MULTI-LABEL OPEN-SET AUDIO CLASSIFICATION

Sripathi Sridhar[*], *Mark Cartwright*

Sound Interaction and Computing (SInC) Lab, New Jersey Institute of Technology
{ss645, mark.cartwright}@njit.edu

ABSTRACT

Current audio classification models have small class vocabularies relative to the large number of sound event classes of interest in the real world. Thus, they provide a limited view of the world that may miss important yet unexpected or unknown sound events. To address this issue, open-set audio classification techniques have been developed to detect sound events from unknown classes. Although these methods have been applied to a multi-class context in audio, such as sound scene classification, they have yet to be investigated for polyphonic audio in which sound events overlap, requiring the use of multi-label models. In this study, we establish the problem of multi-label open-set audio classification by creating a dataset with varying unknown class distributions and evaluating baseline approaches built upon existing techniques.

***Index Terms*—** Open-set, multi-label, audio classification, dataset

1. INTRODUCTION

Audio classification (AC), the machine listening task of identifying sound events in an audio recording, has typically been studied as two task variants, i.e. multi-class AC, where the input recordings are expected to contain only one event, and multi-label AC, where the input recordings may contain multiple overlapping sound events. Real-world audio recordings in typical urban, domestic or environmental settings often contain multiple sound sources of anthrophony, biophony, and geophony, and thus, are better modeled as a multi-label AC task.

Multi-label AC is a common machine listening task that has been applied to various scenarios such as urban sound data [1], everyday environments [2], and music [3]. Much of this work however assumes a small fixed class vocabulary, a closed-set task, which does not reflect real-world scenarios. Everyday sound scenes consist of sources drawn from hundreds if not thousands of classes depending on the class granularity of interest, and people are constantly exposed to novel classes, e.g., those from new or uncommon technology and animal vocalizations. To the “ears” of these models, unknown sound classes simply do not exist or — possibly worse — are confused with known classes. This limited class vocabulary size can be attributed to the cost and difficulty of annotating large-scale audio datasets. However, the result of this barrier is a limited view of the acoustic world by AC models that may miss important yet unexpected or unknown sound events, hindering machine listening’s transformative potential.

One solution to this problem is to build models with a dynamic vocabulary that can be updated in a lightweight manner without having to retrain the model from scratch. An example of this approach is few-shot classification [4], which is often formulated within a meta-learning framework where a model can learn a new class from a small ‘support set’ of examples [5]. Prior work has applied this to tasks such as instrument recognition [6], multi-label audio classification [7], and multi-label drum transcription [8]. However, this method still requires the user or researcher to supply a support set for unseen or novel classes [7], and thus, such supervised approaches are only useful if you know what you are hoping to find and have examples of it. In many situations — e.g., urban noise monitoring, audio accessibility, bioacoustic monitoring — it is the rare events and unexpected events that are arguably the most important to detect, i.e., the machine listening equivalent of a “black swan event” [9]. To this end, we focus on detecting the presence of unknown classes in addition to known classes, referred to as open-set modeling.

Open-set modeling has seen research interest in the image domain for several years [10, 11], but it has only more recently gained interest in the audio domain and been applied to tasks such as domestic sound classification [12], acoustic scene classification [13, 14], and the related yet distinct task of anomalous sound detection [15]. However, all of these tasks are binary or multi-class AC — to the best of our knowledge, open-set modeling has not been applied to multi-label AC.

As in [12, 10], we define *known known* (KK) classes as known (i.e., in-vocabularly) classes seen during training and inference, *known unknown* (KU) classes as unknown (i.e., out-of-vocabulary) classes seen during training and inference, and *unknown unknown* (UU) as unknown classes seen only at inference. A fourth category, unknown known (UK) classes, are classes in which only semantic or metadata information is available in the absence of discrete labels — this category is not considered in this work. We collectively refer to KU and UU as unknown classes, and KK as known classes.

We define multi-label open-set AC (MLOS) as the task of assigning between 0 and $|\text{KK}| + 1$ class labels to an audio recording, where $|\text{KK}|$ is the cardinality of the set of known classes and $+1$ refers to the label indicating the presence of an unknown sound class. Thus, an MLOS model needs to both estimate which known classes are present as well as decide whether at least one unknown class is present. This is in contrast to multi-class open-set AC models which assign only 1 of $|\text{KK}| + 1$ class labels to an audio recording.

In this paper, we (1) establish the problem of MLOS, (2) introduce a new dataset with varying unknown class distributions to investigate this problem, and (3) evaluate baseline approaches comprised of combinations of existing machine listening techniques.

2. DATASET

Prior open-set AC datasets are either multi-class [12] or focused on binary anomalous sound detection [16]. In order to establish the

[*] This work is partially supported by NSF award #1955357

MLOS task, we are interested in exploring the effects of polyphony, and levels of "openness" while working with a large class vocabulary. While few-shot datasets like FSD-MIX-CLIPS [7] meet the polyphony criteria, they do not have varying levels of "openness" nor dataset variants where different classes are assigned to the KK, KU and UU categories. As in [17], we define "openness" as

$$O^{*} = 1 - \sqrt{(2 \times C_{tr}) / (C_{tr} + C_{te})}, \qquad (1)$$

where $C_{tr} = |\text{KK} \cup \text{KU}|$ is the number of classes seen during training and $C_{te} = |\text{KK} \cup \text{KU} \cup \text{UU}|$ the number of classes seen during testing. Thus, for larger C_{tr}, we assign lower values of openness.

To this end, we develop a new dataset of synthetic soundscapes using open-set criteria. As in FSD-MIX-CLIPS, we use a subset of FSD50K where each clip has a single 'present and predominant' label, i.e., the labeled sound event is the only type of sound present with the exception of mild background noise [18]. This gives us 7600 source events from 89 classes, each between 0.5s and 4s in duration. We use only the leaf node labels according to the Audioset ontology [19]. Hereafter we refer to this subset of FSD50K as the *source dataset*.

First, we split the classes into 5 subsets of 18 classes each (except for one subset with 17 classes), and from these subsets, we create 10 variations of class assignments into KK, KU, and UU as shown in Table 1 — 5 with a low degree of openness and 5 with a high degree of openness, i.e. no KU classes. The openness coefficients are $O^{*} = 0.05$ or 0.06 for low openness ($C_{tr} = 72$ or 71) and $O^{*} = 0.13$ or 0.14 for high openness ($C_{tr} = 54$ or 53). For each class assignment variation i, we generate an intermediate dataset called 'Open-Set Soundscape-i' (OSS-i), consisting of 10s 44.1kHz synthetic soundscapes using Scaper [20] — 200k training, 30k validation, and 30k test with no source overlap between splits. The training and validation sets are synthesized from only the known class subsets, e.g. in dataset variant 1, from L1-L4 in the low openness case and H1-H3 in the high openness case (Var. 1 in Table 1). In both openness cases, the test set is synthesized using all the subsets. Additionally, we also create a small tuning validation set using all the subsets for hyperparameter tuning, ensuring no example overlap with the test set.

In each OSS-i, we maintain the class distribution of the source dataset as closely as possible while enforcing a minimum of 200 examples per class. Each soundscape has one to four overlapping source sound events in the foreground, which we place between 0 to 9s in the soundscape. We augment each source with pitch shifting (-2 to +2 semitones) and time stretching (by a factor of 0.8 to 1.2). We use uniform random sampling for all augmentations during generation.

For each OSS-i dataset variant, we generate a dataset of 1s clips by centering a window on each event in the 10s soundscape and labeling a class as present if it overlaps with this window. This yields 10 datasets (5 high, 5 low openness) with ∼500k clips each.

We refer to this as the Open-Set Tagging (OST) dataset and use it to train and evaluate our models. Both OSS and OST datasets are publicly available [1].

3. MODELS

In this study, for the sake of brevity we focus on the high openness MLOS task, as it is the more challenging scenario. Therefore in the following we use D_k to denote the set of known classes seen during training, and D_u for the set of unknown classes seen only during inference.

Openness	Low					High				
Subset	L1	L2	L3	L4	L5	H1	H2	H3	H4	H5
Var. 1	KK	KK	KK	KU	UU	KK	KK	KK	UU	UU
Var. 2	UU	KK	KK	KK	KU	UU	KK	KK	KK	UU
Var. 3	KU	UU	KK	KK	KK	UU	UU	KK	KK	KK
Var. 4	KK	KU	UU	KK	KK	KK	UU	UU	KK	KK
Var. 5	KK	KK	KU	UU	KK	KK	KK	UU	UU	KK

Table 1: Class splits for high and low openness dataset variations

In this section, we present five baseline models, two of which use oracle sources as a way of further exploring the limitations of these approaches.

3.1. Multi-label

Given a multi-label input example x, the classifier C generates a logit vector $\mathbf{v} = C(x) \in \mathbb{R}^{N}$, where $N := |D_k|$ i.e. KK classes present during training. To estimate whether the input contains a class in D_k, we take the indices above a threshold λ, i.e. $\{j : v_j > \lambda; j \in [0, N-1]\}$.

Our baseline approach to the MLOS task is to run inference using a standard multi-label classifier. Then, to predict the unknown class we use the open-set decision criteria discussed later in this section.

The classifier consists of two stages. The first stage is a frozen OpenL3 encoder pre-trained on the environmental subset of Audioset [21], which has shown competitive performance across a variety of audio and music classification tasks in the NeurIPS HEAR 2021 challenge [22]. The encoder input is a 256 frequency bin log-melspectrogram input, with output embeddings of dimension 6144.

The second stage is a multi-layer perceptron (MLP) with five dense layers. Each layer consists of 1024 units and ReLU activation. The number of output units depends on the number of classes in the dataset variant, i.e. $|D_k|$ classes. This system is depicted in Figure 1.

The multi-label classifier output has sigmoid activations and is trained using binary cross-entropy loss. Instead of using a threshold in our experiments, we used an overly-optimistic oracle strategy, picking the m sources with the highest logits, where m is the polyphony from the ground-truth data. We use the checkpoint with the best validation loss for evaluation.

3.2. Combinatorial multi-class

In order to isolate the effect of multi-label training, we include a 'combinatorial multi-class' model. Here we map each unique label combination in the OST training set to a class ID, effectively creating a multi-class model training setup. While OST has around 8000 unique class combinations, we note that this approach would lead to a 'combinatorial explosion' and may be infeasible as the number of classes and unique combinations increase.

Apart from a categorical cross-entropy loss function and different number of output layer units, we use the same architecture and training setup as described in Section 3.1.

3.3. Source estimates multi-class PIT

Since prior work on open-set AC has been in the multi-class setup, we include a model with a universal source separation front-end

[1] 10.5281/zenodo.7241704

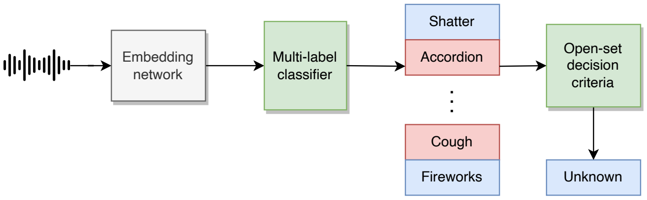

Figure 1: Multi-label model consisting of a pre-trained frozen OpenL3 embedding network and a MLP classifier.

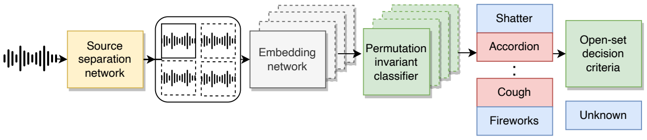
Source separation network
Embedding network
Permutation invariant classifier
Shatter
Accordion
Cough
Fireworks
Open-set decision criteria
Unknown

Figure 2: Source estimates multi-class PIT model consisting of a separation network and multi-class classifier trained using a permutation invariant loss. The separation network is trained separately using MixIT, then its weights are frozen as the classifier is trained.

module to convert the MLOS task to a set of multi-class open-set classification tasks and leverage existing approaches for these subtasks. Related prior work successfully used such a universal source separation pre-processing step to improve classification precision in multi-label closed-set birdsong classification [23]. A separation model generates source estimates for a multi-class classifier that generates predictions. We hypothesize that the separation model will also improve performance on the MLOS task, particularly on unseen known class combinations (which could be misclassified by an open-set model as *unknown* if estimated as a whole) and for clips with high polyphony. However, this approach does come with the risk of error propagation from the separation model to the classifier caused by poor source estimates.

Given an input example x, the separation model S generates eight source estimates s_i. Using m of these eight source estimates as input, the multi-class classifier C generates logits $\mathbf{v}_i = C(s_i) \in \mathbb{R}^N$ for each source s_i, where again $N := |D_k|$, and we use the class with the max logit as our known class prediction for that source, i.e. $\mathrm{argmax}(\mathbf{v}_i)$.

The separation network is a TDCN++ model trained on unlabeled polyphonic mixtures using mixture invariant training (MixIT) [24]. As the authors of [23, 24] note the importance of training MixIT on the target domain for quality source estimates, we train from scratch on data from all variants of the OST dataset for 1M steps and use the checkpoint with the best validation performance. Estimating the number of actual sources from the 8 fixed outputs is a challenging task and a potential failure point. In this paper, we opt for an overly optimistic scenario and use an oracle pruning strategy for testing. We pick the m source estimates with the highest energy, where m is the number of ground-truth sources. We follow this protocol both during training and inference. An existing risk of this approach is that the chosen source estimate may only contain background for input examples with low SNR. Additionally, this protocol may be sub-optimal if the model over-separates, especially in examples with low polyphony.

The multi-class classifier has the same architecture as in Section 3.1, and is trained using a permutation invariant cross-entropy loss [24]. Since the label assignment is only available at the clip level, we generate a prediction for each source estimate and compute the total loss for $m!$ label-source combinations. The best match that minimizes the total loss is used to update the model weights. We use the suffix permutation invariant training (PIT) to denote that a model is trained this way. The model is depicted in Figure 2.

3.4. Oracle sources multi-class PIT

In order to understand the effects of error propagation due to the separation network, we train a model with a perfect separation model, i.e. with the oracle sources. These oracle sources when re-combined yield the OST clips used to train the multi-label classifier model. We use the same model and training setup as in Section 3.3.

3.5. Oracle sources multi-class model

A key limitation of PIT is that it does not guarantee accurate source-label matching during training. In order to further isolate the effect that PIT may have on performance, we evaluate a reference multi-class model with the same architecture trained with oracle sources using standard cross-entropy loss. Given our modeling choices, this serves as an expected upper bound in terms of performance, as it is a true multi-class model.

3.6. Open-set decision criteria

We evaluate two simple open-set decision criteria that have been used previously in multi-class open-set studies. Here, we use these techniques both in the multi-class and multi-label configurations, however, the latter would suffer from false positives in scenarios with no activity or background noise events.

The first approach is softmax thresholding, where the maximum softmax probability (MSP) is compared against a threshold δ [25]–where a model predicts *unknown* if it is below and *known* otherwise. Let $\hat{\mathbf{y}}$ be the classifier output for models without separation, e.g. $\hat{\mathbf{y}} = \mathrm{sigmoid}(\mathbf{v})$, and $\hat{y}_o \in \{0, 1\}$ the open-set prediction, with 0 and 1 denoting a known and unknown class prediction respectively, then

$$\hat{y}_o = \Big\{ 1 \text{ if } \max(\hat{\mathbf{y}}) < \delta; \text{ else } 0 \qquad (2)$$

For PIT models and the oracle sources multi-class model, we predict unknown if any of the *m* source estimates contain an unknown class:

$$\hat{y}_o = \Big\{ 1 \text{ if } \max(\hat{\mathbf{y}}_i) < \delta, \text{ for } i \in [0, m-1]; \text{ else } 0 \qquad (3)$$

where $\hat{\mathbf{y}}_i$ is the classifier output for a source estimate.

The second approach is Openmax [26], which aims to correct ‘overconfident’ model predictions when the example is less likely to belong to the training distribution of the predicted class. Openmax re-weights the logit vector by penalizing the top α ranked logits using models of the training distribution tail for each class. The class-specific models are parameterized by the Weibull distribution tail size τ and logit rank limit α. It also computes an unknown class probability p_u based on the degree of recalibration needed, which is then appended to the updated classifier output. We refer the reader to [26] for further details.

For models without separation, we compute the updated classifier output $\hat{\mathbf{y}}_w$ using the re-weighted logit vector $\mathbf{v}_w$, e.g. $\hat{\mathbf{y}}_w = \mathrm{sigmoid}(\mathbf{v}_w)$. Then, similar to Equation 2–

$$\hat{y}_o = \Big\{ 1 \text{ if } \max(\hat{\mathbf{y}}_w) < \delta \text{ or } \max(\hat{\mathbf{y}}_w) = p_u; \text{ else } 0 \qquad (4)$$

	Accuracy (SD)	
	MSP	Openmax
Multi-label	57.4 (2.9)	–
Source estimates PIT	54.3 (3.0)	–
Oracle sources PIT	**59.7** (4.7)	**61.3** (3.0)
Combinatorial multi-class	59.1 (1.5)	–
Oracle sources multi-class	**61.1** (3.8)	**61.2** (3.8)

Table 2: Unknown detection results using maximum softmax probability thresholding (MSP) and openmax. All results are accuracy averaged over the five dataset variants, with standard deviations in parentheses.

For models with separation we apply this re-weighting and thresholding protocol to $\mathbf{v}_i$, the source estimate logit vectors.

We tune δ, τ, and α on the tuning validation set using Optuna, a Python package for efficient hyperparameter optimization [27], and use hyperparameters from the trial that maximizes unknown detection accuracy.

4. EVALUATION

We evaluate the models separately on closed-set classification and unknown detection. For the former, we evaluate the model only on examples without unknown classes. For the latter, we evaluate the models on all examples at the clip level for a binary classification task. We present the unknown detection results in Table 2 and closed set classification results in Table 3.

From Table 2, we note that the multi-label model is worse than the oracle sources multi-class model. In this dataset, every example has at least one source, however, in scenarios where no event may be present we expect this gap to be larger, as the multi-label model may generate more false positives during silence or background noise events.

Combinatorial multi-class is only slightly worse than oracle sources multi-class. While this is an interesting finding, there are two key limitations. This model does not scale well as the number of classes increases, leading to the 'combinatorial explosion' issue [28]. Furthermore, this dataset follows the imbalanced source dataset distribution making certain known classes more likely than others, meaning that the model does not encounter new class combinations in the test set, leading to an optimistic view of its unknown detection accuracy. We expect this model to perform poorly in scenarios with unseen combinations of known classes, potentially generating false positives.

Oracle sources PIT does better than the multi-label model by about 4%, which suggests that a perfect universal source separator could improve performance on this task. However, the gap is smaller than expected, potentially due to false positives caused by overconfident model predictions [26]. We see some evidence of this in Table 2 where Openmax accuracy for the oracle sources PIT model is better than its MSP accuracy, suggesting that this model is falsely overconfident for examples containing unknown class events.

We also note that oracle sources multi-class is better than oracle sources PIT by about 2%– since they are both trained on the same data, the difference must be due to PIT.

Finally, source estimates PIT is not as good as the oracle sources PIT model, and in fact, performs worse than the multi-label model. This indicates that more research may be needed for universal source separation models to be useful in this task. Some prior results suggest that training the classifier together on the input mixture and source estimates may improve closed-set classification [23], but it remains to be seen whether this translates to unknown detection where the model needs to separate out unknown class events as well.

	Micro F1	**Macro F1**	**mAP**
Multi-label	0.449 (0.01)	0.349 (0.02)	0.400 (0.02)
Source Estimates PIT	0.407 (0.01)	0.332 (0.01)	0.347 (0.01)
Oracle Sources PIT	**0.511** (0.02)	**0.461** (0.04)	**0.501** (0.04)
Oracle sources multi-class	**0.581** (0.01)	**0.541** (0.01)	**0.590** (0.01)

Table 3: Closed-set classification results on 53 or 54 classes, depending on the dataset variant. All metrics are averaged over the five dataset variants, with standard deviations in parentheses.

We notice similar trends in closed-set classification (Table 3) as in unknown detection MSP accuracy. The multi-label model as well as the oracle sources PIT model perform significantly worse than the oracle sources multi-class model, which is in line with the expectation of multi-label classification being a more challenging task. Oracle sources PIT does better than the multi-label model, which suggests that a perfect source separation model would be useful. Lastly, the overall modest performance of the oracle sources multi-class model on both closed- and open-set tasks suggests that better audio representations are also needed to improve performance.

5. DISCUSSION AND CONCLUSION

In this work, we introduced the multi-label open-set audio classification (MLOS) task and developed a synthetic dataset with varying unknown class distributions. We then presented several baseline models using combinations of existing machine listening techniques and evaluated their performance on known class and unknown class metrics.

We show that MLOS is a challenging task that existing approaches alone cannot adequately solve. In our study, we find that a perfect source separation model may be useful for MLOS, but further research is needed for universal source separation models to provide similar improvements in open-set classification.

While we see some interesting results, some other questions were raised, such as how unseen known class combinations might affect unknown class detection, particularly for the multi-label and combinatorial multi-class models. We plan to evaluate this by varying vocabulary and dataset size to control the ratio of seen and unseen known class combinations in the test set.

Moreover, we consider here a simplistic data scenario where there is always at least one sound present. We plan to investigate how the inclusion of background event classes would affect some of the models discussed here, such as the multi-label and source estimates multi-class PIT.

By sharing the dataset and these baseline results, we hope to invite further interest from the community to this under-explored area of research.

6. ACKNOWLEDGMENT

We would like to thank Simran Kaur for helping to generate the initial version of the dataset.

7. REFERENCES

[1] M. Cartwright, A. E. M. Mendez, J. Cramer, V. Lostanlen, G. Dove, H.-H. Wu, J. Salamon, O. Nov, and J. P. Bello, "SONYC Urban Sound Tagging (SONYC-UST): A Multilabel dataset from an urban acoustic sensor network," *Proc. of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019).*

[2] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *2015 International Joint Conference on Neural Networks (IJCNN)*, July 2015, pp. 1–7, iSSN: 2161-4407.

[3] S. Gururani, C. Summers, and A. Lerch, "Instrument activity detection in polyphonic music using deep neural networks." in *ISMIR*, 2018, pp. 569–576.

[4] M. Fink, "Object classification from a single example utilizing class relevance metrics," *Advances in neural information processing systems*, vol. 17, 2004.

[5] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[6] H. Flores Garcia, A. Aguilar, E. Manilow, and B. Pardo, "Leveraging hierarchical structures for few-shot musical instrument recognition," in *Proc. of the 22nd Int. Society for Music Information Retrieval Conference*, 2021.

[7] Y. Wang, N. J. Bryan, M. Cartwright, J. Pablo Bello, and J. Salamon, "Few-Shot Continual Learning for Audio Classification," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

[8] Y. Wang, J. Salamon, M. Cartwright, N. J. Bryan, and J. P. Bello, "Few-Shot Drum Transcription in Polyphonic Music," *International Society for Music Information Retrieval Conference*, 2020.

[9] N. N. Taleb, *The black swan: The impact of the highly improbable*. Random house, 2007, vol. 2.

[10] A. Bendale and T. Boult, "Towards Open World Recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA.

[11] H. Zhang, A. Li, J. Guo, and Y. Guo, "Hybrid models for open set recognition," in *European Conference on Computer Vision*. Springer, 2020, pp. 102–117.

[12] J. Naranjo-Alcazar, S. Perez-Castanos, P. Zuccarello, F. Antonacci, and M. Cobos, "Open Set Audio Classification Using Autoencoders Trained on Few Data," *Sensors*, vol. 20, no. 13, p. 3741, July 2020.

[13] Z. Kwiatkowska, B. Kalinowski, M. Kośmider, and K. Rykaczewski, "Deep learning based open set acoustic scene classification," *Proc. Interspeech 2020.*

[14] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic scene classification in dcase 2019 challenge: Closed and open set classification and data mismatch setups," *Proc. of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE)*, 2019.

[15] 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," *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE)*, 2021.

[16] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, "Unsupervised Detection of Anomalous Sound Based on Deep Learning and the Neyman–Pearson Lemma," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[17] C. Geng, S.-J. Huang, and S. Chen, "Recent Advances in Open Set Recognition: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[18] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[19] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP).* IEEE, 2017, pp. 776–780.

[20] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).*

[21] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

[22] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally, *et al.*, "Hear: Holistic evaluation of audio representations," in *NeurIPS 2021 Competitions and Demonstrations Track*. PMLR, 2022, pp. 125–145.

[23] T. Denton, S. Wisdom, and J. R. Hershey, "Improving bird classification with unsupervised sound separation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE.

[24] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3846–3857, 2020.

[25] B. Dubuisson and M. Masson, "A statistical decision rule with incomplete knowledge about classes," *Pattern recognition*, vol. 26, no. 1, pp. 155–165, 1993.

[26] A. Bendale and T. E. Boult, "Towards Open Set Deep Networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Las Vegas, NV, USA: IEEE, June 2016, pp. 1563–1572.

[27] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[28] H. Phan, T. N. T. Nguyen, P. Koch, and A. Mertins, "Polyphonic audio event detection: multi-label or multi-class multitask classification problem?" in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, pp. 8877–8881.