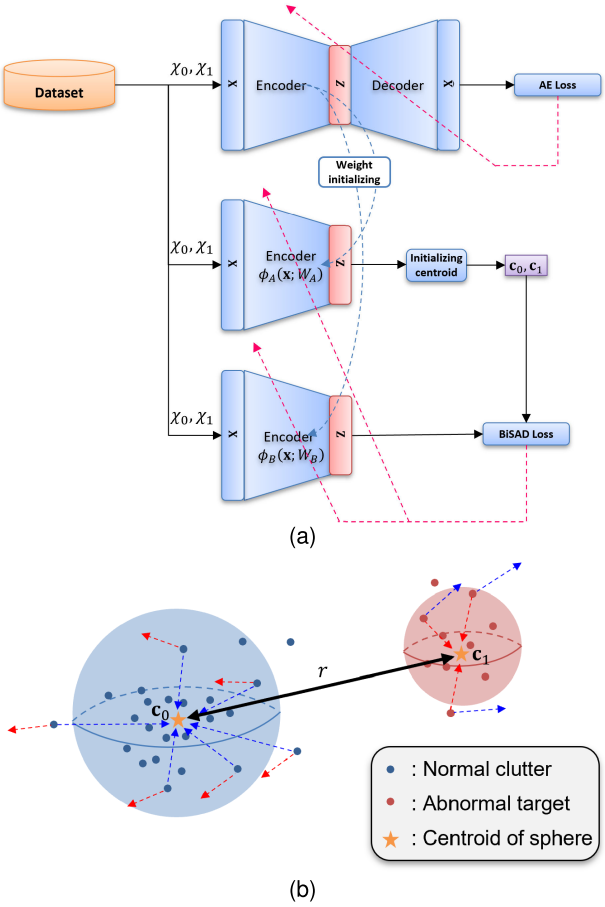


FIGURE 6. Description of the BiSAD. (a) Network architecture of BiSAD. AE is pre-trained using normal and abnormal data samples both. In the main training process, BiSAD uses two encoders: Encoder ϕ_A which learns the centroids of two spheres and encoder ϕ_B which learns the manifold. Following the training, the anomaly score can be calculated the same as conventional deep AD approaches. (b) The concept of manifold learning of the BiSAD. Normal clutter data samples (blue dots) are concentrated on the centroid $\mathbf{c}_0$ of a normal sphere (blue sphere) while moving away from the centroid $\mathbf{c}_1$ of an abnormal sphere (red sphere). Abnormal target data samples are used to train BiSAD in the opposite direction to normal clutter data samples. Simultaneously, BiSAD is trained so that the centroids of the two spheres move away from each other.

pre-trained to form a latent space using both normal and abnormal data samples. It is because BiSAD attempts to form individual spheres for both the normal and abnormal data samples. In the following step, BiSAD uses two encoders, unlike conventional AD approaches that use a fixed single centroid of a normal data sphere calculated by a pre-trained AE encoder (conventional AD approaches, therefore, require a single encoder function to learn manifold). In contrast, in BiSAD, one of the encoders learns the variable centroids of two spheres (ϕ_A), and the other encoder learns the manifold (ϕ_B).

Using the first encoder ϕ_A, the centroids of normal data samples $\mathbf{c}_0$ and abnormal data samples $\mathbf{c}_1$ in latent space can be calculated as follows:

$$\mathbf{c}_0(\mathbf{W}_{\phi_A}) = \frac{1}{N_0} \sum_{i=1}^{N_0} \phi_A(\mathbf{x}_i; \mathbf{W}_{\phi_A}), \quad \mathbf{x}_i \in \chi_0 \quad (12)$$

$$\mathbf{c}_1(\mathbf{W}_{\phi_A}) = \frac{1}{N_1}\sum_{i=1}^{N_1}\phi_A(\mathbf{x}_i;\mathbf{W}_{\phi_A}), \quad \mathbf{x}_i \in \chi_1 \tag{13}$$

where N_0 and N_1 represent the numbers of normal and abnormal data samples, respectively.

Using the second encoder ϕ_B, BiSAD tries to find the manifold. BiSAD concentrates the normal and abnormal data samples near the corresponding sphere centroids of $\mathbf{c}_0$ and $\mathbf{c}_1$ (relevant to the first and last terms in (14)). Simultaneously, BiSAD penalizes normal (or abnormal) data samples from $\mathbf{c}_1$ (or $\mathbf{c}_0$) (relevant to the second and third terms in (14)). BiSAD finds the manifold of two spheres and attempts to increase the distance r between the centroids, simultaneously. Thus, the loss function in BiSAD is defined as follows:

$$\{\mathbf{W}^*_{\phi_A}, \mathbf{W}^*_{\phi_B}\} = \underset{\mathbf{W}^*_{\phi_A},\mathbf{W}^*_{\phi_B}}{\operatorname{argmin}} \frac{\kappa_{00}d_{00} + \kappa_{01}d_{01} + \kappa_{10}d_{10} + \kappa_{11}d_{11}}{r^2}. \tag{14}$$

where

$$d_{00} = \frac{1}{N_0}\sum_{i=1}^{N_0}\left(||\phi_B(\mathbf{x}_i;\mathbf{W}_{\phi_B}) - \mathbf{c}_0(\mathbf{W}_{\phi_A})||^2\right)^{+1}, \quad \mathbf{x}_i \in \chi_0$$

$$d_{01} = \frac{1}{N_0}\sum_{i=1}^{N_0}\left(||\phi_B(\mathbf{x}_i;\mathbf{W}_{\phi_B}) - \mathbf{c}_1(\mathbf{W}_{\phi_A})||^2\right)^{-1}, \quad \mathbf{x}_i \in \chi_0$$

$$d_{10} = \frac{1}{N_1}\sum_{i=1}^{N_1}\left(||\phi_B(\mathbf{x}_i;\mathbf{W}_{\phi_B}) - \mathbf{c}_0(\mathbf{W}_{\phi_A})||^2\right)^{-1}, \quad \mathbf{x}_i \in \chi_1$$

$$d_{11} = \frac{1}{N_1}\sum_{i=1}^{N_1}\left(||\phi_B(\mathbf{x}_i;\mathbf{W}_{\phi_B}) - \mathbf{c}_1(\mathbf{W}_{\phi_A})||^2\right)^{+1}, \quad \mathbf{x}_i \in \chi_1, \tag{15}$$

$$r = ||\mathbf{c}_0(\mathbf{W}_{\phi_A}) - \mathbf{c}_1(\mathbf{W}_{\phi_A})||, \tag{16}$$

and $\mathbf{k} = [\kappa_{00}, \kappa_{01}, \kappa_{10}, \kappa_{11}]$ represents the weighting parameter for determining the strategy in manifold learning (the effects of these parameters will be discussed in the Sec. VI-B).

Following training, anomaly score $s(\mathbf{x})$ can be calculated using (9) as in conventional deep AD approaches. Fig. 6(b) shows the manifold concept of the BiSAD. The manifolds are formed such that the normal and abnormal data samples are concentrated on the corresponding sphere centroids and repelled from the opposite sphere centroids. Concurrently, the centroids of the spheres were trained to be distant from each other.

V. PRELIMINARIES FOR ML-BASED ACTIVE SONAR CLASSIFICATION

A. ACTIVE SONAR DATASET

To verify the BiSAD, we use sea experimental data with one artificial underwater target which is collected by an active sonar system. In these experiments, a linear chirped pulse was transmitted multiple times with a pause and received through a linear sensor array. In conclusion, we acquired raw beam signal data measured along azimuth angles (beam angles) and time (ping numbers).

To generate the active sonar dataset for the training and testing, we use the classify-after-detect strategy (MF, normalization, and clustering are sequentially performed to generate contacts) in Sec II-A. In total, we achieved contacts (whose number was in the order of 10^3) and extracted the contact signals from the raw audio signals. The length of the contact signal was set to twice the length of the pulse by considering the multipath propagation effect, which elongates the transmitted signal.

Each contact signal was annotated with target or nontarget by three experienced sonar experts. The number of target and nontarget data samples was in the order of 10^1 and 10^3, respectively. As expected, the target is a minor class which is abnormal when using the AD approaches. Note that the signal-to-noise ratio (SNR) of the target echo differed from ping to ping owing to the varying transmitter-target-receiver geometry and ocean environments.

To complete the active sonar dataset, we need to divide it into a training dataset and a test dataset. Unlike ordinary ML data split, which is randomly divided into training and test datasets [24], the active sonar dataset should be split by considering physical characteristics for meaningful experiments. Based on the observation of target echo variation from ping to ping, we divide the dataset by temporal change with the ping. More specifically, we define the data samples from the ten first pings (early ten pings) as the training dataset and data samples from the remaining pings as the test dataset. It is noteworthy that the number of targets in the training dataset was less than ten whereas the number of the clutter appeared about five hundred.

B. PREPROCESSING OF ACTIVE SONAR DATASET FOR CLASSIFICATION

In general, it is hard to extract the aural and perceptual information directly from raw audio signals because of their complexity due to high dimensions [33]. Therefore, in conventional audio signal processing, it is natural to transform raw audio signals into time-frequency (TF) data using short-time Fourier transform (STFT) for further processing [34]. Likewise, it is natural to transform the raw beam signal into TF data before applying ML-based active sonar classifiers. Fig. 7 illustrates the transformation process. First, the raw audio signal is extracted and STFT is applied. Following STFT, frequency constraining (considering the bandwidth of the transmitted pulse) and resizing are applied sequentially. The pre-processed data are called TF images. In conclusion, ML-based active sonar classifiers attempt to discriminate the target from clutter using TF images as inputs.

Fig. 8 shows examples of the TF images. Figs. 8(a) and (b) show three different TF images of the target and clutter, respectively. It is difficult to distinguish visually because the target and clutter signals arise from scattering by the same transmitted signal (they have a similar frequency band with the transmitted signal).

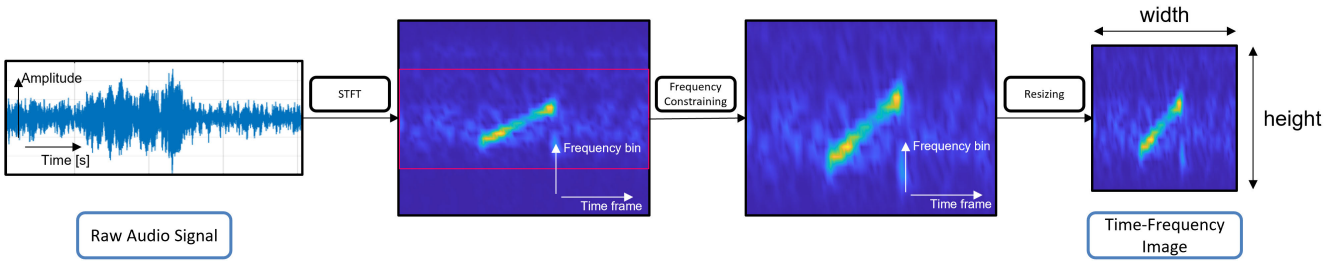


FIGURE 7. The pre-processing of active sonar dataset for classification. The raw audio beam signal is transformed into time-frequency (TF) data through short-time Fourier transform (STFT). Following STFT, bandwidth constraining and resizing were applied. The TF image presents the input of active sonar classifier.

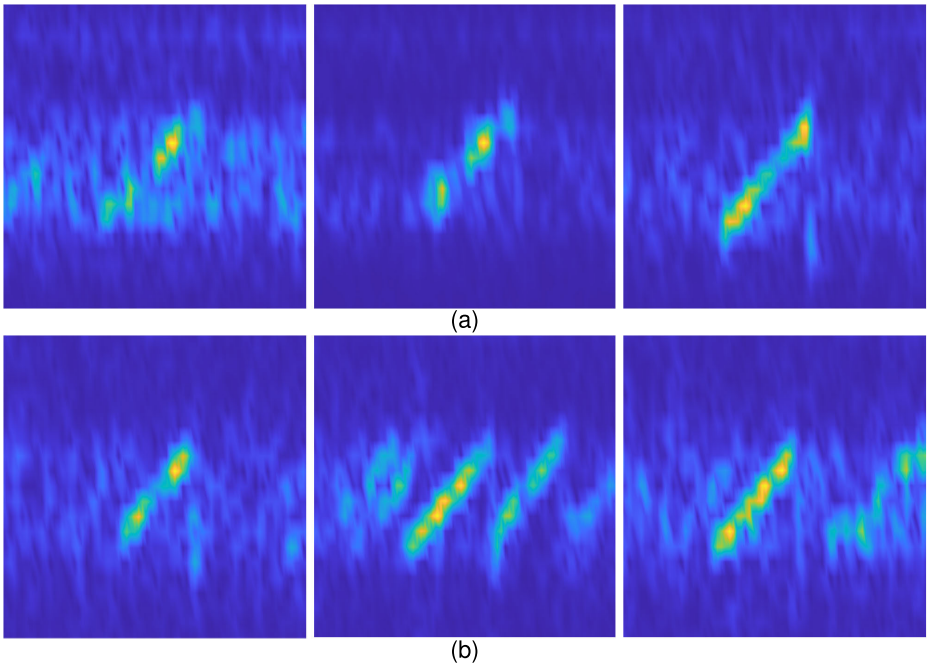


FIGURE 8. The examples of time-frequency (TF) image of (a) TF images of target and (b) are TF images of nontarget in the active sonar dataset. Because target and nontarget TF images are generated from similar physical experiences, it is hard to discriminate. The purpose of this study is to develop an active sonar classifier that can classify these TF images.

C. SETUP FOR TRAINING AND TESTING

The encoder networks for feature learning are composed of two CNN layers followed by two fully connected (FC) layers after flattening. The decoder networks for AE have a symmetrical structure to that of the encoder networks. The network architecture is summarized in Table 1. The size of the input TF image was set to 64×64.

All supervised learning and deep AD approaches are initialized using pre-trained AE weights. For training (including pre-training), the Adam optimizer was used with an epoch of 100, batch size of 64, and learning rate of 0.001. Five-fold cross-validation was conducted [35] and an early stopping technique was employed using the validation loss in each fold to prevent overfitting. The final output was calculated using the average of five model outputs trained in the five-fold cross-validation. For supervised learning, the softmax output, which is the mean probability of each class, is averaged, and for deep AD approaches, the anomaly scores are averaged [36], [37].

Owing to small amount of the imbalanced active sonar dataset, ordinary supervised learning makes decision with a bias toward to dominant normal class, which results in excessive false alarms and missing targets. To overcome this problem, data argumentation can be used to complement abnormal data samples by modifying the existing abnormal data samples [24]. However, general data argumentation schemes such as rotating, shifting, and shearing are unsuitable for active sonar dataset because it does not consider the physical characteristics of underwater sound propagation and scattering.

Inevitably, down-sampling is used, which reduces the number of dominant normal data samples to the number of abnormal data samples when the network is trained [24].

TABLE 1. Networks architecture.

(a) Encoder

Input	Shape: $1 \times 64 \times 64$.
CNN layer	Channel: 32, 3×3 Convolution. Stride: 1, Padding: 1, ReLU activation function, 2×2 Maxpooling.
CNN layer	Channel: 64, 3×3 Convolution, Stride: 1, Padding: 1, ReLU activation function, 2×2 Maxpooling.
Flatten	Shape: 1×16384.
FC layer	Input channel: 16384, Output channel: 512, ReLU activation function.
FC layer	Input channel: 512, Output channel: 128.
Output	Shape: 1×128.

(b) Latent space

Latent vector	Shape: 1×128.

(c) Decoder

Input	Shape: 1×128.
FC layer	Input channel: 128, Output channel: 512, ReLU activation function.
FC layer	Input channel: 512, Output channel: 16384.
Reshape	Shape: $64 \times 16 \times 16$.
Transposed CNN layer	2×2 Unpooling, Stride: 1, Padding: 1, Channel: 32, 3×3 Transposed convolution, ReLU activation function.
Transposed CNN layer	2×2 Unpooling, Stride: 1, Padding: 1, Channel: 1, 3×3 Transposed convolution.
Output	Shape: $1 \times 64 \times 64$.

VI. ML-BASED ACTIVE SONAR CLASSIFICATION

We performed a generalization test using active sonar dataset for the BiSAD along with the supervised learning, deep SVDD, and deep SAD. For a qualitative analysis, the receiver operating characteristic (ROC) curve is used and is calculated using the anomaly score according to ranging thresholds in the probability of detection P_D and probability of false alarm P_{fa}. P_D and P_{fa} are defined as:

$$P_D = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{17}$$

$$P_{fa} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{18}$$

where TP, TN, FP, and FN indicate true positive (predict actually abnormal target as abnormal target), true negative (predict actually normal clutter as normal clutter), false positive (predict actually normal clutter as abnormal target), and false negative (predict actually abnormal target as normal clutter), respectively.

TABLE 2. Performance analysis of average ROC curves.

ML-based approach	$P_D{}^*$	AUC
Supervised learning	0.650	0.900
Deep SVDD	0.333	0.942
Deep SAD	0.721	0.976
BiSAD	**0.800**	**0.989**

*P_D at the fixed P_{fa} of 0.01

For quantitative analysis, we calculated P_D at the specific P_{fa} of 0.01, and the area under the curve (AUC) of the average ROC curve according to the considered ML-based approaches.

A. COMPARATIVE ANALYSIS

Fig. 9(a) to (d) illustrate the ROC curves of supervised learning, deep SVDD, deep SAD, and BiSAD, in the test dataset after training using the training dataset (i.e., the ten first pings). Because the number of data samples is limited in the active sonar dataset, we cannot guarantee performance using the results obtained from a single trial. Therefore, in our analysis, we conducted 30 trials and analyze 30 ROC curves to ensure the reliability of the results. Fig. 9 shows the average, minimum, and maximum ROC curves, which represent the average, minimum, and maximum values at a specific P_{fa} among the 30 ROC curves. Fig. 10 presents a close-up view of the average ROC curve plotted on a logarithmic scale of the lower P_{fa} part for a distinct comparison of performance.

The supervised learning exhibits the lowest performance and largest variability than AD approaches, owing to the small training dataset whose size is in the order of 10^1 after the down-sampling. The BiSAD shows high P_D than the other ML-based approaches at low P_{fa}. Among the ML-based classifiers, the BiSAD exhibits the best performance. The calculated quantitative values are presented in Table 2. The P_D and AUC of BiSAD were both higher than those of the others. Compared to supervised learning, BiSAD exhibited superior P_D and AUC values of 0.150 and 0.089, respectively.

B. ANALYSIS ON THE κ IN THE LOSS FUNCTION

In the BiSAD loss function of (14), we adopted four weights κ_{00}, κ_{01}, κ_{10}, and κ_{11} to control the manifold learning of BiSAD. In this subsection, we analyze the generalization test performance according to the weight vector $\mathbf{k}$ was set to [1, 0, 0, 1], [0, 1, 1, 0], and [1, 1, 1, 1].

The $\mathbf{k}$ of [1, 0, 0, 1] attempts to concentrate normal and abnormal data samples on the individual spheres with less penalization; the penalization is solely considered by r in the denominator of (14). Meanwhile, the $\mathbf{k}$ of [0, 1, 1, 0] emphasize the penalizing without considering the

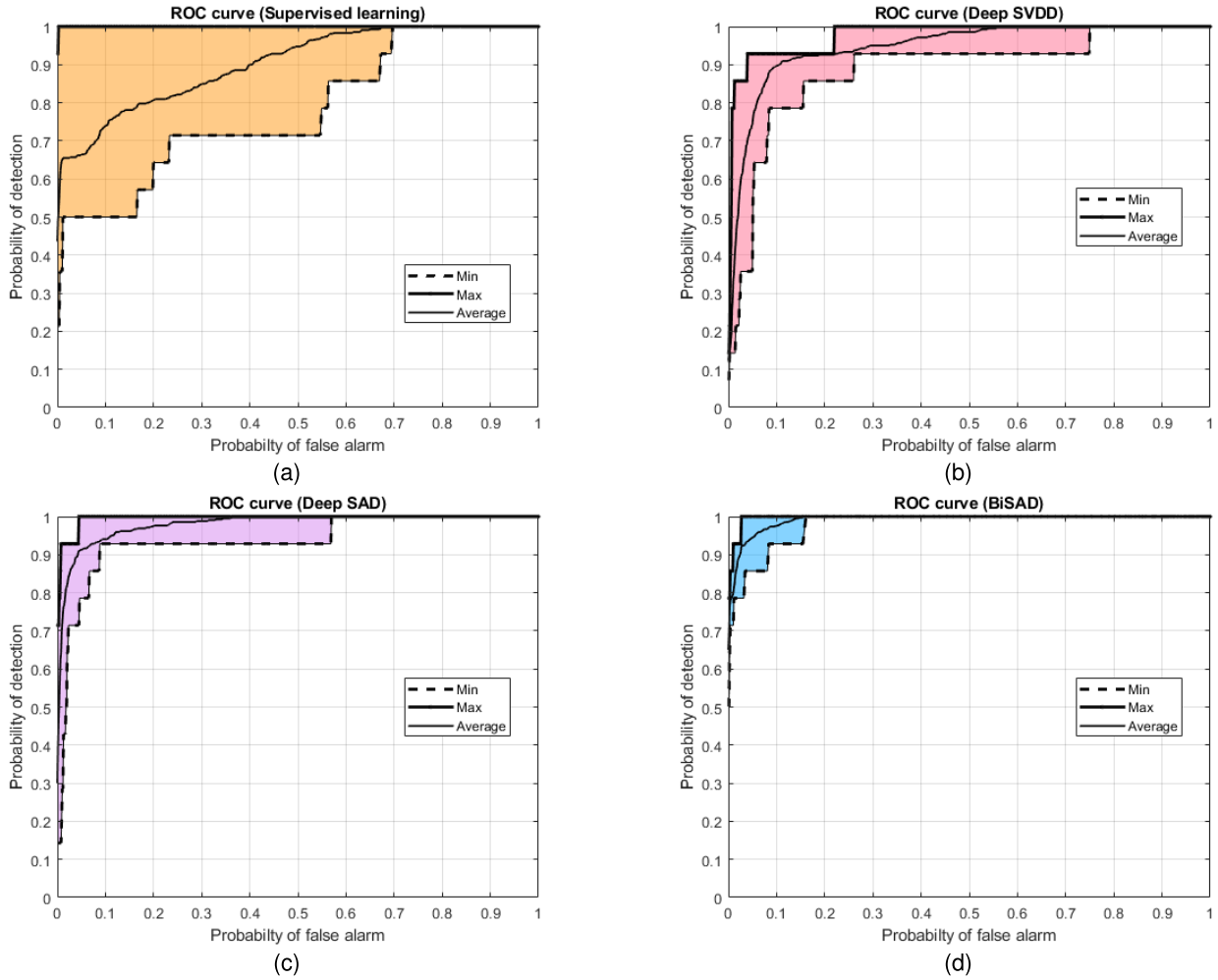


FIGURE 9. Comparison of the ROC curves of (a) supervised learning, (b) deep SVDD, (c) deep SAD, and (d) BiSAD, respectively. The performance variability of supervised learning is high. BiSAD exhibits superior performance and stable ROC curves.

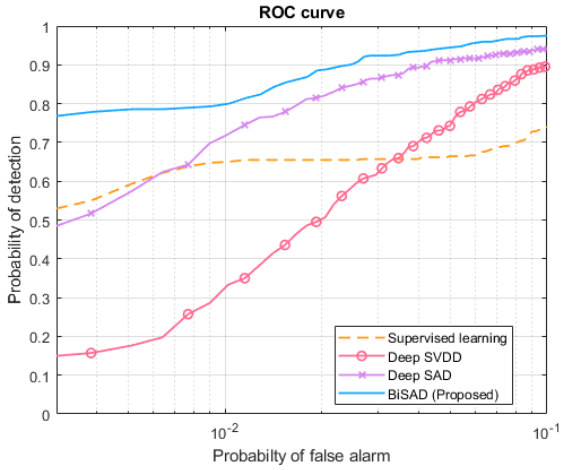


FIGURE 10. Comparison of the average ROC curves. The BiSAD exhibits superior performance than other approaches.

concentration. When we set **k** as [1, 1, 1, 1], BiSAD uses both concentration and penalization; therefore, it finds compact spheres while considering penalization.

Fig. 11(a), (b), and (c) show a comparison of the average, maximum, and minimum ROC curves, and Fig. 12 shows a comparison of average ROC curves at P_{fa} less than 0.1. The maximum ROC curves of **k** of [1, 0, 0, 1] and [0, 1, 1, 0] are superior than that of **k** of [1, 1, 1, 1], and the average ROC curve is good in the order of [0, 1, 1, 0], [1, 1, 1, 1], and [1, 0, 0, 1]. However, the variability of the classification performance of **k** of [1, 0, 0, 1] and [0,1,1 0] is significant whereas that of **k** of [1, 1, 1, 1] is moderately consistent.

The results imply that solely focusing on concentrating or penalizing can mislead the manifold learning of the BiSAD. On the other hand, if concentrating and penalizing are used simultaneously, stability can be improved because the latent space is explored while complementing each other. Therefore, we used **k** of [1, 1, 1, 1] for robust performance in the following experiment.

C. GENERALIZATION TEST ON THE BEAM SIGNAL INCLUDING TARGET ECHO

We conducted a generalization test on the beam signal (classify-before-detect strategy in Fig. 1 (b)) including target

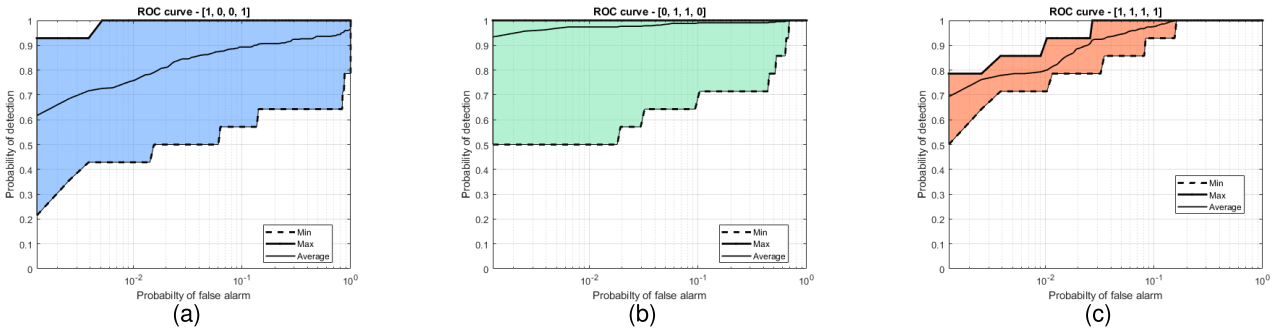


FIGURE 11. Comparison of the ROC curves according to k of (a) [1, 0, 0, 1], (b) [0, 1, 1, 0], and (c) [1, 1, 1, 1]. The case of k of [1, 1, 1, 1] shows stable results than others.

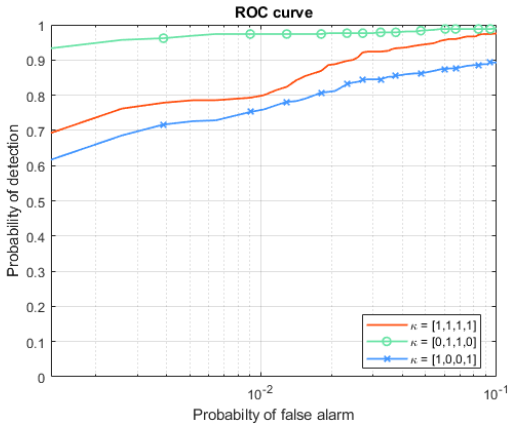


FIGURE 12. Comparison of the average ROC curves according to k. The case of k of [0, 1, 1, 0] shows superior performance; however, variability is high (as shown in the Fig. 11).

echo at a specific ping and azimuth angle using ML-based classifiers trained with the active sonar dataset. Fig. 13 illustrates the spectrogram of the beam signal containing the target echo and the average anomaly scores of supervised learning, deep SVDD, deep SAD, and BiSAD. The beam signals of the 11^{th} and the 29^{th} pings were utilized. As mentioned in Sec. V-A, the training dataset consists of data samples from the ten first pings. Because the ocean environment changes over time, the beam signal of the 11^{th} ping had similar characteristics to the training dataset, whereas the beam signal of the 29^{th} ping had different characteristics.

The average anomaly scores in Fig. 13 are calculated by averaging anomaly scores after normalization over 30 trials as in the previous experiment; the normalization was conducted by dividing the anomaly score by its maximum value. By comparing the anomaly scores in Figs. 13(a) and (b), the classification performances of the considered ML-based active sonar classifiers diminish for the beam signal at the 29^{th} ping, which deviates from those in the training dataset. Particularly, the maximum value of averaged anomaly score from the supervised learning is less than 0.8 because the anomaly scores from supervised learning vary along 30 trials; the inconsistent prediction of supervised learning deteriorates reliability for target and clutter classification. However, BiSAD shows a more robust and accurate classification performance than others. Furthermore, the anomaly score of BiSAD is narrower than those of the others and shows a low level in nontarget locations; particularly in the earlier time corresponding to the reverberation region. The generalization test in Fig. 13 implies that BiSAD can accurately distinguish a target from a nontarget even for the unexperienced data samples and the generalization performance of BiSAD is superior to those of the conventional ML-based classifiers.

D. COMPARATIVE ANALYSIS WITH VARIOUS DEEP NEURAL NETWORKS

It would be interesting to analyze the performance of the well-known deep neural networks in a small dataset of the active sonar classification problem and compare them with BiSAD. We trained and tested four networks, VGG16, ResNet50, ResNeXt, and SwinViT achieving remarkable performance in the vision with a supervised learning approach [25], [26], [27], [28]. These networks with complex models (deeper layers with more parameters) may have a better capacity to fit datasets than the shallow networks used in previous approaches, including BiSAD. However, the deeper layers of these networks require more memory sizes and cause overfitting problems on small data sets [24].

Fig. 14 shows a comparison of the average, maximum, and minimum ROC curves of VGG16, ResNet50, ResNeXt, and SwinViT, respectively. The supervised learning approaches using VGG16, ResNet50, and ResNeXt show stable and high detection probabilities at false alarm rates greater than 0.01. However, the probability of detection decreases sharply and shows large variability with the decrease of false alarm rate. SwinViT shows poor performance compared to the previously proposed networks, contrary to general expectations. This phenomenon is because transformer-based networks require a large amount of data and do not fit the small active sonar dataset.

Fig. 15 shows the probability of detection at a false alarm rate of 0.001, where the red line means the median value of the probability of detection, and the upper and lower bounds of the blue box mean the first and third quartile, respectively.

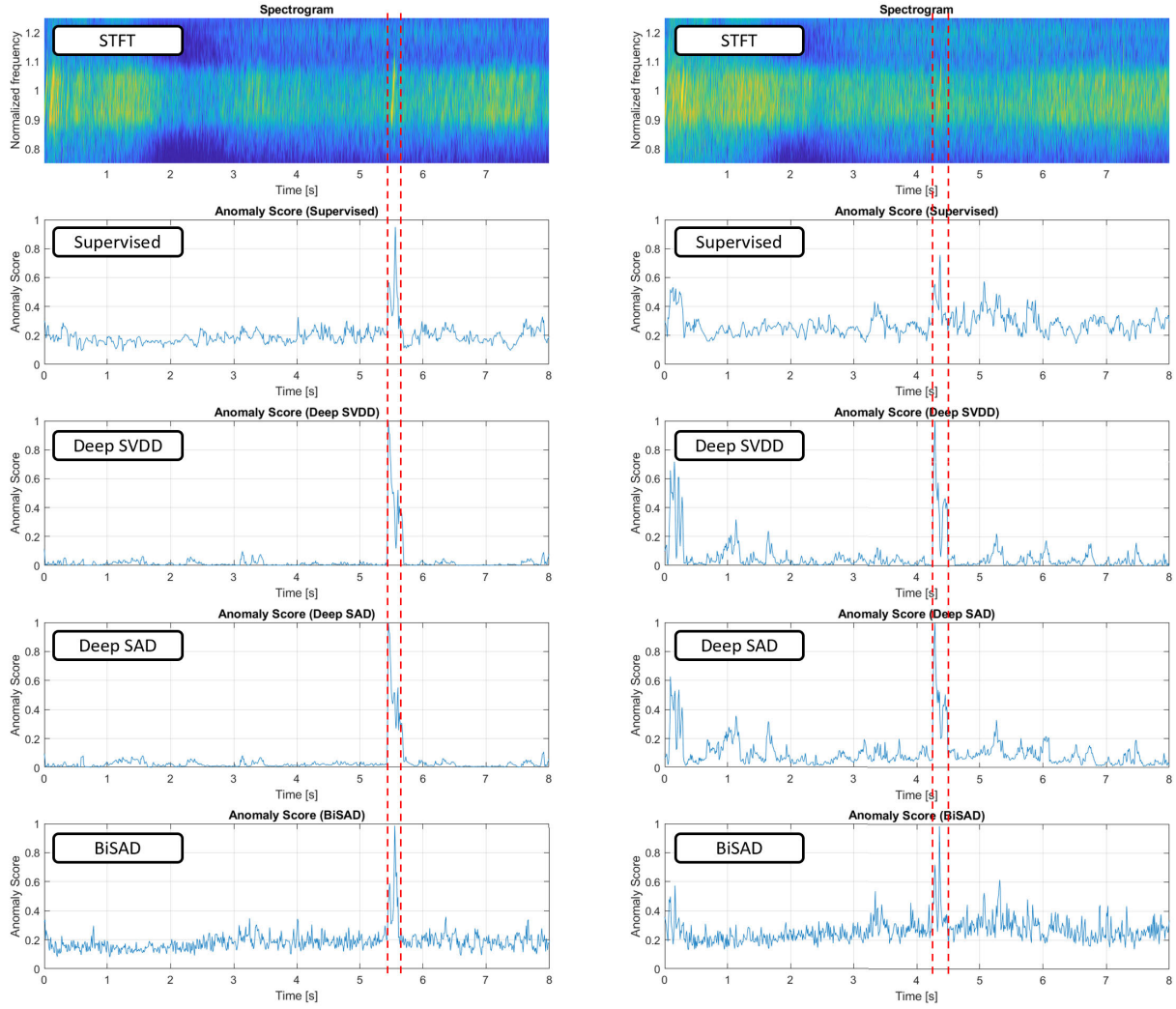


FIGURE 13. The generalization test on the whole beam signal including target echo of (a) the 11^{th} ping, and (b) the 29^{th} ping. The spectrogram of beam signal and average anomaly scores are presented. The red dot line indicates the location of the target echo. The average anomaly scores of BiSAD show narrow and accurate results. Anomaly score of BiSAD classified target echo accurately and shows a low level for nontarget locations. These results imply that the generalization performance of BiSAD is superior to conventional ML-based classifiers.

The upper and lower black lines mean the maximum and minimum probabilities of detection, and the red plus markers mean outliers. Deep SVDD and SAD show low probabilities of detection (median values of 0.14 and 0.50, respectively) with large variabilities. On the other hand, BiSAD shows a high probability of detection (median value of 0.79) with small variability. In the supervised learning approaches, shallow CNN shows a low probability of detection (median value of 0.43) although its maximum performance is high. Furthermore, it shows the largest variability. SwinViT shows a low probability of detection (median value of 0.07) with large variability and its lower bound reaches the probability of detection of zero. VGG16, ResNet50, and ResNeXt show high probabilities of detection (median value of 0.79) with relatively small variabilities, however, they have outliers that are lower than 0.5. These outliers might be caused by the overfitting problem owing to the small active sonar dataset used to train the deep neural networks.

It is noteworthy that the performance of shallow networks trained with supervised learning is significantly enhanced by BiSAD using the AD strategy modified based on the active sonar characteristics. When we compare three supervised learning-based approaches (VGG16, ResNet50, ResNeXt) with the BiSAD, they show a similar median value of detection probability, but BiSAD shows more stable results than supervised learning-based approaches having outliers. Also, the first quartile of BiSAD equals its median value owing to consistent results across multiple trials and it implies that BiSAD has strong reliability. From the results in Fig. 15, we could confirm that BiSAD using the

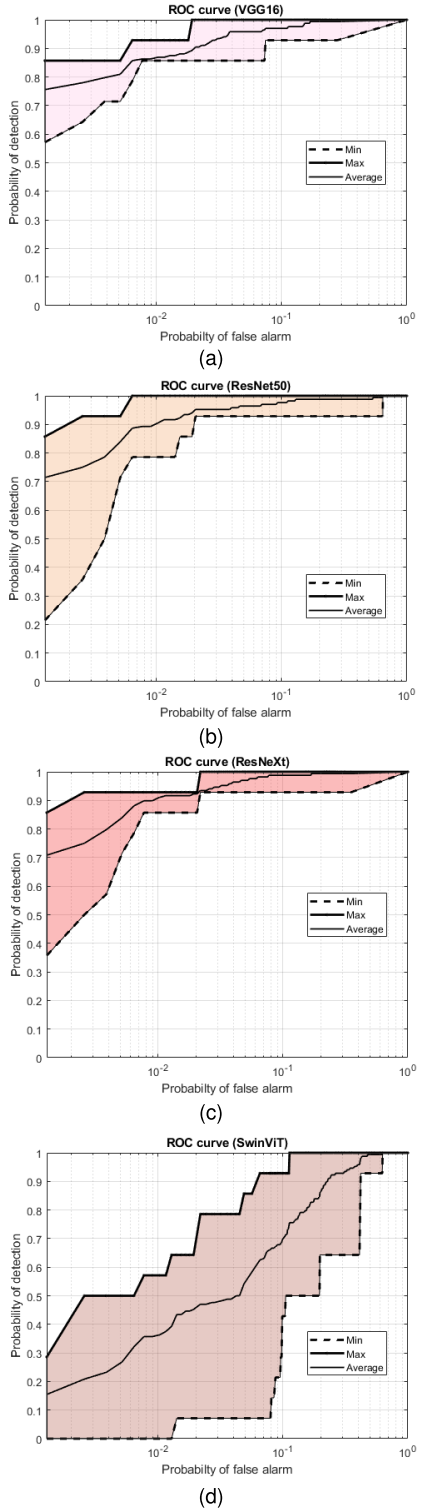


FIGURE 14. Comparison of the ROC curves of supervised learning with (a) VGG16, (b) ResNet50, (c) ResNeXt, and (d) SwinViT, respectively. At the low false alarm rate, deep neural network-based supervised learning shows low detection probabilities with large variability. SwinViT, which requires a large amount of data, shows the lowest performance in active sonar classification problems using small datasets.

shallow networks with much fewer parameters has better generalization performance than deep neural networks-based

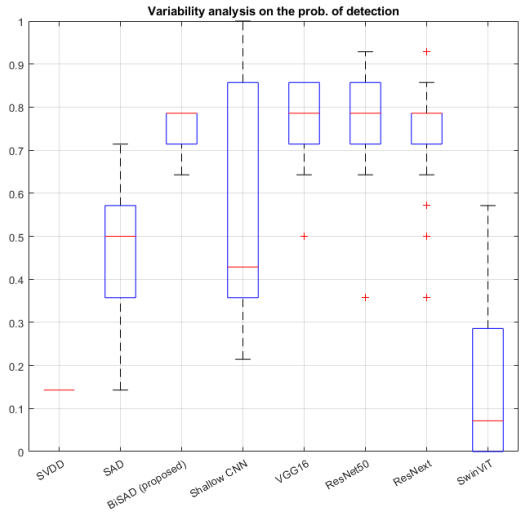


FIGURE 15. The probability of detection of various ML-based approaches at a false alarm rate of 0.001. BiSAD shows better generalization performance than conventional AD approaches and deep neural networks-based supervised learning approaches even though it uses a smaller size of networks.

supervised learning approaches. These results show that the prior assumption of BiSAD is fit for the active sonar classification problem. Therefore, BiSAD is more suitable for practical application to active sonar systems than the deep neural networks.

VII. CONCLUSION

In this study, we proposed the BiSAD, which is a modified deep AD approach. The difference between BiSAD and the conventional deep AD approach is that BiSAD assumes that abnormal data samples have characteristics similar to those of normal data samples because all abnormal data samples are caused by artificial objects. This assumption leads BiSAD to learn two individual sphere manifolds: normal (clutter) and abnormal (target). The loss function of BiSAD consists of two concentration terms and two penalizing terms based on the relationship between the spheres and data class. Simultaneously, BiSAD also searches for the latent space where the distance between the two centroids of spheres is maximized.

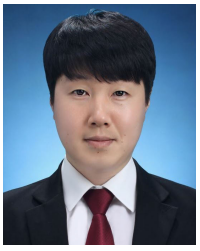
To verify the BiSAD, we use sea experimental data that include scattered signals from underwater artificial objects. We divided the training and test data samples according to the ping. Subsequently, ROC curves were calculated to evaluate the qualitative performance, and P_D and AUC were presented as quantitative evaluations. The results show that BiSAD has superior classification performance and stability compared to conventional deep AD and supervised learning approaches. We also analyzed the performance based on the weight of the loss function setting. The analysis demonstrated that robust performance was obtained when all terms in the loss function were used. Furthermore, we conduct a generalization test on the beam signal including target echo and compared the BiSAD with supervised learning

approaches with various deep neural networks that have deeper layers and more parameters. The results reveal that BiSAD has superior generalization performance compared to conventional ML-based classifiers including deep neural networks-based approaches.

REFERENCES

[1] A. D. Waite, *Sonar for Practising Engineers*. Hoboken, NJ, USA: Wiley, 2002.
[2] R. O. Nielsen, *Sonar Signal Processing*. Norwood, MA, USA: Artech House, 1991.
[3] D. A. Abraham, *Underwater Acoustic Signal Processing: Modeling, Detection, and Estimation*. Cham, Switzerland: Springer, 2019.
[4] W. S. Burdic, *Underwater Acoustic System Analysis*. Baileys Harbor, WI, USA: Peninsula Pub, 2002.
[5] R. P. Gorman and T. J. Sejnowski, "Analysis of hidden units in a layered network trained to classify sonar targets," *Neural Netw.*, vol. 1, no. 1, pp. 75–89, 1988.
[6] R. P. Gorman and T. J. Sejnowski, "Learned classification of sonar targets using a massively parallel network," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-36, no. 7, pp. 1135–1140, Jul. 1988.
[7] F. B. Shin, D. H. Kil, and R. F. Wayland, "Active impulsive echo discrimination in shallow water by mapping target physics-derived features to classifiers," *IEEE J. Ocean. Eng.*, vol. 22, no. 1, pp. 66–80, Jan. 1997.
[8] A. Trucco, "Detection of objects buried in the seafloor by a pattern-recognition approach," *IEEE J. Ocean. Eng.*, vol. 26, no. 4, pp. 769–782, Oct. 2001.
[9] S. M. Murphy and P. C. Hines, "Examining the robustness of automated aural classification of active sonar echoes," *J. Acoust. Soc. Amer.*, vol. 135, no. 2, pp. 626–636, Feb. 2014.
[10] I. Seo, S. Kim, Y. Ryu, J. Park, and D. S. Han, "Underwater moving target classification using multilayer processing of active sonar system," *Appl. Sci.*, vol. 9, no. 21, p. 4617, Oct. 2019.
[11] T. Sun, J. Jin, T. Liu, and J. Zhang, "Active sonar target classification method based on Fisher's dictionary learning," *Appl. Sci.*, vol. 11, no. 22, p. 10635, Nov. 2021.
[12] S. Lee, I. Seo, J. Seok, Y. Kim, and D. S. Han, "Active sonar target classification with power-normalized Cepstral coefficients and convolutional neural network," *Appl. Sci.*, vol. 10, no. 23, p. 8450, Nov. 2020.
[13] G. de Magistris, P. Stinco, J. R. Bates, J. M. Topple, G. Canepa, G. Ferri, A. Tesei, and K. le Page, "Automatic object classification for low-frequency active sonar using convolutional neural networks," in *Proc. OCEANS MTS/IEEE*, Oct. 2019, pp. 1–6.
[14] Y. Chen, H. Liang, and S. Pang, "Study on small samples active sonar target recognition based on deep learning," *J. Mar. Sci. Eng.*, vol. 10, no. 8, p. 1144, Aug. 2022.
[15] P. Stinco, G. De Magistris, A. Tesei, and K. D. LePage, "Automatic object classification with active sonar using unsupervised anomaly detection," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 46–50.
[16] Q. Wang, S. Du, F. Wang, and Y. Chen, "Underwater target recognition method based on multi-domain active sonar echo images," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Aug. 2021, pp. 1–5.
[17] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–38, Mar. 2021.
[18] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. AhmedSiddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4393–4402.
[19] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller, and M. Kloft, "Deep semi-supervised anomaly detection," 2019, *arXiv:1906.02694*.
[20] K. M. Kim, C. Lee, and D. H. Youn, "Adaptive processing technique for enhanced CFAR detecting performance in active sonar systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 36, no. 2, pp. 693–700, Apr. 2000.
[21] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, vol. 96, 1996, pp. 226–231.
[22] F. B. Shin and D. H. Kil, "Full-spectrum signal processing using a classify-before-detect paradigm," *J. Acoust. Soc. Amer.*, vol. 99, no. 4, pp. 2188–2197, Apr. 1996.
[23] N. Allen, P. C. Hines, and V. W. Young, "Performances of human listeners and an automatic aural classifier in discriminating between sonar target echoes and clutter," *J. Acoust. Soc. Amer.*, vol. 130, no. 3, pp. 1287–1298, Sep. 2011.
[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[27] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.
[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
[29] B. Kim, S. Yang, J. Kim, and S. Chang, "QTI submission to DCASE 2021: Residual normalization for device-imbalanced acoustic scene classification with efficient design," 2021, *arXiv:2206.13909*.
[30] J. Horváth, D. Güera, S. K. Yarlagadda, P. Bestagini, F. M. Zhu, S. Tubaro, and E. J. Delp, "Anomaly-based manipulation detection in satellite images," *Networks*, vol. 29, p. 21, Jan. 2019.
[31] L. Sun, J. Liu, Y. Liu, and B. Li, "HRRP target recognition based on soft-boundary deep SVDD with LSTM," in *Proc. Int. Conf. Control, Autom. Inf. Sci. (ICCAIS)*, Oct. 2021, pp. 1047–1052.
[32] T. Inoue, P. Vinayavekhin, S. Morikuni, S. Wang, T. H. Trong, D. Wood, M. Tatsubori, and R. Tachibana, "Detection of anomalous sounds for machine condition monitoring using classification confidence," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, 2020, pp. 66–70.
[33] S. Dieleman, A. V. D. Oord, and K. Simonyan, "The challenge of realistic music generation: Modelling raw audio at scale," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, p. 1–15.
[34] S. K. Mitra and Y. Kuo, *Digital Signal Processing: A Computer-Based Approach*, vol. 2. New York, NY, USA: McGraw-Hill, 2006.
[35] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 4. Cham, Switzerland: Springer, 2006.
[36] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," 2021, *arXiv:2104.02395*.
[37] J. Chen, S. Sathe, C. Aggarwal, and D. Turaga, "Outlier detection with autoencoder ensembles," in *Proc. SIAM Int. Conf. Data Mining*, 2017, pp. 90–98.

GEUNHWAN KIM received the B.S. and M.S. degrees in electronics engineering and the Ph.D. degree in electrical and electronics engineering from Kyungpook National University, Daegu, in 2015, 2017, and 2022, respectively. Since 2022, he has been a Postdoctoral Researcher with the Department of Ocean Systems Engineering, Sejong University, Seoul, Republic of Korea. His research interests include underwater acoustic signal processing, high-duty cycle sonar systems, and deep learning applications for sonar systems.



YOUNGMIN CHOO (Member, IEEE) received the B.A. degree in engineering and the Ph.D. degree in underwater acoustics from Seoul National University, Republic of Korea, in 2005 and 2012, respectively. He was a Postdoctoral Research Fellow at Seoul National University and the Scripps Institution of Oceanography, from 2012 to 2016. Since 2016, he has been an Assistant Professor at the Department of Defense Systems Engineering, Sejong University, Republic of Korea, where he is currently an Associate Professor. He has particular research interest in the areas of signal processing and underwater acoustics. His main research interests include scattering and reverberation from rough boundaries, sonar signal processing of beamforming, and ocean parameter inversion from data.