

SEED: SOUND EVENT EARLY DETECTION VIA EVIDENTIAL UNCERTAINTY

Xujiang Zhao[†], Xuchao Zhang[‡], Wei Cheng[‡], Wenchao Yu[‡], Yuncong Chen[‡], Haifeng Chen[‡], Feng Chen[†]

[†]University of Texas at Dallas, {xujiang.zhao, feng.chen}@utdallas.edu

[‡]NEC Laboratories America, {xuczhang, weicheng, wyu, yuncong, haifeng}@nec-labs.com

ABSTRACT

Sound Event Early Detection (SEED) is an essential task in recognizing the acoustic environments and soundscapes. However, most of the existing methods focus on the offline sound event detection, which suffers from the over-confidence issue of early-stage event detection and usually yield unreliable results. To solve the problem, we propose a novel Polyphonic Evidential Neural Network (PENet) to model the evidential uncertainty of the class probability with Beta distribution. Specifically, we use a Beta distribution to model the distribution of class probabilities, and the evidential uncertainty enriches uncertainty representation with evidence information, which plays a central role in reliable prediction. To further improve the event detection performance, we design the backtrack inference method that utilizes both the forward and backward audio features of an ongoing event. Experiments on the DESED database show that the proposed method can simultaneously improve 13.0% and 3.8% in time delay and detection F1 score compared to the state-of-the-art methods.

Index Terms— Sound event detection, early detection, uncertainty, evidence

1. INTRODUCTION

Sound event detection (SED), as a fundamental task to recognize the acoustic events, has achieved significant progress in a variety of applications, such as unobtrusive monitoring in health care, and surveillance. Recently, Deep Neural Network (DNN) based methods such as CRNN [1] and Conformer [2] significantly improve the event detection performance. However, these methods are usually designed in an offline setting that the entire audio clip containing sound events is fully observed. This assumption may not hold in many real-world applications that require real-time event detection. For example, the event detection in audio surveillance [3] requires low latency reaction to potentially dangerous circumstances for life saving and protection. In this paper, we will focus on the sound event early detection (SEED) problem, which is designed in an online setting that requires ongoing events to be recognized as early as possible.

Despite the importance of the SEED problem, few existing focus on detecting sound events with short delays from

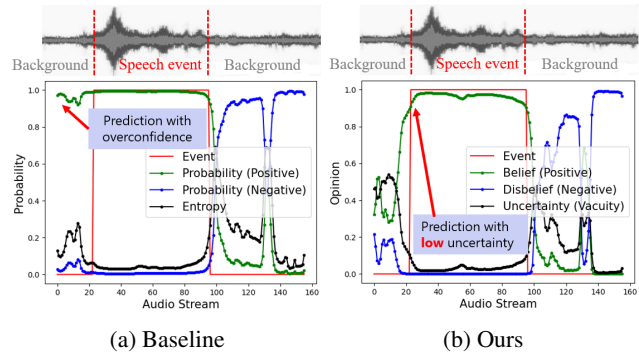


Fig. 1: An example of sound event early detection. (a) Baseline (CRNN) model might give a prediction with over-confidence, and result in a false positive detection; (b) To avoid over-confidence, our framework estimates vacuity uncertainty instead of entropy, and make a reliable prediction with low vacuity.

audio streaming data. Some works design a monotonous detection function to achieve early detection, such as random regression forests algorithm[4], Dual-DNN [5]. Some work [6] proposes a detection front-end to identify seed regions from spectrogram features to detect events at the early stage. However, the prediction of these methods are based on probability, which could be not reliable (over-confidence) [7, 8]. Especially during the early stage of an ongoing event, we only collect a small number of stream audios that may not be enough to compose a clear event sound to support a reliable prediction. Figure 1 (a) shows an example that prediction based on probability is over-confidence at the early stage.

To solve the issue discussed above, we propose a novel Polyphonic Evidential Neural Network (PENet) to estimate the Beta distribution instead of the class probability such that we can estimate evidential uncertainty for each prediction. The attached evidential uncertainty is able to detect the “over-confidence” prediction and achieve a reliable prediction. To further improve the SEED performance, we propose the backtrack inference method that consider the forward information (waiting for the future information) of an ongoing event.

2. METHODOLOGY

In this section, we begin with the essential concepts of evidential uncertainty. Then, we introduce the proposed Poly-

phonic Evidential Neural Network with its backtrack inference method.

2.1. Subjective Logic and Evidential Uncertainty

Subjective Logic (SL) [9] defines a subjective opinion by explicitly considering the dimension of uncertainty derived from vacuity (i.e., a lack of evidence). For a given binomial opinion towards proposition (e.g., an audio segment) \mathbf{x} , an opinion is expressed by two belief masses (i.e., belief b and disbelief d) and one uncertainty mass (i.e., vacuity, u). Denote an opinion by $\omega = (b, d, u)$, where b and d can be thought as positive (event happen) vs. negative (event not happen) on a given audio segment. We have the property $b + d + u = 1$ and $b, d, u \in [0, 1]$.

An opinion, ω , can always be projected onto a single probability distribution by removing the uncertainty mass. To this end, the expected belief probability p is defined by: $p = b + a \cdot u$, where a refers to a base rate representing a prior knowledge without commitment such as neither agree nor disagree. A binomial opinion follows a Beta pdf (probability density function), denoted by $\text{Beta}(p|\alpha, \beta)$ in Eq (1), where α and β represents the positive and negative evidence.

$$\text{Beta}(p|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1}, \quad (1)$$

where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ and $\Gamma(\cdot)$ is the gamma function. In SL, α and β are received over time. An opinion w can be obtained based on α and β as $w = (\alpha, \beta)$. This can be translated to $w = (b, d, u)$ using the mapping rule in SL:

$$b = \frac{\alpha - 1}{\alpha + \beta}, \quad d = \frac{\beta - 1}{\alpha + \beta}, \quad u = \frac{W}{\alpha + \beta}, \quad (2)$$

where W is an amount of uncertainty evidence. In practice we set $W = 2$ for binary case.

Evidential Uncertainty. The concept of evidential uncertainty has been discussed differently depending on domains [10, 7, 11, 12, 13]. In this work, we adopt the concept of uncertainty based on SL in developing an uncertainty-based SEED framework when the input is a streaming audio signal. *Vacuity* refers to a lack of evidence, meaning that uncertainty is introduced because of no or insufficient information. High vacuity might happen at the early stage of an ongoing event due to the small amount of collected stream audios, resulting in an over-confidence estimation. Table 1 illustrates the difference between probability and evidence. For example, at the early stage of an ongoing event, we only collect 1 positive evidence and 4 negative evidence. And we can calculate its expected probability $p = [0.2, 0.8]$, which result in high confidence of negative prediction. However, prediction based on small amount of evidence (i.e., high vacuity) is not reliable. With collecting more evidence (e.g., $[\alpha, \beta] = [200, 4]$), we would have a reliable prediction with low vacuity.

Table 1: Difference between evidence and probability. Prediction with less evidence (high vacuity) is not reliable.

Evidence	$[\alpha, \beta] = [1, 4]$	$[\alpha, \beta] = [4, 4]$	$[\alpha, \beta] = [200, 4]$
Probability	$p = [0.2, 0.8]$	$p = [0.5, 0.5]$	$p = [0.98, 0.02]$
Uncertainty	High Vacuity	High Vacuity	Low

2.2. Polyphonic Evidential Neural Network

Based on the intuition of evidential uncertainty in SEED, we propose a novel Polyphonic Evidential Neural Network (PENet) with reliable prediction. The overall description of the framework is shown in Figure 2. For SEED setting, the audio signal is collected in a streaming way. At each timestamp t , we collect an audio segment \mathbf{x}^t . The corresponding label of \mathbf{x}^t is denoted as $y^t = [y_1^t, \dots, y_K^t]$, where $y_k^t = \{0, 1\}$.

PENet. For polyphonic sound event detection, most existing methods would consider a binary classification for each class, such as softmax output [14, 15]. As discussed in Section 2.1, evidential uncertainty can be derived from binomial opinions or equivalently Beta distributions to model an event distribution for each class. Therefore, we design a Polyphonic Evidential Neural Network (PENet) f to form their binomial opinions for the class-level Beta distribution of a given audio segment \mathbf{x}^t . In addition, we considered a context of m segments for sequential input purpose. Then, the conditional probability $P(p_k^t | \mathbf{x}^{[t-m, t]}; \theta)$ of class k can be obtained by:

$$P(p_k^t | \mathbf{x}^{[t-m, t]}; \theta) = \text{Beta}(p_k^t | \alpha_k^t, \beta_k^t) \quad (3)$$

$$\alpha_k^t, \beta_k^t = f_k(\mathbf{x}^{[t-m, t]}; \theta) \quad (4)$$

where $\mathbf{x}^{[t-m, t]}$ means a sequence of audio segments, i.e., $[\mathbf{x}^{t-m}, \mathbf{x}^{t-m+1}, \dots, \mathbf{x}^t]$, f_k is the output of PENet for class k , and θ is the model parameters. The Beta probability function $\text{Beta}(p_k^t | \alpha_k^t, \beta_k^t)$ is defined by Eq. (1). Note that PENet is similar to the classical polyphonic sound event detection model (e.g., CRNN [14]), except that we use an activation layer (e.g., ReLU) instead of the softmax layer (only outputs class probabilities). This ensures that PENet would output non-negative values taken as the evidence for the predicted Beta distribution.

Training with Beta Loss. In this paper, we design and train neural networks to form their binomial opinions for the classification of a given audio segment as a Beta distribution. For the binary cross-entropy loss, we have the Beta loss by computing its Bayes risk for the class predictor,

$$\begin{aligned} \mathcal{L}_{Beta} &= \sum_{t=1}^T \sum_{k=1}^K \int [\text{BCE}(y_k^t, p_k^t)] \text{Beta}(p_k^t; \alpha_k^t, \beta_k^t) dp_k^t \\ &= \sum_{t=1}^T \sum_{k=1}^K \left[y_k^t \left(\psi(\alpha_k^t + \beta_k^t) - \psi(\alpha_k^t) \right) + \right. \\ &\quad \left. (1 - y_k^t) \left(\psi(\alpha_k^t + \beta_k^t) - \psi(\beta_k^t) \right) \right], \end{aligned} \quad (5)$$

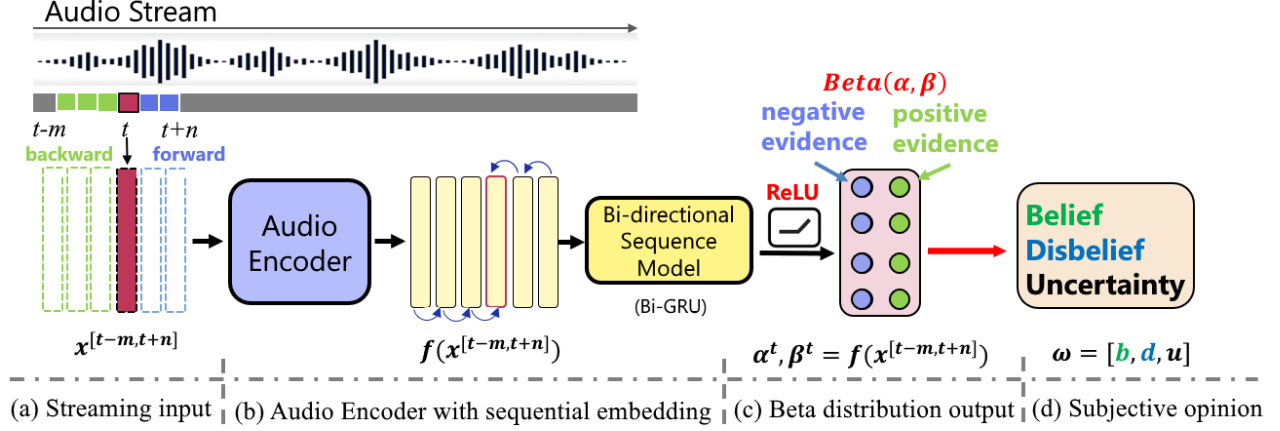


Fig. 2: PENet Overview. Given the streaming audio segments (a), PENet is designed for estimating the Beta distribution (c), which can be transferred to subjective opinion (d) with vacuity uncertainty.

where T is the number of segments decomposed from an audio, K is the number of class, $\text{BCE}(y_k^t, p_k^t) = -y_k^t \ln(p_k^t) - (1 - y_k^t) \ln(1 - p_k^t)$ is the binary cross-entropy loss, and $\psi(\cdot)$ is the *digamma* function. The log expectation of Beta distribution derives the second equality.

Uncertainty-based Inference. At the test stage, we consider a simple strategy to make a reliable prediction. For each class, we predict sound events happened only when belief larger than disbelief with a small vacuity,

$$\hat{y}_k^t = \begin{cases} 1, & \text{if } b_k^t > d_k^t \text{ and } u_k^t < V \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where $\hat{y}_k^t \in \{0, 1\}$ is the model prediction for class k in segment t , V is the vacuity threshold.

2.3. Backtrack Inference

We propose a backtrack inference method that considers forward and backward information to feed into PENet as a sequential input to further improve early detection performance. Figure 2 (a) illustrate the backtrack input. Then, we can rewrite Eq. (4) as

$$\alpha_k^t, \beta_k^t = f_k(\mathbf{x}^{[t-m, t+n]}; \theta), \quad (7)$$

where m is the backward steps, and n is the forward steps. The additional previous and future information is critical for the prediction of the current audio segment. We show that backtrack inference improves SEED detection accuracy, but the waiting process (consider forward information) causes a higher detection delay.

3. EXPERIMENTS

3.1. Experiment Details

Dataset. We conduct the experiments on DESED2021 dataset [14] The dataset for this task is composed of 10 sec

audio clips recorded in domestic environments or synthesized using Scaper [16] to simulate a domestic environment. The dataset includes 10 classes of sound events that represent a subset of Audioset [15]. In DESED2021 dataset, the training set contains 10,000 synthetic audio clips with strong-label, 1578 weak-label audio clips, and 14,412 unlabeled audio clips. The validation set includes 1168 audio clips that are annotated with strong-label (timestamps obtained by human annotators). The test set includes 1,016 real-world audio clips.

Features. The input features used in the experiments are log-mel spectrograms extracted from the audio signal resampled to 16000 Hz. The log-mel spectrogram uses 2048 STFT windows with a hop size of 256 and 128 Mel-scale filters. At the training stage, the input is the full observed 10-second sound clip. As a result, each 10-second sound clip is transformed into a 2D time-frequency representation with a size of (626×128). At the test stage, we collect an audio segment at each timestamp, which can be transformed into a 2D time-frequency representation with a size of (4×128).

Comparing Methods. To evaluate the effectiveness of our proposed approach (PENet), we compare it with one state-of-the-art SEED method: Dual DNN [5]; two SED methods: CRNN [14] and Conformer [2]; and three different uncertainty methods: *Entropy*, *Epistemic* uncertainty [17] (represents the uncertainty of model parameters), and *Aleatoric* uncertainty [18] (represents the uncertainty of data noise). We use MC-drop [19] to estimate epistemic and aleatoric uncertainties in the experiments.

Evaluation Metrics. Since the traditional offline sound event detection metrics cannot early detection performance, we use both early detection F1 and detection delay to evaluate our performance for the onset of sound events at the early stage. We first define the true positive prediction for the event k only happened when the first prediction timestamp d_p is located into event happened intervals. In addition, we set an early

predict tolerance L that if the first prediction is earlier than true event occurred. Otherwise, we consider the prediction for this event is false positive,

$$TP_k = \begin{cases} 1, & \text{if } y_k^{d_p} == 1 \text{ and } d_p - d_t \geq L \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where d_t is the onset timestamp of the predicted event. For detection delay, it's only measured when we have a true positive prediction. The detection delay is defined as follows,

$$\text{delay} = \begin{cases} d_p - d_t, & \text{if } d_p \geq d_t \\ 0, & \text{if } d_p < d_t \end{cases} \quad (9)$$

Set up. For all experiments, we use CRNN [20] as the backbone except Conformer. We use the Adam optimizer for all methods and follow the same training setting as [20]. For the uncertainty threshold, we set 0.5 for epistemic uncertainty and 0.9 for other uncertainties (entropy, vacuity, aleatoric).

3.2. Results and Analysis

Early Detection Performance. Table 2 shows that our Evidence model with vacuity outperforms all baseline models under the detection delay and early detection F1 score for sound event early detection. The outperformance of vacuity-based detection is fairly impressive. This confirms that low vacuity (a large amount of evidence) is the key to maximize the performance of early detection. In addition, we observe that backtrack technical can significantly improve the early detection F1, demonstrating that backtrack information is essential in SEED. However, using the backtrack technical would increase the detection delay as well. Furthermore, the test inference time of our approach is around 5ms, less than the streaming segment duration (60ms), which indicates that our method satisfies the real-time requirement.

Table 2: Early detection performance on DESED dataset.

Model	Delay ↓	F1 ↑	Time
Conformer	0.372	0.639	6.6ms
Dual DNN	0.386	0.682	5.1ms
CRNN	0.284	0.687	5.0ms
CRNN + entropy	0.312	0.679	5.0ms
CRNN + epistemic	0.278	0.647	27ms
CRNN + aleatoric	0.281	0.643	27ms
PENet	0.247	0.670	5ms
PENet + vacuity	0.252	0.691	5ms
PENet + vacuity + backtrack	0.310	0.725	5.2ms

Uncertainty Analysis. We explore the sensitivity of vacuity threshold used in the evidence model. Figure 3 plots the detection delay and early detection F1 score with the varying vacuity threshold values. When the vacuity threshold increases, the detection delay decreases continuously, and the

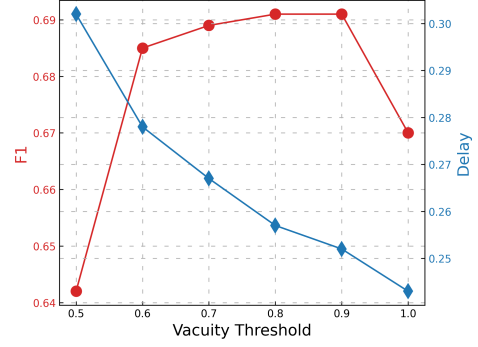


Fig. 3: Effect of vacuity threshold.

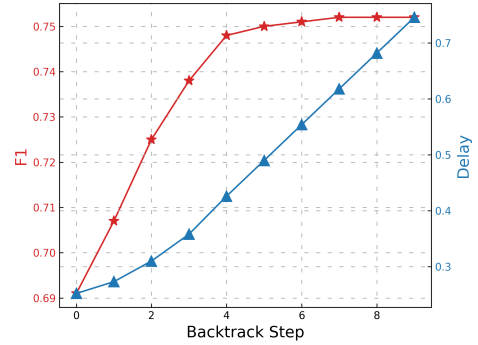


Fig. 4: Detection delay and F1 score for different numbers of backtrack steps.

early detection F1 score reaches the highest when vacuity is 0.9.

Trade off of Backtrack. We analyzed the sensitivity of our proposed backtrack method to the number of backtrack steps. Table 4 shows a trade-off between detection delay and F1 score with the varying numbers of steps. When the backtrack step increase, the detection delay is continuously increased, and detection F1 increases until backtrack step equal to 6. The results demonstrate that backtrack information is critical to improving the detection accuracy in SEED.

4. CONCLUSION

In this paper, we propose a novel Polyphonic Evidential Neural Network to model the evidential uncertainty of the class probability with Beta distribution. Specifically, we use a Beta distribution to model the distribution of class probabilities, and the evidential uncertainty enriches uncertainty representation with evidence information, which plays a central role in reliable prediction. And the proposed backtrack inference method can further improve the event detection performance. The experiment results demonstrate that the proposed method outperformed in the SEED task compared with other competitive counterparts.

5. REFERENCES

- [1] Annamaria Mesaros, Aleksandr Diment, Benjamin Elizalde, Toni Heittola, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen, “Sound event detection in the dcase 2017 challenge,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 992–1006, 2019.
- [2] Koichi Miyazaki, Tatsuya Komatsu, Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, and Kazuya Takeda, “Conformer-based sound event detection with semi-supervised learning and data augmentation,” *dim*, vol. 1, pp. 4, 2020.
- [3] Quoc Nguyen Viet, HoSeok Kang, Sun-Tae Chung, Seongwon Cho, Keeseong Lee, and Taein Seol, “Real-time audio surveillance system for ptz camera,” in *2013 International Conference on Advanced Technologies for Communications (ATC 2013)*. IEEE, 2013, pp. 392–397.
- [4] Huy Phan, Marco Maass, Radoslaw Mazur, and Alfred Mertins, “Early event detection in audio streams,” in *2015 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2015, pp. 1–6.
- [5] Huy Phan, Philipp Koch, Ian McLoughlin, and Alfred Mertins, “Enabling early audio event detection with neural networks,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 141–145.
- [6] Ian Vince McLoughlin, Yan Song, Lam Dang Pham, Ramaswamy Palaniappan, Huy Phan, and Yue Lang, “Early detection of continuous and partial audio events using cnn,” in *Proceedings of Interspeech*. International Speech Communication Association, 2018, vol. 2018, pp. 3314–3318.
- [7] Xujiang Zhao, Feng Chen, Shu Hu, and Jin-Hee Cho, “Uncertainty aware semi-supervised learning on graph data,” *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [8] Murat Sensoy, Lance Kaplan, and Melih Kandemir, “Evidential deep learning to quantify classification uncertainty,” *arXiv preprint arXiv:1806.01768*, 2018.
- [9] Audun Jøsang, *Subjective Logic: A Formalism for Reasoning Under Uncertainty*, Springer, 2016.
- [10] Audun Josang, Jin-Hee Cho, and Feng Chen, “Uncertainty characteristics of subjective opinions,” in *FUSION*. IEEE, 2018, pp. 1998–2005.
- [11] Liyan Xu, Xuchao Zhang, Xujiang Zhao, Haifeng Chen, Feng Chen, and Jinho D. Choi, “Boosting cross-lingual transfer via self-learning with uncertainty estimation,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Nov. 2021, pp. 6716–6723.
- [12] Weishi Shi, Xujiang Zhao, Feng Chen, and Qi Yu, “Multifaceted uncertainty estimation for label-efficient deep learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 17247–17257, 2020.
- [13] Yibo Hu, Yuzhe Ou, Xujiang Zhao, Jin-Hee Cho, and Feng Chen, “Multidimensional uncertainty-aware evidential neural networks,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, pp. 7815–7822, May 2021.
- [14] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon, “Sound event detection in domestic environments with weakly labeled data and soundscape synthesis,” in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019.
- [15] Shawn Hershey, Daniel PW Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal, “The benefit of temporally-strong labels in audio event classification,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 366–370.
- [16] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello, “Scaper: A library for soundscape synthesis and augmentation,” in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 344–348.
- [17] Yarin Gal and Zoubin Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.
- [18] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft, “Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 1184–1193.
- [19] Yarin Gal and Zoubin Ghahramani, “Bayesian convolutional neural networks with bernoulli approximate variational inference,” *arXiv preprint arXiv:1506.02158*, 2015.
- [20] Nicolas Turpault and Romain Serizel, “Training sound event detection on a heterogeneous dataset,” *arXiv preprint arXiv:2007.03931*, 2020.