ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096305

MULTI-LABEL TEMPORAL EVIDENTIAL NEURAL NETWORKS FOR EARLY EVENT DETECTION

Xujiang Zhao[1], *Xuchao Zhang*[2], *Chen Zhao*[3], *Jin-Hee Cho*[4], *Lance Kaplan*[5], *Dong Hyun Jeong*[6], *Audun Jøsang*[7], *Haifeng Chen*[1], *Feng Chen*[3]

[1]NEC Lab, [2]Microsoft, [3]UT-Dallas, [4]Virginia Tech, [5]US Army Research Lab, [6]University of the District of Columbia, [7]University of Oslo

ABSTRACT

Early event detection aims to detect events even before the event is complete. However, most of the existing methods focus on an event with a single label but fail to be applied to cases with multiple labels. Another non-negligible issue for early event detection is a prediction with overconfidence due to the high vacuity uncertainty that exists in the early time series. It results in an over-confidence estimation and hence unreliable predictions. To this end, technically, we propose a novel framework, Multi-Label Temporal Evidential Neural Network (MTENN), for multi-label uncertainty estimation in temporal data. MTENN is able to quality predictive uncertainty due to the lack of evidence for multi-label classifications at each time stamp based on belief/evidence theory. In addition, we introduce a novel uncertainty estimation head (weighted binomial comultiplication (WBC)) to quantify the fused uncertainty of a sub-sequence for early event detection. We validate the performance of our approach with state-of-the-art techniques on real-world audio datasets.

Index Terms— early event detection, uncertainty

1. INTRODUCTION

In recent decades, early detection of temporal events has aroused a lot of attention and has applications in a variety of industries, including security [1, 2], quality monitoring [3], medical diagnostic [4], transportation [5]. According to the time series, an event can be viewed with three components, pre-event, ongoing event, and post-event. Early event detection in machine learning identifies an event during its initial ongoing phase after it has begun but before it concludes [6, 7].

To achieve the earliness of event detection, existing approaches can be broadly divided into several major categories. Prefix-based techniques [8, 5] aim to learn a minimum prefix length of the time series from the training instances and utilize it to classify a testing time series. Shapelet-based approaches [9, 4] focus on obtaining a set of key shapelets from the training dataset and utilizing them as class discriminatory features. Dual-DNN [7] is proposed for the sound event early detection via a monotonous function design. [10] identifies seed regions from spectrogram features to detect events at the early stage. Other algorithms have considered epistemic uncertainty for reliable event prediction [11]. Although these approaches address the importance of early detection, they primarily focus on an event with a single label but fail to be applied to situations with multiple labels.

Another non-negligible issue for early event detection is a prediction with overconfidence [12, 13, 14]. In general, the occurrence of an event is determined by its predicted probability. An event with a high probability is considered as an occurrence. This, however, may not be reliable. Figure 1 shows an example that the prediction of the occurrence of an event in an audio clip with a binary class (occurs or not) based on its predicted probability is overconfident at the pre-event stage. In this case, the ground truth (red line) demonstrates that the ongoing stage starts at the 20th frame. Nevertheless, the event is falsely detected, prior to it actually occurring (Figure 1, left), because a greater probability (*i.e.* 0.9 indicated on the green line) is given by positive evidence. Here, the evidence indicates data samples (*i.e.* actions) that are closest to the predicted one in the feature space and used to support the decision-making. Positive (negative) evidence is the observed samples that have the same (opposite) class labels. The event prediction with overconfidence at its early stage is due to high vacuity uncertainty [15] which is a terminology representing a lack of evidence. Therefore, it makes event detection based on probability unreliable. To overcome this flaw, methods developed on uncertainty estimation using evidence are desirable for early event detection.

To address the aforementioned issues, we proposed a novel framework, Multi-Label Temporal Evidential Neural Network (MTENN), which is composed of two phases: in phase one, a time series data is viewed as a sequence of segments with equal temporal length, where each segment comes one after another. Instead of predicting occurrence probabilities for all events, their positive and negative evidence are estimated through the proposed MTENN. In the second phase, a sliding window spanning the most recent collected segments is designed to validate whether an event is successfully detected through a novel uncertainty estimation head: Weighted Binomial Comultiplication.

2. PRELIMINARIES

In this section, we introduce some essential concepts in subjective logic and evidential uncertainty.

Subjective Logic (SL) was defined [15] by explicitly consid-

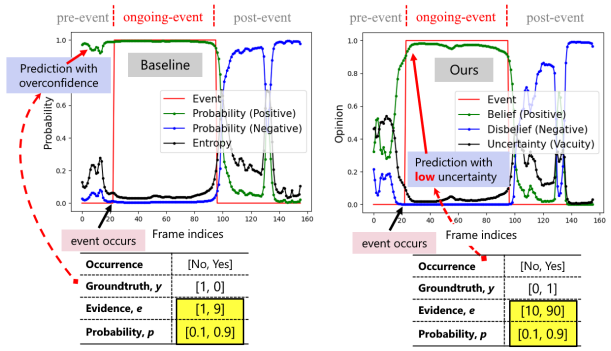


Fig. 1: Illustration of overconfidence issue. (Left) The occurrence of the event is falsely detected at the pre-event stage prior to its starting. This indicates that predicted probabilities are not reliable due to insufficient evidence. (Right) Instead of probabilities, subjective opinions (*e.g.,* belief, disbelief, uncertainty) are used in the proposed method for early event detection.

ering the dimension of uncertainty derived from vacuity (*i.e.,* a lack of evidence). Given a 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 $\omega = (b, d, u, a)$, where b and d can be thought as positive (an event occurs) vs. negative (does not occurs) on a given segment, and a refers to a base rate representing a prior knowledge. We have the property $b + d + u = 1$ where $b, d, u, a \in [0, 1]$. To this end, the expected belief probability p is defined by: $p = b + a \cdot u$. A binomial opinion can be translated into a Beta distribution, denoted by $\text{Beta}(p|\alpha, \beta)$, where α and β represent the positive and negative evidence.

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

where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$, and $\Gamma(\cdot)$ is the gamma function. Eq. (2) shows the mapping rule between Beta distribution and opinion.

$$b = \frac{\alpha - a \cdot W}{\alpha + \beta}, \; d = \frac{\beta - (1-a) \cdot W}{\alpha + \beta}, \; u = \frac{W}{\alpha + \beta}, \tag{2}$$

where W is an amount of uncertainty evidence. In practice, we set $W = 2$ for a binary case. In addition, subjective logic can extend to the multi-class scenario with a multinomial opinion, which can be translated into a Dirichlet distribution.

Evidential Uncertainty. *Vacuity* refers to the lack of adequate evidence. In other words, this type of uncertainty is due to a lack of or insufficient information of evidence. High vacuity may happen at the early stage of an ongoing event due to the small number of collected stream signals, resulting in an over-confidence estimation. Figure 1 illustrates the implicit relations and differences between probability and evidence. For example, at the pre-event stage, we only collect 1 negative evidence and 4 positive evidence. And we can calculate its expected probability $p = [0.2, 0.8]$, which results in an over-confidence prediction. However, prediction based on a small amount of evidence (*i.e.,* high vacuity) is not reliable. As more evidence is collected (*e.g.,* $[\alpha, \beta] = [4, 200]$), we have a reliable prediction with low uncertainty.

3. METHODOLOGY

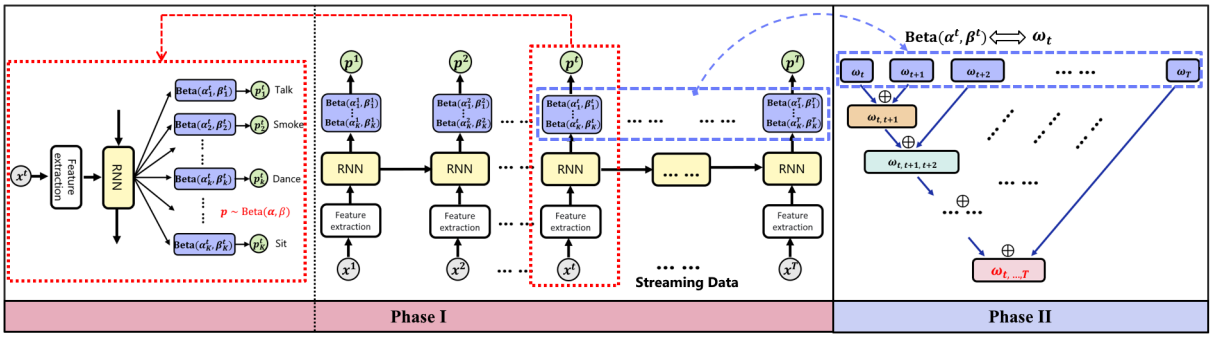
Problem Formulation. Given a time series data with multiple labels where each class label is viewed as an event, let $\mathcal{X} \times \mathcal{Y}$ be the data space, where $\mathcal{X}$ is an input space and $\mathcal{Y} = \{0, 1\}^K$ is an output space. A time series data $\{(\mathbf{x}^t, \mathbf{y}^t)\}_{t=1}^T \in (\mathcal{X} \times \mathcal{Y})$ consists of T segments where each $(\mathbf{x}^t, \mathbf{y}^t)$ is collected one after another over time. $\mathbf{x}^t$ represents the feature vector. $\mathbf{y}^t = [y_1^t, \ldots, y_K^t]^T$ denotes the multi-label formula with $y_k^t = \{0, 1\}, \forall k \in \{1, \cdots K\}$ representing an event occurs or not and K is the number of classes. A segment buffer $\mathcal{B}$ is initialized as empty. It is maintained by adding each segment one at a time. That is, at timestamp t, the buffer includes all segments from previous $\mathcal{B} = \{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^t$ and $|\mathcal{B}| = t$. At each time, a predictive model $f : \mathcal{X} \to \mathcal{Y}$ parameterized by $\boldsymbol{\theta}$ takes segments in $\mathcal{B}$ as the input and outputs a event prediction vector $\hat{\mathbf{y}}^t = [\hat{y}_1^t, \ldots, \hat{y}_K^t]^T$ where $\hat{y}_k^t \in \{0, 1\}$ represents the predicted result of the k-th event at time t . Therefore, for some events which are predicted as occurrences, one may conclude that they can be detected at time t.

MTENN Framework. For multi-label early event detection, most existing methods would consider a binary classification for each class, such as sigmoid output [16, 17]. As discussed in Section 2, evidential uncertainty can be derived from binomial opinions or equivalently Beta distributions to model an event distribution for each class. Therefore, we proposed a novel a Multi-label Evidential Temporal Neural Network (MTENN) $f(\cdot)$ to form their binomial opinions for the class-level Beta distribution of a given time series segments $[\mathbf{x}^1, \ldots, \mathbf{x}^t]$. Then, the conditional probability $P(p_k^t|\mathbf{x}^1, \ldots, \mathbf{x}^t; \boldsymbol{\theta})$ of class k at timestamp t can be obtained by:

$$\begin{aligned} [f^1, \ldots, f^t] &\leftarrow f(\mathbf{x}^1, \ldots, \mathbf{x}^t; \boldsymbol{\theta}) \\ (\alpha_1^t, \beta_1^t), \ldots, (\alpha_K^t, \beta_K^t), &\leftarrow f^t(\mathbf{x}^1, \ldots, \mathbf{x}^t; \boldsymbol{\theta}), \\ p_k^t &\sim \textbf{Beta}(p_k^t|\alpha_k^t, \beta_k^t), \\ y_k^t &\sim \textbf{Bernoulli}(p_k^t), \end{aligned}$$

where $k \in \{1, \cdots, K\}$, $t \in \{1, \cdots, T\}$, f^t is the output of MTENN at timestamp t, and $\boldsymbol{\theta}$ refers to model parameters. The Beta probability function $\textbf{Beta}(p^t|\alpha_k^t, \beta_k^t)$ is defined by Eq. (1). Therefore, MTENN is able to quality predictive uncertainty (vacuity) due to a lack of evidence for multi-label classifications at each time stamp based on belief/evidence theory, and vacuity can be calculated based on Eq. (2) from the estimated Beta distribution.

In this paper, we design and train MTENN to form their binomial opinions for the classification of a given streaming segment as a Beta distribution. For the binary cross-entropy loss, we have the MTENN loss by computing its Bayes risk



(a) Multi-Label Evidence (b) Multi-Label Temporal Evidential Neural Network (c) Sequential Uncertainty Estimation

Fig. 2: **Framework Overview.** Given the streaming data, (b) MTENN is able to quality predictive uncertainty for multi-label classifications at each time stamp. Specifically, (a) at each time step with data segment x^t, MTENN is able to predict Beta distribution for each class, which can be equivalent transfer to subjective opinion ω_t; (c) based on a sliding window, a novel fusion operator is introduced to quantify the fused uncertainty of a sub-sequence for early event detection.

for the class predictor,

$$\mathcal{L}_{MTENN} = \sum_{t=1}^{T}\sum_{k=1}^{K}\int \Big[\mathbf{BCE}(y_k^t, p_k^t)\Big]\mathbf{Beta}(p_k^t;\alpha_k^t,\beta_k^t)dp_k^t$$
$$= \sum_{t=1}^{T}\sum_{k=1}^{K}\Big[y_k^t\Big(\psi(\alpha_k^t+\beta_k^t)-\psi(\alpha_k^t)\Big)+(1-y_k^t)\Big(\psi(\alpha_k^t+\beta_k^t)-\psi(\beta_k^t)\Big)\Big],$$

where $\mathbf{BCE}(y_k^t, p_k^t) = -y_k^t\log(p_k^t) - (1-y_k^t)\log(1-p_k^t)$ is the binary cross-entropy loss, and $\psi(\cdot)$ is the *digamma* function. The log expectation of Beta distribution derives the second equality.

Multi-label Sequential Uncertainty Quantitation. In the second phase, for early event detection, at time t, a subset including m most recent collected segments are considered to validate whether an event is successfully detected or not, as shown in Fig 2 (c). We name the subset as a sliding window, as it dynamically restructures a small sequence of segments from $t-m$ to t and performs validation through an early detection function at each time. Based on the sliding window, we introduce a novel uncertainty fusion operator based on MTENN to quantify the fused uncertainty of a sub-sequence for early event detection.

Weighted Binomial Comultiplication. After we get the sequential Beta distribution output, a sequential fusional opinion can be estimated via a subjective operator (e.g., union operator). As shown in Fig 2 (b), we can use subjective operator $\oplus$ to fuse the opinions. Here we consider to use comultiplication operator [18] to fusion two opinion ω_i and ω_j via Eq. (3),

$$\begin{aligned} b_{i\oplus j} &= b_i + b_j - b_i b_j \\ d_{i\oplus j} &= d_i d_j + \frac{a_i(1-a_j)d_i u_j + (1-a_i)a_j u_i d_j}{a_i + a_j - a_i a_j} \\ u_{i\oplus j} &= u_i u_j + \frac{a_j d_i u_j + a_i u_i d_j}{a_i + a_j - a_i a_j} \\ a_{i\oplus j} &= a_i + a_j - a_i a_j \end{aligned} \tag{3}$$

Based m sliding windows, the sequential fusional opinion can be calculated by

$$\hat{\omega}^t = c_{t-m}\cdot\omega^{t-m} \oplus c_{t-m+1}\cdot\omega^{t-m+1} \oplus \ldots \oplus c_t\cdot\omega^t \tag{4}$$

where $\mathbf{c}$ is the weight for each opinion when executing the operator, which is designed for the order information and emphasizes the importance of current time step t. We consider the vacuity from $\hat{\omega}^t$ as sequential uncertainty for a sub-sequence for early event detection (e.g., filter the over-confidence prediction by large vacuity).

4. EXPERIMENTS

Dataset. We conduct the experiments on DESED2021 dataset [16] and AudioSet-Strong-Labeled dataset [17]. DESED2021 dataset is composed of 10-sec audio clips recorded in domestic environments or synthesized using Scaper to simulate a domestic environment. The original AudioSet-Strong-Labeled dataset consists of an expanding ontology of 632 audio event classes and a collection of 2,084,320 human-labeled 10-second sound clips drawn from YouTube videos. To simulate the application of early event detection in industries, we select some subsets from AudioSet-Strong-Labeled dataset to form early event detection datasets. Specifically, we select four subsets from AudioSet-Strong-Labeled dataset, including explosion, alarm, liquid, and engine subclasses. The details of each dataset are shown in Table 1.

Comparing Methods. To evaluate the effectiveness of our proposed approach, we compare it with two state-of-the-art early sound event detection methods: Dual DNN [7] and SEED [19]; two sound event detection methods: CRNN [16] and Conformer [20]; In addition, we consider three different uncertainty methods as the baselines, which include *Entropy*, *Epistemic* uncertainty [21], and *Aleatoric* uncertainty [22] . We consider using uncertainty to filter the high uncertainty prediction for three uncertainty-based methods. We use MC-drop [21] to estimate epistemic and aleatoric uncertainties in the experiments.

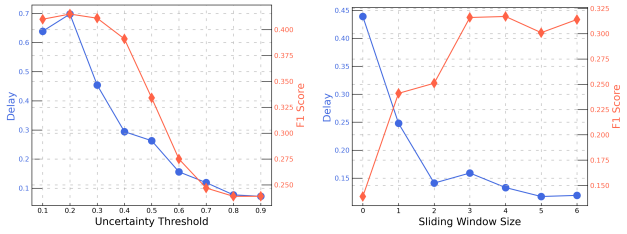
Evaluation Metrics. We consider both early detection F1 score and detection delay as the evaluation metrics. We first define the true positive prediction for the event k, which only happens when the first prediction timestamp d_p is located in the ongoing event region. In contrast, the false positive prediction happened when the first prediction timestamp d_p is not located in the ongoing event region. Then we can calculate

Table 1: Early sound event detection performance on Audio datasets.

Datasets	**DESED2021**	**AudioSet(Explosion)**	**AudioSet(Alarm)**	**AudioSet(Liquid)**	**AudioSet(Engine)**
# Classes	10	5	7	9	8
# Training samples	10,000	5,518	8,085	6,517	18,741
# Validation samples	1,168	788	1,355	931	2,677
# Test samples	1,016	1,577	2,311	1,862	5,354
	Detection Delay ↓ / Detection F1 Score ↑				
Dual DNN	0.386 / 0.682	0.325 / 0.295	0.257 / 0.221	0.467 / 0.162	0.324 / 0.323
SEED	0.252 / 0.691	0.339 / 0.288	0.293 / 0.407	0.334 / 0.172	0.428 / 0.342
Conformer	0.372 / 0.639	0.444 / 0.268	0.292 / 0.429	0.463 / 0.166	0.427 / 0.323
CRNN	0.284 / 0.677	0.415 / 0.278	0.273 / 0.408	0.451 / 0.144	0.404 / 0.301
CRNN + entropy	0.312 / 0.669	0.422 / 0.272	0.282 / 0.406	0.465 / 0.142	0.423 / 0.313
CRNN + epistemic	0.278 / 0.647	0.401 / 0.28	0.244 / 0.413	0.411 / 0.152	0.356 / 0.31
CRNN + aleatoric	0.281 / 0.643	0.404 / 0.288	0.252 / 0.419	0.421 / 0.157	0.377 / 0.312
MTENN	**0.206 / 0.727**	**0.119 / 0.314**	**0.217 / 0.470**	**0.059 / 0.200**	**0.294 / 0.391**

precision, recall, and F1 score based on true positive prediction and false positive prediction for each event. For detection delay, it's only measured when we have a true positive prediction. Then the detection delay is defined as delay $= d_p - d_t$ if $d_p \geq d_t$, otherwise, delay $= 0$, where d_t is the onset timestamp of the predicted event.

Settings. We use CRNN [23] as the backbone except for Conformer. We use the Adam optimizer for all methods and follow the same training setting as [23].



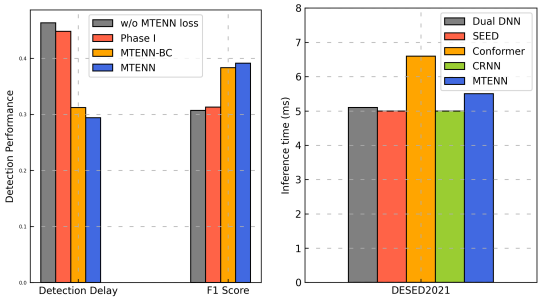
(a) AudioSet (Engine) (b) AudioSet (Explosion)

Fig. 3: Sensitive analysis.

Early Event Detection Performance. Table 1 shows that our proposed methods outperform all baseline models under the detection delay and early detection F1 score for the sound event early detection. The outperformance of MTENN is fairly impressive. This confirms that the belief comultiplication operator is the key to improving the sequential uncertainty estimation.

Sensitive Analysis. (1) Uncertainty threshold. Figure 3 (a) shows the detection delay and early detection F1 score with varying vacuity threshold values. There is a tradeoff between detection delay and detection accuracy. The higher the uncertainty threshold increase, the more overconfident predictions (predictions with high uncertainty) result in an aggressive early prediction (may predict event happen early but may cause a false positive prediction). (2) Effect of sliding window size. Fig 3 (b) shows the performance with the varying size of sliding windows. When the sliding window size increases, the detection delay continuously decreases, and detection F1 increases until the sliding window size is large enough.

Ablation study. We conducted additional experiments (see Fig 4) to demonstrate the contributions of the key technical components, including MTENN loss and WBC. Specifically, we consider three ablated models: (a) MTENN-BC, a variant of MTENN that uses binomial comultiplication without weight; (b) MTENN (Phase I): only consider phase I to predict event without any sequential uncertainty head; (c) w/o MTENN loss: a variant of MTENN (Phase I) that consider BCE loss, where the probability can be calculated based on the expected probability of Beta distribution. The key findings obtained from this experiment: MTENN loss and sequential uncertainty can enhance early event detection in detection delay and accuracy.



(a) Ablation Study (b) Inference Time

Fig. 4: Ablation study and inference time.

5. CONCLUSION

This work proposes a novel framework, Multi-Label Temporal Evidential Neural Network (MTENN), for early event detection in temporal data. MTENN is able to quality predictive uncertainty for multi-label classifications at each time stamp. In addition, we introduce a novel uncertainty fusion operator based on MTENN to quantify the fused uncertainty of a sub-sequence for early event detection.

6. ACKNOWLEDGMENT

This work is partly supported by the Army Research Office under Grant Contract Number W91NF-20-2-0140 and NSF under Grant Numbers 2107449, 2107450, and 2107451. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein.

References

[1] TK Sai and KA Reddy, "Application of acoustic techniques in thermal power plants," *Sensors & Transducers*, vol. 210, no. 3, pp. 42, 2017.

[2] Junfeng Guo, Ang Li, and Cong Liu, "Aeva: Black-box backdoor detection using adversarial extreme value analysis," *arXiv preprint arXiv:2110.14880*, 2021.

[3] Guoliang He, Wen Zhao, and Xuewen Xia, "Confidence-based early classification of multivariate time series with multiple interpretable rules," *Pattern Analysis and Applications*, 2020.

[4] Lei Zhao, Huiying Liang, Daming Yu, Xinming Wang, and Gansen Zhao, "Asynchronous multivariate time series early prediction for icu transfer," in *ICIMH*, 2019, pp. 17–22.

[5] Ashish Gupta, Hari Prabhat Gupta, and etc., "An early classification approach for multivariate time series of on-vehicle sensors in transportation," *TITS*, 2020.

[6] Minh Hoai and Fernando De la Torre, "Max-margin early event detectors," *IJCV*, vol. 107, no. 2, pp. 191–202, 2014.

[7] Huy Phan, Philipp Koch, Ian McLoughlin, and Alfred Mertins, "Enabling early audio event detection with neural networks," in *ICASSP*. IEEE, 2018, pp. 141–145.

[8] Ashish Gupta, Hari Prabhat Gupta, Bhaskar Biswas, and Tanima Dutta, "A fault-tolerant early classification approach for human activities using multivariate time series," *IEEE Transactions on Mobile Computing*, vol. 20, no. 5, pp. 1747–1760, 2020.

[9] Wenhe Yan, Guiling Li, Zongda Wu, Senzhang Wang, and Philip S Yu, "Extracting diverse-shapelets for early classification on time series," *World Wide Web*, vol. 23, no. 6, pp. 3055–3081, 2020.

[10] Ian Vince McLoughlin, Yan Song, and etc., "Early detection of continuous and partial audio events using cnn," in *Interspeech*. International Speech Communication Association, 2018, vol. 2018, pp. 3314–3318.

[11] Hossein Soleimani, James Hensman, and Suchi Saria, "Scalable joint models for reliable uncertainty-aware event prediction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 8, pp. 1948–1963, 2017.

[12] Xujiang Zhao, Feng Chen, Shu Hu, and Jin-Hee Cho, "Uncertainty aware semi-supervised learning on graph data," *NeurIPS*, vol. 33, 2020.

[13] Murat Sensoy, Lance Kaplan, and etc., "Evidential deep learning to quantify classification uncertainty," *NeurIPS*, 2018.

[14] Liyan Xu, Xuchao Zhang, Xujiang Zhao, Haifeng Chen, Feng Chen, and Jinho D Choi, "Boosting cross-lingual transfer via self-learning with uncertainty estimation," *arXiv preprint arXiv:2109.00194*, 2021.

[15] Audun Jøsang, *Subjective Logic: A Formalism for Reasoning Under Uncertainty*, Springer, 2016.

[16] Nicolas Turpault, Romain Serizel, and etc., "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.

[17] 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*. IEEE, 2021, pp. 366–370.

[18] Audun Jøsang, "Belief calculus," *arXiv preprint arXiv:0606029*, 2006.

[19] Xujiang Zhao, Xuchao Zhang, and etc., "Seed: Sound event early detection via evidential uncertainty," in *ICASSP*. IEEE, 2022, pp. 3618–3622.

[20] Koichi Miyazaki, Tatsuya Komatsu, and etc., "Conformer-based sound event detection with semi-supervised learning and data augmentation," *dim*, vol. 1, pp. 4, 2020.

[21] Yarin Gal and Zoubin Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*. PMLR, 2016, pp. 1050–1059.

[22] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, and etc., "Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning," in *ICML*. PMLR, 2018, pp. 1184–1193.

[23] Nicolas Turpault and Romain Serizel, "Training sound event detection on a heterogeneous dataset," *arXiv preprint arXiv:2007.03931*, 2020.