# THE DCASE2022 CHALLENGE TASK 2 SYSTEM: ANOMALOUS SOUND DETECTION WITH SELF-SUPERVISED ATTRIBUTE CLASSIFICATION AND GMM-BASED CLUSTERING

Technical Report

*Feiyang Xiao*[1], *Youde Liu*[2], *Jian Guan*[1*], *Yuming Wei*[1], *Qiaoxi Zhu*[3], *Tieran Zheng*[2], *and Jiqing Han*[2]

[1]Group of Intelligent Signal Processing, College of Computer Science and Technology,
Harbin Engineering University, Harbin, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[3]Centre for Audio, Acoustic and Vibration, University of Technology Sydney, Ultimo, Australia

## ABSTRACT

This report describes our submission for DCASE2022 Challenge Task 2, an ensemble system for unsupervised anomalous sound detection (ASD), allowing domain shifts. It integrates two domain generalization methods, a self-supervised attribute classification and a GMM-based clustering for unsupervised ASD. Experiments were conducted on the development dataset of DCASE2022 Challenge Task 2. The results show that our ensemble system can achieve 88.5% in average AUC under the source domain, 78.5% in average AUC under the target domain, and 68.8% in average pAUC.

***Index Terms—*** Anomalous Sound Detection, Spectral-temporal Fusion, Domain Generalization, Self-supervised Classification, Gaussian Mixture Model

## 1. INTRODUCTION

The subject of DCASE2022 Challenge Task 2 is “Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques” [1]. Compared with DCASE2020 Challenge Task 2 [2] and DCASE2021 Challenge Task 2 [3], this task focuses on the domain generalization techniques for anomalous sound detection. Various factors can cause domain shift of machine sounds in the real factory environment, such as differences in operation speed, machine load, and environmental noise. Thus, using domain adaptation techniques is costly or even impractical. Therefore, domain generalization techniques are adopted, using the source domain data to learn common features across domains, which will provide the generalization ability for the model under both source and target domains.

Our submission system is the integration of two ASD methods. Here, a self-supervised attribute classification method is proposed, where the attributes describing audio files are employed as the classification labels. In addition, we also introduce a Gaussian mixture model (GMM) based clustering method for ASD under domain shift conditions. An ensemble strategy [4] is applied to integrate the two methods as our final ensemble system to further improve the detection performance.

## 2. PROPOSED METHOD

### 2.1. Self-supervised Attribute Classification

We adopt our previous work STgram-MFN [5] as the backbone, which presents a spectral-temporal fusion feature, STgram, as the input feature and uses MobileFaceNet (MFN) [6] as the self-supervised classifier. Different from STgram-MFN, we introduce two classifiers after MFN, i.e., attribute classifier and label classifier, to classify attributes and labels of the audio signal, respectively.

Focal loss [7] is employed for the label classifier to mitigate the sample imbalance problem, and the binary cross-entropy loss is adopted for the attribute classifier. We use the negative log mean probability of the labels belonging to the corresponding section as the anomaly score. Following STgram-MFN [5], audio signals with 10 seconds in length as the model input, and our model is trained for all machine types. Based on this basic model, we provide three models as follows:

**Model-1:** We fine-tune our basic model on the samples in the target domain.

**Model-2:** We add section information to attributes for the basic model, which will increase the number of labels.

**Model-3:** We add domain information to attributes for the basic model. Meanwhile, label smoothing is adopted for label classification.

### 2.2. GMM-based Clustering Method

In addition, we introduce a GMM-based clustering system for this task, where global weighted rank pooling (GWRP) [8] is adopted for audio feature representation. Let $X \in \mathbf{R}^{F\times T}$ be the log-Mel spectrogram, where $F$ denotes the frequency dimension, and $T$ is the time dimension. The GWRP of $X$ can be calculated as:

$$gwrp(X_i) = \frac{1}{z(r)} \sum_{j=1}^{T} r^{j-1} X_{i,j}, \tag{1}$$

where $0 \leq r \leq 1$ is a hyper-parameter, and $z(r) = \sum_{j=1}^{T} r^{j-1}$ is a normalization term. Note that $r$ can be set at different values according to machine type. This method is named gwrp-GMM.

Meanwhile, SMOTE [9] is employed to deal with the sample insufficiency by over-sampling the samples in the target domain. In our experiments, the over-sampling ratio between the target and

[*] Corresponding author.

This work was partly supported by the Natural Science Foundation of Heilongjiang Province under Grant No. YQ2020F010.

source domains is set as 0.2. The parameters setting (i.e., the clustering centres of GMM, $r$, and using SMOTE or not) is provided in Table 1.

### 2.3. Ensemble

We adopt the ensemble learning strategy to combine the proposed self-supervised attribute classification method and GMM-based clustering method as our ensemble systems:

**Ensemble-1:** Integration of gwrp-GMM, Model-1, Model-2, and Model-3.

**Ensemble-2:** Integration of gwrp-SMOTE-GMM, Model-1, Model-2, and Model-3.

The weights for system integration is shown in Table 2.

## 3. EXPERIMENTS

Regarding the self-supervised method, the experiments on the development and additional training dataset from the ToyADMOS dataset [10] and MIMII DG dataset [11]. We train our models on training data of the development dataset and additional training dataset. The experimental setup for model training is the same as STgram-MFN [5]. For the clustering method, experiments are performed on the development dataset to train GMM models. **We submit four systems to DCASE2022 Challenge task 2, namely gwrp-GMM (system-1), gwrp-SMOTE-GMM (system-2), Ensemble-1 (system-3), and Ensemble-2 (system-4).**

These systems are evaluated on the test data of the development dataset. The average (harmonic mean) results of AUC under source domain (AUC-S), AUC under target domain (AUC-T) and partial AUC (pAUC) for each machine type is shown in Table 3, 4 and 5, respectively. All of our proposed models and systems outperform the baseline systems.

## 4. CONCLUSION

We presented our submission systems for DCASE2022 Challenge Task 2 in this technical report, using two proposed methods (i.e., self-supervised attribute classification method and GMM-based clustering method) and model integrations applying the ensemble learning strategy. Experimental results show that our proposed systems outperformed the baseline systems.

## 5. REFERENCES

[1] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on DCASE 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," *In arXiv e-prints: 2206.05876*, 2022.

[2] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, November 2020, pp. 81–85.

[3] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions," *arXiv e-prints: 2106.04492, 1–5*, 2021.

[4] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[5] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 816–820.

[6] S. Chen, Y. Liu, X. Gao, and Z. Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," in *Proceedings of Chinese Conference on Biometric Recognition (CCBR)*. Springer, 2018, pp. 428–438.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[8] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Proceedings of European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 695–711.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research (JAIR)*, vol. 16, pp. 321–357, 2002.

[10] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," in *Proceedings of Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, Barcelona, Spain, November 2021, pp. 1–5.

[11] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," *arXiv e-prints: 2205.13879*, 2022.

Table 1: The parameters (cluster centers / $r$ / using SMOTE ) for GMM models.
("N" means not using SMOTE, "Y" means using SMOTE)

| Methods | ToyCar | ToyTrain | bearing | fan | gearbox | slider | valve |
|---|---|---|---|---|---|---|---|
| gwrp-GMM | 2 / 0.99 / N | 2 / 0.81 / N | 2 / 1.00 / N | 2 / 1.00 / N | 2 / 0.99 / N | 1 / 0.88 / N | 2 / 0.45 / N |
| gwrp-SMOTE-GMM | 1 / 1.00 / Y | 2 / 0.81 / N | 2 / 1.00 / N | 2 / 1.00 / N | 2 / 0.98 / Y | 1 / 0.90 / Y | 2 / 0.50 / Y |

Table 2: The weights (gwrp-GMM / gwrp-SMOTE-GMM : Model-1 : Model-2 : Model-3) for system integration with ensemble learning strategy.

| Systems | ToyCar | ToyTrain | bearing | fan | gearbox | slider | valve |
|---|---|---|---|---|---|---|---|
| Ensemble-1 | 0.3: 0: 0.7: 0 | 0.4: 0: 0.6: 0 | 0.1: 0.2: 0: 0.7 | 0.8: 0: 0.1: 0.1 | 0.8: 0.1: 0: 0.1 | 0.7: 0: 0.3: 0 | 0.4: 0: 0: 0.6 |
| Ensemble-2 | 0.4: 0.1: 0.5: 0 | 0.4: 0: 0.6: 0 | 0.1: 0.1: 0: 0.8 | 0.8: 0: 0.1: 0.1 | 0.7: 0.2: 0: 0.1 | 0.8: 0: 0.1: 0.1 | 0.4: 0: 0.1: 0.5 |

Table 3: The average (harmonic mean) results of AUC-S (%) for different machine types.

| Methods | ToyCar | ToyTrain | bearing | fan | gearbox | slider | valve | Average |
|---|---|---|---|---|---|---|---|---|
| *Baseline* | | | | | | | | |
| AE-baseline | 91.7(90.4) | 77.0(76.3) | 57.0(54.4) | 79.0(78.6) | 69.2(68.9) | 78.8(78.0) | 52.1(52.0) | 72.1(71.2) |
| MobileNetV2-baseline | 61.2(59.1) | 60.4(57.3) | 63.1(60.6) | 71.5(70.8) | 70.4(69.2) | 69.8(65.2) | 68.8(67.1) | 66.5(64.2) |
| *Self-supervised attribute classification* | | | | | | | | |
| Model-1 | 68.4(66.3) | 85.7(85.5) | 57.2(55.6) | 73.8(69.9) | 73.7(71.4) | 90.1(89.5) | 75.0(71.2) | 74.8(72.8) |
| Model-2 | 46.9(41.6) | 64.9(63.8) | 77.7(77.5) | 73.6(71.7) | 76.5(74.0) | 93.3(93.2) | 73.4(71.1) | 72.3(70.4) |
| Model-3 | 84.5(82.1) | 93.3(93.0) | 63.1(61.8) | 81.8(80.3) | 80.9(80.5) | 93.3(93.1) | 73.3(70.5) | 81.5(80.2) |
| *GMM-based Clustering* | | | | | | | | |
| **gwrp-GMM (system-1)** | 90.2(89.7) | 91.4(90.8) | 60.2(52.5) | 80.2(79.7) | 84.6(83.5) | 95.7(95.6) | 95.7(95.7) | 85.4(83.9) |
| **gwrp-SMOTE-GMM (system-2)** | 89.7(89.1) | 91.5(90.9) | 62.1(56.9) | 80.2(79.7) | 84.6(83.4) | 95.1(95.0) | 93.6(93.5) | 85.3(84.1) |
| *Ensemble* | | | | | | | | |
| **Ensemble-1 (system-3)** | 82.7(80.7) | 92.8(92.4) | 77.2(77.2) | 86.0(85.8) | 88.8(88.3) | 96.4(96.3) | 96.3(96.2) | 88.6(88.1) |
| **Ensemble-2 (system-4)** | 84.5(83.0) | 92.8(92.5) | 77.2(77.2) | 86.0(85.8) | 88.1(87.6) | 96.0(95.9) | 95.2(95.1) | 88.5(88.1) |

Table 4: The average (harmonic mean) results of AUC-T (%) for different machine types.

| Methods | ToyCar | ToyTrain | bearing | fan | gearbox | slider | valve | Average |
|---|---|---|---|---|---|---|---|---|
| *Baseline* | | | | | | | | |
| AE-baseline | 36.6(34.8) | 26.4(23.4) | 59.0(58.4) | 49.2(47.2) | 62.8(62.4) | 49.0(47.7) | 49.9(49.5) | 47.6(46.2) |
| MobileNetV2-baseline | 52.8(52.0) | 46.3(45.9) | 61.8(59.9) | 51.8(48.2) | 59.0(56.2) | 48.6(38.2) | 60.9(57.2) | 54.5(51.1) |
| *Self-supervised attribute classification* | | | | | | | | |
| Model-1 | 83.6(82.7) | 58.2(52.1) | 74.8(74.7) | 62.8(62.7) | 64.5(62.7) | 76.2(75.5) | 57.3(52.6) | 68.2(66.1) |
| Model-2 | 71.4(66.1) | 53.5(48.8) | 72.2(72.0) | 48.3(41.3) | 61.2(56.9) | 71.5(65.5) | 57.5(44.7) | 62.2(56.5) |
| Model-3 | 73.1(67.0) | 43.2(39.0) | 68.6(67.2) | 53.0(52.5) | 69.0(59.8) | 69.3(59.8) | 50.7(32.3) | 61.0(54.9) |
| *GMM-based Clustering* | | | | | | | | |
| **gwrp-GMM (system-1)** | 73.7(71.6) | 51.2(44.8) | 87.4(86.3) | 65.9(63.2) | 78.7(77.0) | 81.6(80.3) | 89.6(89.6) | 75.4(73.3) |
| **gwrp-SMOTE-GMM (system-2)** | 81.1(80.3) | 52.0(45.1) | 85.4(84.7) | 65.9(63.2) | 81.8(81.0) | 85.6(84.8) | 89.9(89.9) | 77.4(75.6) |
| *Ensemble* | | | | | | | | |
| **Ensemble-1 (system-3)** | 84.7(83.6) | 56.8(49.6) | 77.0(77.0) | 69.8(69.2) | 79.9(78.6) | 84.1(82.3) | 90.8(90.7) | 77.6(75.9) |
| **Ensemble-2 (system-4)** | 86.0(85.3) | 57.1(49.6) | 77.3(77.1) | 69.8(69.2) | 81.6(80.7) | 86.3(85.4) | 91.4(91.3) | 78.5(77.0) |

Table 5: The average (harmonic mean) results of pAUC (%) for different machine types.

| Methods | ToyCar | ToyTrain | bearing | fan | gearbox | slider | valve | Average |
|---|---|---|---|---|---|---|---|---|
| *Baseline* | | | | | | | | |
| AE-baseline | 52.8(52.7) | 50.6(50.5) | 52.2(52.0) | 58.0(57.5) | 58.7(58.5) | 56.0(55.8) | 50.4(50.4) | 54.1(53.9) |
| MobileNetV2-baseline | 52.5(52.3) | 51.6(51.5) | 57.9(57.1) | 57.6(56.9) | 56.5(56.0) | 56.5(54.7) | 65.3(62.4) | 56.9(55.8) |
| *Self-supervised attribute classification* | | | | | | | | |
| Model-1 | 62.0(59.9) | 58.7(57.8) | 52.7(52.6) | 59.1(58.5) | 56.1(56.0) | 61.8(61.7) | 63.5(60.3) | 59.1(58.1) |
| Model-2 | 52.4(52.3) | 50.6(50.5) | 65.6(65.4) | 59.8(58.9) | 58.3(57.9) | 66.2(63.8) | 62.9(60.3) | 59.4(58.4) |
| Model-3 | 66.8(62.2) | 55.1(54.4) | 58.4(58.0) | 60.5(60.5) | 56.7(56.0) | 66.0(63.9) | 62.6(58.9) | 60.9(58.9) |
| *GMM-based Clustering* | | | | | | | | |
| **gwrp-GMM (system-1)** | 59.4(58.1) | 60.2(59.1) | 55.8(54.7) | 63.0(62.1) | 65.9(64.7) | 74.3(73.3) | 70.6(70.2) | 64.2(63.2) |
| **gwrp-SMOTE-GMM (system-2)** | 62.5(60.5) | 61.1(59.7) | 55.3(54.2) | 63.0(62.1) | 66.9(65.8) | 77.4(76.7) | 66.1(66.0) | 64.6(63.6) |
| *Ensemble* | | | | | | | | |
| **Ensemble-1 (system-3)** | 66.6(63.1) | 61.5(60.0) | 67.3(67.0) | 66.2(65.9) | 66.2(64.8) | 76.2(74.9) | 74.8(73.5) | 68.4(67.0) |
| **Ensemble-2 (system-4)** | 66.5(63.5) | 61.9(60.2) | 66.8(66.6) | 66.2(65.9) | 67.8(66.7) | 79.1(78.2) | 73.2(72.5) | 68.8(67.6) |