ENSEMBLE SYSTEM INCLUDING PRE-TRAINED MODEL BASED PROTOTYPES FOR DCASE2026 TASK2

Technical Report

Yanfei Wang

Shanghai Wangshuo Technology Co., Ltd
Shanghai, China
yf116800@126.com

Yongji Sun, FangPing Xie

Shanghai Wangshuo Technology Co., Ltd
Shanghai, China
fandefda@163.com, 823351370@qq.com

ABSTRACT

In this report, we propose EAT based ensemble system to address the Dcase2026 Task 2. We used the pre-trained EAT model and fine-tuned it in the development set. Then, we built a prototype classifier and use the distance to prototypes to get the anomaly score. The system is well generalized and are easy to deploy. The final results are obtained through model ensemble by combining several models including the aforementioned ones, the official baseline and so on. Our final ensemble system has achieved 62.12% in the official scores calculated as a harmonic mean of the area under the curve (AUC) and partial AUC (p = 0.1) over all machine types and domains in the development set.

***Index Terms*—** Anomalous sound detection, ensemble learning, prototype classifier, EAT model, DCASE2026

1 1. INTRODUCTION

DCASE2026 Task 2 addresses Noise-aware Unsupervised Anomalous Sound Detection for Machine Condition Monitoring. The objective is to determine whether a target machine is operating normally or anomalously by using only normal sound clips for training. Compared with previous editions, the 2026 task provides two-channel recordings containing both near-mic and far-mic signals, which introduces additional environmental noise and domain variation. This makes robust feature extraction and domain generalization essential for competitive performance.

We propose an ensemble system built around a pre-trained Efficient Audio Transformer (EAT) model and prototype-based anomaly scoring. The pre-trained model supplies rich audio representations, while the prototype classifier models the distribution of normal samples in the embedding space. Finally, an ensemble of complementary subsystems is fused to obtain the final anomaly score.

1 2. PROPOSED SYSTEM

The overall framework consists of three stages. First, log-mel spectrograms are extracted from the two-channel audio and fed into an EAT backbone that has been fine-tuned on the development data. Second, clip-level embeddings are used to construct normal-class prototypes, and an anomaly score is computed from the distance to the nearest prototype. Third, the EAT-based score is combined with scores from the official baseline autoencoder and other auxiliary models through an ensemble fusion module.

1.1 2.1 EAT Fine-tuning

EAT is a self-supervised audio transformer that has been shown to learn strong utterance-level and frame-level representations efficiently. We initialize the model with weights pre-trained on large-scale audio data and fine-tune all parameters on the DCASE2026 Task 2 development set. During fine-tuning, only normal clips are used, and data augmentation such as mixup and SpecAugment is applied to improve generalization across domains.

1.2 2.2 Prototype Classifier

After fine-tuning, we extract embeddings from the [CLS] token or pooled output for each normal training clip. For each machine type, section, and domain, a prototype is computed as the centroid of the corresponding normal embeddings. At inference time, the embedding of a test clip is compared with the prototypes, and the minimum distance is taken as the anomaly score. Larger distances indicate a higher likelihood of anomaly.

1 3. ENSEMBLE STRATEGY

To enhance robustness, we aggregate anomaly scores from multiple subsystems. In addition to the EAT prototype classifier, the ensemble includes the official DCASE baseline autoencoder and other complementary models. The individual scores are normalized and fused by weighted averaging, where the weights are determined empirically on the development set. This combination leverages the complementary strengths of pre-trained transformer embeddings and classic reconstruction-based detectors.

1 4. EXPERIMENTAL RESULTS

We evaluate the proposed system on the DCASE2026 Task 2 development set using the official metrics: area under the ROC curve (AUC) and partial AUC (pAUC) with p = 0.1. The official score is computed as the harmonic mean of AUC and pAUC over all machine types and domains.
Our final ensemble system achieved an official score of 62.12% on the development set. This result demonstrates that the combination of pre-trained EAT representations, prototype-based anomaly scoring, and model ensemble is effective for the noise-aware unsupervised anomalous sound detection task.

1 5. CONCLUSION

In this report, we presented an ensemble system for DCASE2026 Task 2. The system leverages a pre-trained EAT model fine-tuned on the development data and a prototype classifier for anomaly scoring. By combining multiple models including the official baseline, the final ensemble achieved a competitive official score of 62.12%. The proposed approach is well generalized and easy to deploy, making it suitable for practical machine condition monitoring applications.

1 REFERENCES

[1] http://dcase.community/workshop2026/
[2] T. Nishida, N. Harada, D. Takeuchi, D. Niizumi, K. Imoto, K. Dohi, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "Description and Discussion on DCASE 2026 Challenge Task 2: Noise-aware Unsupervised Anomalous Sound Detection for Machine Condition Monitoring," arXiv e-prints: 2606.01578, 2026.
[3] W. Chen, et al., "Self-Supervised Pre-Training with Efficient Audio Transformer," in Proc. IJCAI, 2024.
[4] NTT CSLab, "dcase2023_task2_baseline_ae," https://github.com/nttcslab/dcase2023_task2_baseline_ae
[5] Jake Snell, Kevin Swersky, Richard S. Zemel, "Prototypical Networks for Few-shot Learning, "Advances in Neural Information Processing Systems (NeurIPS), 2017.
[6]Harada Noboru, Niizumi Daisuke, Takeuchi Daiki, Ohishi, Yasunori, Yasuda Masahiro and Saito Shoichiro, "ToyADMOS2: Another Dataset of Miniature-Machine Operating Sounds for Anomalous Sound Detection under Domain Shift Conditions," in Proc. Detection and Classification of Acous-tic Scenes and Events Workshop (DCASE), 2021, pp. 1-5.
[7]Dohi Kota, Nishida Tomoya, Purohit Harsh, Tanabe Ryo, Endo Takashi, Yamamoto Masaaki, Nikaido Yuki and Kawaguchi Yohei, "MIMII DG: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection for Domain Generalization Task," in Proc. 7th Detec-tion and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022), Nancy, France, Nov. 2022.
[8]N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Yasuda, "First-Shot Anomaly Detection for Machine Condition Monitor-ing: A Domain Generalization Baseline," in Proc. 31st European Signal Processing Conference (EUSIPCO), pp. 191-195, 2023.