A DOMAIN COMPATIBILITY AND CONFIDENCE-BASED MODEL SELECTION SYSTEM FOR DOMAIN-AGNOSTIC INCREMENTAL LEARNING

Technical Report

Haruto Takami

Okayama University
Okayama, Japan
p58q2pku@s.okayama-u.ac.jp

Sunao Hara

Okayama University
Okayama, Japan
hara@okayama-u.ac.jp

ABSTRACT

Domain-agnostic Incremental Learning requires a model to adapt to newly introduced domains while preserving previously acquired knowledge under the constraint that past-domain training data cannot be accessed again. In the baseline method, multiple domain-specific classification models are maintained, and the model used for inference is selected based on output entropy. However, using entropy as a confidence criterion often yields high confidence even when the prediction is incorrect. To address this issue, we propose a model selection method that combines a VAE-based domain rejection mechanism with a ConfidNet-based confidence estimation mechanism. First, domain-specific VAEs are used to reject domain models that are not compatible with the input sample. Next, ConfidNet estimates the confidence of the remaining candidate models, and the classification result of the model with the highest confidence score is selected as the final output. Experimental results on the DCASE 2026 Task 7 development dataset demonstrate that the proposed method consistently outperforms the baseline system under all evaluation conditions. In particular, the classification accuracy on the D2 task improved from 54.77% to 70.11%, while the average accuracy on the D3 task improved from 45.50% to 59.09%. These results confirm that the proposed method provides an effective model selection strategy for Domain-agnostic Incremental Learning.

***Index Terms*—** Domain-agnostic Incremental Learning, Sound Classification, Model Selection, Variational Autoencoder, ConfidNet

1. INTRODUCTION

In this report, we describe the system submitted to DCASE 2026 Challenge Task 7 [1]. This task focuses on Domain-agnostic Incremental Learning (DAIL) [2] for sound classification across multiple acoustic domains. Incremental Learning aims to adapt to newly introduced data or domains while preserving previously acquired knowledge. In this task, classifiers are incrementally trained for three different domains (D1, D2, and D3) to classify ten sound event classes. During training, only data from the current domain are available, and access to data from previously learned domains is prohibited. Furthermore, domain labels are unavailable during inference, requiring the system to perform accurate classification across all known domains. Under these conditions, it is essential not only to adapt to new domains while retaining previously acquired knowledge, but also to select an appropriate domain-specific model for each input sample. Therefore, participants are required to develop methods that achieve accurate model selection and sound classification in a domain-agnostic setting.

In the baseline system[2], an input sample is independently fed into all domain-specific models, and the entropy of the resulting class probability distribution is calculated for each model. The model with the lowest entropy is selected, and its classification result is used as the final output. However, using entropy as a confidence criterion often yields high confidence even when the prediction is incorrect. In practice, a model may produce highly confident predictions even when the prediction is incorrect, making accurate domain estimation difficult when relying solely on entropy-based model selection.

To address this issue, we propose a model selection framework that improves the inference-time model selection mechanism. First, a Variational Autoencoder (VAE) is trained for each domain to evaluate the consistency between an input sample and the corresponding domain distribution. Specifically, KL divergence computed from the latent distribution of the VAE is used to reject domain models whose training distributions are substantially inconsistent with the input sample. Next, ConfidNet [3] is applied to the remaining candidate models to estimate prediction confidence. Unlike output probabilities from the classification model, ConfidNet estimates the likelihood that a prediction is correct. The model with the highest estimated confidence is then selected, and its classification result is used as the final output.

2. PROPOSED METHOD

Figure 1 illustrates the overall architecture of the proposed method. For each input audio sample, a VAE-based gating mechanism is first applied to evaluate its domain compatibility with each domain-specific model. Models judged to be substantially inconsistent with their corresponding training distributions are rejected from the candidate set, and classification is performed only using the remaining models. Next, feature representations extracted immediately before the classification layer of each candidate model are fed into the corresponding ConfidNet to estimate prediction confidence. Finally, the classification result produced by the model with the highest estimated confidence is selected as the final prediction.

The proposed method consists of two main components: (1) a VAE-based gating mechanism for rejecting domain models with

This work was supported by JSPS KAKENHI Grant Number JP23K11335.

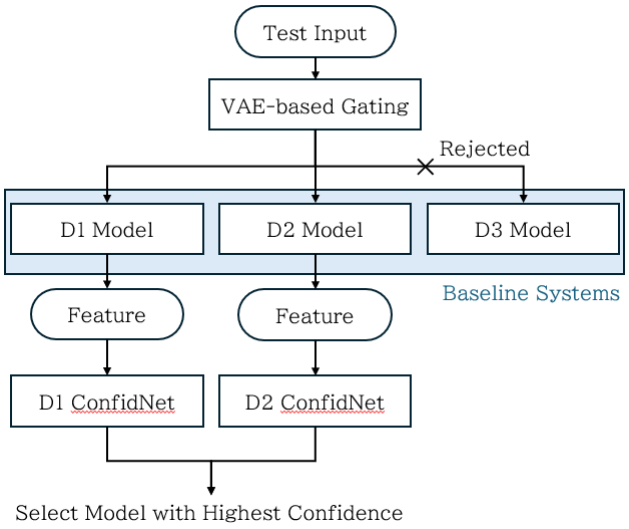

Figure 1: Overview of the proposed system

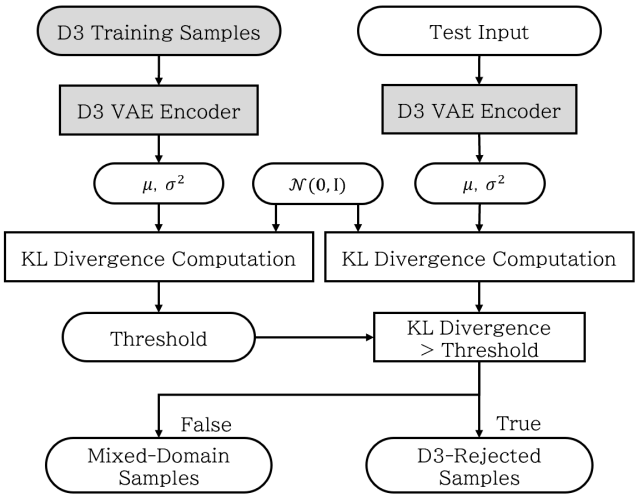

Figure 2: VAE-based gating system

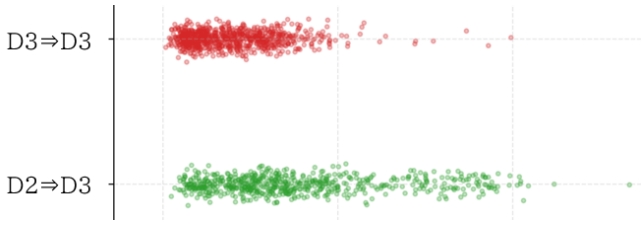

Figure 3: Distribution of KL divergence

low domain compatibility, and (2) a ConfidNet-based model selection mechanism for selecting the most reliable prediction among the remaining candidate models. The details of these components are described in the following sections.

2.1. VAE-based Gating Mechanism

In the proposed method, a VAE is trained for each domain to construct a gating mechanism that determines whether an input sample is unlikely to belong to the corresponding domain.

First, a VAE is trained using the training data of each domain. The VAE consists of an Encoder and a Decoder composed of multiple convolutional layers. The input log-mel spectrogram is mapped into a latent space and then reconstructed by the Decoder. After training, KL divergence is computed from the mean ($\boldsymbol{\mu}$) and variance vectors ($\boldsymbol{\sigma}^2$) of the latent distribution produced by the Encoder. The KL divergence between the latent distribution and the standard normal prior is computed as

$$D_{\mathrm{KL}} = -\frac{1}{2}\sum_{i=1}^{d}\left(1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2\right) \quad (1)$$

where i is the latent dimension index and d is the dimensionality of the latent space.

Figure 2 illustrates the decision process using the D3 model as an example. First, a threshold is estimated from the distribution of KL divergence values obtained from the D3 training data and stored for later use. During inference, an input sample is fed into the D3 VAE Encoder, and its KL divergence is computed and compared with the stored threshold. If the KL divergence is lower than the threshold, the sample is regarded as a mixed-domain sample and forwarded to the subsequent model selection stage. In contrast, if the KL divergence exceeds the threshold, the sample is considered unlikely to belong to the D3 domain, and the D3 model is rejected as a candidate model. The same procedure is applied to the D2 VAE to determine whether the D2 model should be rejected.

Figure 3 shows the distribution of KL divergence values obtained by feeding test samples from each domain into the D3 VAE. When D3 samples are used as input, the KL divergence values tend to be concentrated within a relatively narrow range except for a small number of outliers. In contrast, when D2 samples are used as input, the distribution becomes wider and a larger number of samples exhibit high KL divergence values. This behavior suggests that the VAE strongly reflects the characteristics of the domain observed during training. Furthermore, D1 samples, which are not used during training, are also expected to produce larger KL divergence values. Based on these observations, samples with large KL divergence values are regarded as unlikely to belong to the corresponding domain, and the associated domain model is rejected from the candidate set. This gating mechanism eliminates models with low domain compatibility before model selection, thereby reducing incorrect model selection in the subsequent stage.

Because the task prohibits re-accessing training data from previously learned domains, D2 training data are unavailable when learning the D3 domain. Therefore, the threshold estimated from the training data of each domain is stored during VAE training and reused during inference for domain rejection.

2.2. ConfidNet-based Model Selection

After domain compatibility evaluation by the VAE-based gating mechanism, ConfidNet is applied to the remaining candidate models to determine the final model used for classification. ConfidNet is a multilayer fully connected network designed to estimate the correctness of a prediction from feature representations extracted by a classification model. For each domain-specific baseline model, the

Table 1: Class-wise accuracy (%) of the baseline (left) and proposed (right) methods.

	Last learned domain		
	D2	D3	
Class	D2 Test	D2 Test	D3 Test
alarm	33.08	32.31	64.52
baby_cry	–	–	54.17
bark	78.26	82.61	32.47
engine	79.71	79.71	57.65
fire	22.22	22.22	37.84
footsteps	61.22	63.27	20.90
knock	75.00	72.22	–
telephone_ringing	–	–	51.61
piano	67.01	67.01	93.15
speech	52.26	52.26	13.78
Average	58.6	59.0	47.3

	Last learned domain		
	D2	D3	
Class	D2 Test	D2 Test	D3 Test
alarm	48.46	40.77	58.06
baby_cry	–	–	33.33
bark	73.91	82.61	46.75
engine	73.91	59.42	64.71
fire	27.78	33.33	35.14
footsteps	83.67	87.76	39.55
knock	86.11	77.78	–
telephone_ringing	–	–	35.48
piano	67.01	63.92	93.15
speech	85.43	84.92	47.35
Average	**68.3**	**66.3**	**50.4**

feature vector immediately before the classification layer is used as input, and the corresponding True Class Probability (TCP) obtained from the baseline model is used as the training target. Through this training process, ConfidNet learns to estimate the probability that a classifier's prediction is correct based on the extracted feature representation.

Because training data for D1 are not provided in this task, the D1 ConfidNet model is trained incrementally using both D2 and D3 training data. Specifically, it is first trained using feature representations obtained by feeding D2 samples into the D1 model and is subsequently fine-tuned using D3 samples. In contrast, the D2 and D3 ConfidNet models are trained only with feature representations extracted from their corresponding domain training data.

During inference, the input sample is fed into all candidate models that have not been rejected by the VAE-based gating mechanism. The feature vector immediately before the classification layer of each candidate model is then provided to the corresponding ConfidNet. Finally, the model that produces the highest ConfidNet output is selected, and the classification result of the selected baseline model is used as the final prediction.

3. EXPERIMENTAL EVALUATIONS

3.1. Experimental setups

Experiments were conducted using the DCASE 2026 Task 7 development dataset. The proposed method was built upon the official baseline system[2] and evaluated using classification models corresponding to domains D1, D2, and D3.

For the VAE-based domain suitability evaluation, 64-dimensional log-mel spectrograms extracted by the baseline system were used as input features. The encoder consisted of four convolutional blocks, each comprising a convolution layer, batch normalization layer, ReLU activation function, and max-pooling layer. The decoder was constructed using corresponding transposed convolution layers. All input samples were adjusted to a fixed length of 401 frames by zero-padding or truncation, and the latent space dimension was set to 128. Separate VAEs were trained for the D2 and D3 domains. The models were trained for 100 epochs using the AdamW optimizer with a learning rate of 1.0×10^{-3} and a batch size of 32. The loss function was defined as the sum of the reconstruction loss and KL divergence. For domain rejection, the threshold was determined as the 95th percentile of the KL divergence distribution obtained from the training data of each domain. During inference, the KL divergence of an input sample was compared with the corresponding threshold, and the associated domain model was excluded from the candidate set when the threshold was exceeded.

For the ConfidNet-based model selection, feature vectors extracted immediately before the classification layer of each baseline model were used as inputs. ConfidNet consisted of three fully connected layers with ReLU activation functions and dropout layers inserted between them. The network was trained to regress the True Class Probability (TCP) obtained from the corresponding baseline model using the mean squared error loss. Training was performed for 100 epochs using the Adam optimizer with a learning rate of 1.0×10^{-4} and a batch size of 64. Since the DCASE 2026 Task 7 development dataset does not include domain D1, the D1 ConfidNet was first trained using D2 data and subsequently fine-tuned using D3 data. In contrast, the D2 and D3 ConfidNets were trained only with data from their respective domains to achieve domain-specific confidence estimation.

All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 4090 GPU.

3.2. Experimental results

Table 1 shows the class-wise accuracies of the baseline system and the proposed method. Note that the entries marked as "–" correspond to cases where the corresponding class is not included in the evaluation data for that specific test condition. The Average values are computed as the arithmetic mean of the class-wise accuracies, excluding such unavailable classes. Overall, the proposed method achieves higher performance than the baseline in most classes under both D2 and D3 conditions. For the D2 Test set, the Average accuracy improves from 58.6% to 68.3% when the last learned domain is D2, and from 59.0% to 66.3% when the last learned domain is D3. The proposed method improves classification performance across several classes, including alarm, footsteps, knock, and speech. For the D3 Test set, the Average accuracy also improves from 47.3% to 50.4%. The proposed method shows higher performance than the baseline for multiple classes, including bark, engine, footsteps, and speech. In particular, notable gains are observed for footsteps (20.90% to 39.55%) and speech (13.78% to 47.35%). On the other hand, both methods still show variability across classes, indicating that certain classes remain more challenging than others.

Table 2: Classification accuracy (%) of baseline and proposed methods.

Method	Tested domain	Last learned domain	
		D2	D3
Baseline	D2	54.77	54.77
	D3	–	36.23
	Average across domains	54.77	45.50
Proposed	D2	**70.11**	**66.82**
	D3	–	**51.36**
	Average across domains	**70.11**	**59.09**

Table 2 shows the classification accuracies over the entire test set for each domain. The proposed method outperformed the baseline under all evaluation conditions. For the D2 task, evaluated on the entire D2 test set, classification accuracy improved from 54.77% to 70.11%. For the D3 task, classification accuracy on the D2 test set increased from 54.77% to 66.82%, and on the D3 test set from 36.23% to 51.36%. The Average across domains is computed as the arithmetic mean of the classification accuracies on the D2 and D3 test sets without considering the number of samples in each domain. As a result, the Average across domains improved from 45.50% to 59.09%.

4. CONCLUSION

In this report, we proposed a model selection method that combines a VAE-based domain rejection mechanism and a ConfidNet-based confidence estimation mechanism to improve model selection performance in DAIL. In the proposed method, domain models that were not suitable for an input sample were first excluded from the candidate set using VAE. Then, the final classification result was determined by applying ConfidNet-based confidence estimation to the remaining candidate models. Experimental results on the DCASE 2026 Task 7 development dataset showed that the proposed method achieved higher Average accuracy and improved class-wise accuracies for most classes under both D2 and D3 conditions. These improvements were also reflected in the overall classification accuracy. In the D2 task, classification accuracy improved from 54.77% to 70.11%. In the D3 task, the Average across domains improved from 45.50% to 59.09%. Without modifying the classification models themselves, the proposed method achieved better performance than the baseline system under all evaluation conditions simply by improving the model selection strategy during inference.

5. REFERENCES

[1] R. Casciotti, M. Mulimani, M. Harju, J. R. Jensen, and A. Mesaros, "Domain-agnostic incremental learning for sound classification: A dcase 2026 challenge task," in *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2026.

[2] M. Mulimani and A. Mesaros, "Domain-incremental learning for audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[3] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, and P. Pérez, "Addressing failure prediction by learning model confidence," in *NeurIPS*, 2019.