A Novel Audio-Based Machine Learning Model for Automated Detection of Collision Hazards at Construction Sites

Khang Dang[a] and Tuyen Le[b]

[a]Glenn Department of Civil Engineering, College of Engineering, Computing and Applied Sciences, Clemson University, United States
[b]Glenn Department of Civil Engineering, College of Engineering, Computing and Applied Sciences, Clemson University, United States
E-mail: kdangho@g.clemson.edu, tuyenl@clemson.edu

Abstract –

Collisions between workers and operating vehicles are the leading source of fatal incidents in the construction industry. One of the most prevalent factors causing contact hazards is the decline in construction workers' auditory situational awareness due to the hearing loss and the complicated nature of construction noises. Thus, a computational technique that can augment the audible sense of a worker can significantly improve safety performance. Since construction machines often generate distinct sound patterns while operating at the construction sites, audio signal processing could be an innovative solution to achieve the goal. Unfortunately, the current body of knowledge regarding automated surveillance in construction still lacks such advanced methods. This paper presents a newly developed auditory surveillance framework using convolutional neural networks (CNNs) that can detect collision hazards by processing acoustic signals in construction sites. The study specifically has two primary contributions: (1) a new labeled dataset of normal and abnormal sound events relating to collision hazards in the construction site, and (2) a novel audio-based machine learning model for automated detection of collision hazards. The model was trained with different network architectures, and its performance was evaluated using various measures, including accuracy, recall, precision, and combined F-measure. The research is expected to help increase the auditory situational awareness of construction workers and consequently enhance construction safety.

Keywords –

Machine Learning; Sound Surveillance; Hazard Detection; Construction Safety

1 Introduction

According to the Occupational Safety and Health Administration (OSHA), the annual fatality rate in the construction industry in the United States is relatively high compared to that in other industrial sectors [1]. Most of these fatalities occurred when workers being struck by a construction vehicle. This is because the nature of construction sites often includes potential hazards during the situation that construction workers and heavy mobile equipment are in proximity [2]. The critical factor leading to collision hazard was reported as the decline in auditory situational awareness of construction workers due to the hearing loss [3] and the complicated nature of construction noises [4]. Therefore, a novel audio-based technique that can augment the audible sense of a worker needs to be developed to improve safety performance.

The use of advanced computational techniques in auditory signal processing for hazard detection is motivated by strong acoustic emissions coming from hazardous situations. Hence, it is possible to extract much useful information from sounds at job sites. For example, construction machines often produce unique sound patterns while performing certain activities [5], [6]. Moreover, the detection of acoustic events is further complicated by the heterogeneous sound types of construction equipment operations generated from diverse working environments [7], [8]. In this case, the detection tends to fall under the categorization of construction equipment-related activities. Therefore, abnormal acoustic events that can cause collision hazards are classified as mobile equipment, and normal acoustic events are identified as stationary equipment. This is especially the case when one of the most common causes of construction accidents was "struck by moving vehicles" [9]. Thus, such auditory

surveillance of potential accidents would significantly improve construction safety.

However, sound sensing in the construction field for safety has received little attention from the academic community in the past decade. A majority number of related studies were only focused on tracking various activities of construction equipment to reduce operating costs and the identification of working and operation activities [5], [6], [10], [11]. To address these existing issues, we propose a novel audio-based machine learning model for automated detection of collision hazards at construction sites. The study specifically has two primary contributions: (1) a new labeled dataset of normal and abnormal sound events relating to collision hazards in the construction site, and (2) a CNN model for automated detection of collision hazards.

The remainder of this paper is organized as follows: Section 2 surveys recent related work on the applications of auditor surveillance, and the use of CNN for the detection of abnormal events; Section 3 describes the novel audio-based machine learning model for automated detection of collision hazards; Section 4 describes the setup of an experiment in which we compare the performance of CNN across various datasets; Section 5 discusses the results of this experiment; and, finally, Section 6 summarizes the paper and proposes directions for future work.

2 Related Work

Auditory surveillance technologies in the construction industry help support the construction industry's safety performance since lack of excellent visibility was the principle factor leading to fatalities [1]. However, there is still a lack of such research in the field. Most of the sound-based surveillance technologies were only focused on monitoring construction work activities and equipment operations. For instance, a hybrid system for recognizing multiple construction equipment activities was proposed [11], and a supervised machine learning-based sound identification algorithm was implemented to enhance construction site activity monitoring and performance evaluation [12]. A few studies attempted to develop new approaches for conducting an audio-based event detection system for safety, but some limitations still existed. Experimental trials were designed to deploy sensing technology to provide alerts to proximity detection when heavy construction equipment and workers are in close proximity [2]. Nonetheless, the devices were installed on construction equipment only, not equipped on construction workers. Another approach using a machine learning algorithm can categorize sound events and make construction workers aware of possible safety risks and hazards [7]. Still, the sound data relating to collision hazard was only collected from a particular worksite. Such an approach is restrained because the sounds emitted by the equipment from various construction sites may differ. To address the gap, there is a need to investigate the surveillance approach to detect collision hazards from the perspective of construction workers using the sound collected from multiple construction sites.

Auditory surveillance has been extensively applied for the detection of abnormal events in various contexts and achieved promising results even in environments with complicated noises. For instance, some researchers developed a technique for detecting shouting events in a real-life railway environment [13]. Additionally, efforts have been made focusing on the detection of crimes in elevators [14], and the detection of human emotions based on verbal sounds during hazardous situations in public spaces [15], [16]. Other researchers also implemented signal processing for the surveillance of healthcare facilities, including a system for medical telesurvey [17], and a framework that detects older adults' falls [18].

Previous machine learning approaches to recognize and classify the auditory events typically used conventional machine learning models such as Gaussian Mixture Models (GMM) [19], [20], Hidden Markov Models (HMM) [12], [21], [22], and Support Vector Machine [11], [18], [23]. However, those techniques have been proved to underperform deep learning methods, such as Deep Neural Networks (DNN), in a variety of tasks for sound classification [24]-[26]. Recently, several studies have used a more complex architecture of DNN, such as Convolutional Neural Networks (CNN), for audio classification [27]–[30]. In sound processing, CNN can learn filters that are shifted in both time and frequency so that it can cover numerous input fields [29]. It was also found that the performance of CNN networks for signal classification is highly regulated by the optimum number of convolutional layers [28], [30]. CNN models can also be trained with a back-propagation mechanism that consists of two processes, including the pattern creation process and the testing process [31]. Motivated by these recent impressive performances in auditory surveillance, we applied CNN to the detection of hazard collisions in raw audio.

3 Novel Audio-Based Machine Learning Model for Automated Detection of Collision Hazards

Hereby, we present an innovative framework using supervised deep learning for sound detection associated with the recognition of mobile equipment. This framework is based on processing audio signals

generated at construction jobsites. The overall process, including three main steps, is illustrated in Figure 1. First, audio files are collected and labelled as abnormal and normal types. Then, acoustic features are extracted using the Fast Fourier Transform (FFT) function. Those features are the input of the CNN model, which is trained on the labelled data to detect acoustic events.

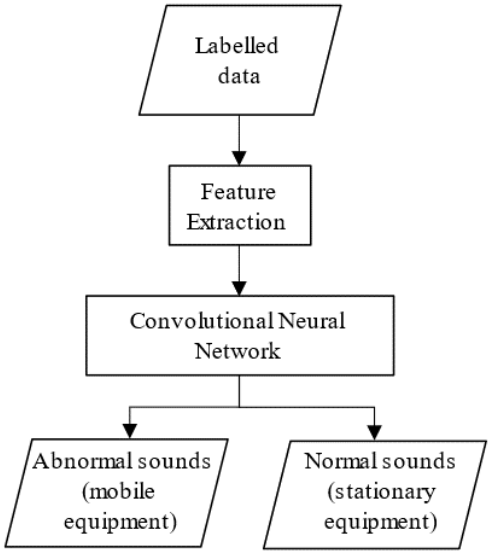

Figure 1. Overall flowchart for automated detection of collision hazards

3.1 Labelled Data

Given the goal of creating an audio event detection of collision hazards, we need to define the set of events the system should recognize. Thus, the sound sources originated from construction equipment-related activities need to be classified as a set of classes. This set of classes will allow us to collect labelled data for training and evaluation. Since the occurrence of abnormal sounds is an essential indicator of dangerous situations requiring quick safety responses, collected sound events are labelled into normal and abnormal types.

3.2 Feature Extraction

This step aims to extract the acoustic features in both time and frequency domains from audio signals by using different extraction functions. In this work, the most commonly used Mel-Frequency Cepstral Coefficients (MFCCs) are extracted by the FFT. They are mainly used to depict the spectral envelope in a significant number of audio processing applications. Through feature extraction, the components of the sound signals that are good for identifying the sound contents are recognized. In other words, the feature extraction process transforms the raw signals into feature vectors in which specific properties of audio signals are emphasized.

To obtain MFCCs from a discrete audio signal, the audio signal undergoes a pre-emphasis process, where the extraction function FFT is employed to convert the signal to the frequency domain. Then, the spectrum of the frequency domain is fed into mel-filter banks. Each filter has a center frequency called the filter bank energies. This compression operation makes the acoustic features match more closely to what humans hear. In the following step, the Discrete Cosine Transform (DCT) is applied to filter bank energies. The output coefficients of DCT are called Mel Frequency Cepstral Coefficients (MFCCs). It is worth noting that only 13 MFCC coefficients are extracted in our work, as recommended in several studies in sound classification [32]–[34]. This is because the higher MFCC coefficients represent fast changes in the filter bank energies, and it turns out that these fast changes actually degrade sound classification performance. The MFCCs extracted from the sound signal are stored as an array of values. The vertical axis represents the number of MFCCs calculated in order and the horizontal axis represents the number of frames available.

3.3 Convolutional Neural Network

After the feature extraction is completed, the CNN model is developed for sound detection with the array of the MFCC values as the input. The deep CNN architecture proposed in this study is comprised of four convolutional layers, as depicted in Figure 2, followed by a max-pooling layer, a dropout layer, a flatten layer, and two fully connected layers to get the output. The activation function used for convolutional layers and dense layers is the Rectified Linear Unit, which is most commonly used in deep learning models. The function returns zero if it receives any negative input, but for any positive value, it returns the same amount back. The Softmax activation function is applied to the output layer. The output is a prediction of which class an audio belongs to.

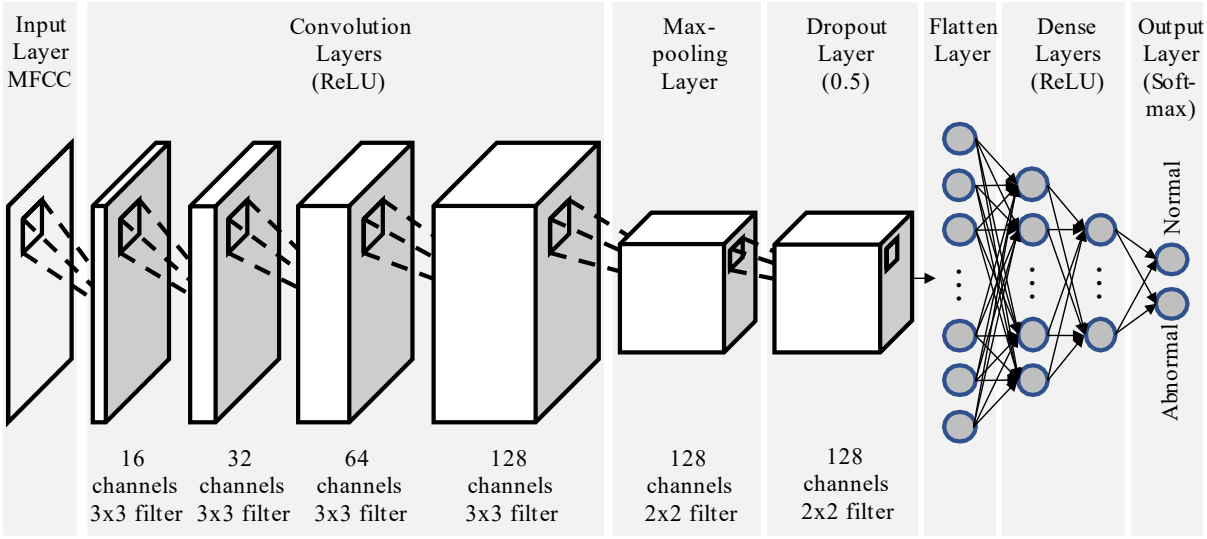

Figure 2. Detailed architecture of the convolutional neural network

The prediction is achieved through several steps in training the neural network. The first step is convolution, which is a process of taking a small matrix of numbers, called filter, then passing it over our input and transform it based on the values from the filter. Subsequent feature map values throughout convolution layers are calculated. The following steps are achieved through Max-pooling and Dropout layers. While the Max-pooling layer selects a maximum value from each region and put it in the corresponding place in the output, Dropout works by randomly setting the outgoing values to 0 at each update of the training phase to prevent overfitting. Then the shape of the data is changed from a two-dimensional matrix to a one-column vector, which is the correct format for dense layers to interpret in the last step. Each dense layer consists of neurons represented by nodes. Each node of the previous layer is connected to all nodes of the next layer. This connection is defined as a scalar value called weights. The model adjusts its weights by a training process called backpropagation. The backpropagation process can be separated into four distinct steps: the feedforward pass, the loss function, the backward pass, and the weight update. At first, the weights are randomized, and the feedforward pass is implemented. Then, the backpropagation is processed with a loss function. In this work, a loss function is represented as categorical cross-entropy, which is a great measure to distinguish two discrete probability distributions. To achieve the correct prediction, the amount of loss needs to be reduced. Therefore, the model finds out which weights most directly contributed to the network's loss and adjusts the weights so that the loss decreases. Finally, all the weights are taken and updated.

4 Experimental Setup

4.1 Datasets

Since the videos on YouTube, which is the most popular video sharing website, have become a treasure of data, the audio files of the dataset prepared for this research were extracted from this abundant source. To extract high-quality sounds from the videos, the authors avoid noisy backgrounds by using only videos that enable a broad view of recordings to ensure that no other irrelevant sound sources affect the sound quality. The audio files were then converted into wav format at 44.1 kHz sampling rate, 16-bit depth, and mono channel. An extensive set of acoustic signals in the construction site were finalized, as shown in Table 1. This dataset consists of two classes: the abnormal class includes sound excerpts from nine mobile equipment, and the normal class includes sound excerpts from seven stationary equipment. The original audio files extracted from YouTube videos were segmented into smaller frames with an equal length of 3 seconds and 2/3 overlapping. The total duration in seconds of audio files in each subset of the abnormal class, the normal class is summarized in Table 1. At this stage, the total number of the whole dataset is 3,629 audio files.

Table 1. Number of original examples in each subset of data

Abnormal class		Normal class	
Type	Total duration (s)	Type	Total duration (s)
Excavator	522	Pneumatic tamper	459

Bulldozer	387	Concrete pumper	180
Grader	1185	Pile driver	174
Front end loader	558	Pneumatic breaker	312
Forklift	123	Steel welding	1533
Compactor roller	855	Hammer	291
Scraper	2712	Saw	807
Water truck	147		
Crane	642		

The audio files were artificially mixed with background audio from two backgrounds, the wind noise, and the street noise. The mixture has the purpose of evaluating the CNN model performance to see if the model is more difficult to detect in one environment than the other. A noisy training dataset was created to enable the model the ability to detect acoustic events among noisy background. Specifically, the audio files were mixed at different "signal-to-noise" ratios (SNRs) (-15dB, -5dB, 5dB, 15dB). As a result, two different datasets with wind-noise background and street-noise background were created. Each dataset has a total number of 14,516 samples.

4.2 Model Training

The authors trained the CNN model on each dataset. 64% of the samples in each dataset were used for training, and the remaining 16% and 20% of the samples were used for validation and testing, respectively. With each dataset, the training procedure was stopped after 15 epochs when the good performance was achieved on the validation set. Various measures, specifically accuracy, recall, precision, and F1-score, are used to evaluate the sound detection performance. Accuracy is the number of correct predictions divided by the total number of predictions. While precision quantifies the number of positive class predictions that actually belong to the positive class, recall quantifies the number of positive class predictions made out of all positive examples in the dataset. F1-score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

5 Results

The system correctly classified almost all audio files. The performance of the proposed CNN model on the two datasets is shown in Table 2. We found that the CNN model performed good predictions on both datasets, with "wind" and "street" backgrounds. Besides, the results show that the ability of the model to detect collision hazards is affected by different acoustic environments. As can be seen that the accuracy of the model slightly varies across the different acoustic scenarios, with 98.52% and 97.49% in the "wind" and "street" scenarios, respectively. This concludes that the model detects the abnormal sounds relating to collision hazards better in the wind-noisy background with no missed "abnormal" detection (the precision of "normal" detection = 1.00).

Table 2. Measures for the performance of the proposed CNN model on each dataset

Dataset	Wind background		Street background	
Class	Ab-normal	Normal	Ab-normal	Normal
Precision	0.98	1.00	0.97	0.99
Recall	1.00	0.96	1.00	0.93
F1-score	0.99	0.98	0.98	0.96
Accuracy	98.52%		97.49%	

6 Conclusion and Future Work

In this paper, the authors applied a machine learning technique for automated detection of collision hazards. The presented framework was tested using multiple audio files collected from YouTube videos, and the results are profound. As the first stage of this research project, we found that the proposed CNN model trained on two datasets accurately recognizes sound patterns of sounds from mobile equipment.

The limitation of this research is that the model was not developed to work on more sorts of background noises provided that a construction site is considered as noisy workplace. Besides, the system could not capture the localization of mobile equipment provided that a mobile vehicle moving toward a worker is a risk and a mobile vehicle moving away the worker is considered as safety. As the plan for future research, the authors intend to develop a model that can operate across different noise backgrounds, such as rain and ocean wave sounds or the noise from people talking, and consider the factor of localizing the sound source. This work is a great start to build more realistic models that can work on detecting collision hazards under the complicated nature of construction noises.

7 Acknowledgments

This research was funded by the National Science Foundation (NSF) through the Award # 1928550. The authors gratefully acknowledge NSF's support. Any opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do

not necessarily reflect the views of NSF.

References

[1] J. W. Hinze and J. Teizer, "Visibility-related fatalities related to construction equipment," Saf. Sci., vol. 49, no. 5, pp. 709–718, Jun. 2011, doi: 10.1016/j.ssci.2011.01.007.

[2] E. D. Marks and J. Teizer, "Method for testing proximity detection and alert technology for safe construction equipment operation," Constr. Manag. Econ., vol. 31, no. 6, pp. 636–646, Jun. 2013, doi: 10.1080/01446193.2013.783705.

[3] T. C. Morata, C. L. Themann, R. F. Randolph, B. L. Verbsky, D. C. Byrne, and E. R. Reeves, "Working in noise with a hearing loss: perceptions from workers, supervisors, and hearing conservation program managers," Ear Hear., vol. 26, no. 6, pp. 529–545, Dec. 2005, doi: 10.1097/01.aud.0000188148.97046.b8.

[4] E. Vinnik, P. M. Itskov, and E. Balaban, "Individual differences in sound-in-noise perception are related to the strength of short-latency neural responses to noise," PLoS One, vol. 6, no. 2, pp. 1–8, 2011, doi: 10.1371/journal.pone.0017266.

[5] C. F. Cheng, A. Rashidi, M. A. Davenport, and D. V. Anderson, "Activity analysis of construction equipment using audio signals and support vector machines," Autom. Constr., vol. 81, pp. 240–253, Sep. 2017, doi: 10.1016/j.autcon.2017.06.005.

[6] C. F. Cheng, A. Rashidi, M. A. Davenport, and D. Anderson, "Audio signal processing for activity recognition of construction heavy equipment," in ISARC 2016 - 33rd International Symposium on Automation and Robotics in Construction, 2016, pp. 642–650, doi: 10.22260/isarc2016/0078.

[7] Y. C. Lee, M. Shariatfar, A. Rashidi, and H. W. Lee, "Evidence-driven sound detection for prenotification and identification of construction safety hazards and accidents," Autom. Constr., vol. 113, p. 103127, May 2020, doi: 10.1016/j.autcon.2020.103127.

[8] Y. Xie et al., "Historical Accident and Injury Database-Driven Audio-Based Autonomous Construction Safety Surveillance."

[9] A. Perlman, R. Sacks, and R. Barak, "Hazard recognition and risk perception in construction," Saf. Sci., vol. 64, pp. 13–21, Apr. 2014, doi: 10.1016/j.ssci.2013.11.019.

[10] C. A. Sabillon, A. Rashidi, B. Samanta, C. F. Cheng, M. A. Davenport, and D. V. Anderson, "A productivity forecasting system for construction cyclic operations using audio signals and a Bayesian approach," in Construction Research Congress 2018: Construction Information Technology - Selected Papers from the Construction Research Congress 2018, 2018, vol. 2018-April, pp. 295–304, doi: 10.1061/9780784481264.029.

[11] Sherafat, Rashidi, Lee, and Ahn, "A Hybrid Kinematic-Acoustic System for Automated Activity Detection of Construction Equipment," Sensors, vol. 19, no. 19, p. 4286, Oct. 2019, doi: 10.3390/s19194286.

[12] T. Zhang, Y. C. Lee, M. Scarpiniti, and A. Uncini, "A supervised machine learning-based sound identification for construction activity monitoring and performance evaluation," in Construction Research Congress 2018: Construction Information Technology - Selected Papers from the Construction Research Congress 2018, 2018, vol. 2018-April, pp. 358–366, doi: 10.1061/9780784481264.035.

[13] J.-L. Rouas, J. Louradour, and S. Ambellouis, "Audio Events Detection in Public Transport Vehicle," in 2006 IEEE Intelligent Transportation Systems Conference, 2006, pp. 733–738, doi: 10.1109/ITSC.2006.1706829.

[14] R. Radhakrishnan and A. Divakaran, "Systematic acquisition of audio classes for elevator surveillance," in Image and Video Communications and Processing 2005, 2005, vol. 5685, p. 64, doi: 10.1117/12.587814.

[15] C. Clavel, I. Vasilescu, L. Devillers, and T. Ehrette, "Fiction database for emotion detection in abnormal situations," in 8th International Conference on Spoken Language Processing, ICSLP 2004, 2004.

[16] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," Speech Commun., vol. 50, no. 6, pp. 487–503, Jun. 2008, doi: 10.1016/j.specom.2008.03.012.

[17] M. Vacher, D. Istrate, L. Besacier, J.-F. Serignat, E. Castelli, and L. Besacier, "Sound Detection and Classification for Medical Telesurvey," 2004.

[18] S. M. Adnan, A. Irtaza, S. Aziz, M. O. Ullah, A. Javed, and M. T. Mahmood, "Fall detection through acoustic Local Ternary Patterns," Appl. Acoust., vol. 140, pp. 296–300, Nov. 2018, doi: 10.1016/j.apacoust.2018.06.013.

[19] W. Choi, J. Rho, D. K. Han, and H. Ko, "Selective background adaptation based abnormal acoustic event recognition for audio surveillance," in Proceedings - 2012 IEEE 9th International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2012, 2012, pp. 118–123, doi: 10.1109/AVSS.2012.65.

[20] Y. Lee, D. K. Han, and H. Ko, "Acoustic signal

based abnormal event detection in indoor environment using multiclass adaboost," IEEE Trans. Consum. Electron., vol. 59, no. 3, pp. 615–622, 2013, doi: 10.1109/TCE.2013.6626247.

[21] "An abnormal sound detection and classification system for surveillance applications - IEEE Conference Publication." [Online]. Available: https://ieeexplore.ieee.org/document/7096526. [Accessed: 19-Nov-2019].

[22] C. V. Cotton and D. P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2011, pp. 69–72, doi: 10.1109/ASPAA.2011.6082331.

[23] Y. Alsouda, S. Pllana, and A. Kurti, "A Machine Learning Driven IoT Solution for Noise Classification in Smart Cities," arXiv Prepr. arXiv1809.00238, Sep. 2018.

[24] M. Asgari, I. Shafran, and A. Bayestehtashk, "Inferring social contexts from audio recordings using deep neural networks," in IEEE International Workshop on Machine Learning for Signal Processing, MLSP, 2014, doi: 10.1109/MLSP.2014.6958853.

[25] M. Ravanelli, B. Elizalde, K. Ni, and G. Friedland, "Audio concept classification with Hierarchical Deep Neural Networks," in European Signal Processing Conference, 2014.

[26] O. Gencoglu, T. Virtanen, and H. Huttunen, "Recognition of acoustic events using deep neural networks," in European Signal Processing Conference, 2014.

[27] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," IEEE Signal Process. Lett., vol. 24, no. 3, pp. 279–283, Mar. 2017, doi: 10.1109/LSP.2017.2657381.

[28] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), 2015, vol. 2015-Novem, pp. 1–6, doi: 10.1109/MLSP.2015.7324337.

[29] A. Salekin, S. Ghaffarzadegan, Z. Feng, and J. Stankovic, "A Real-Time Audio Monitoring Framework with Limited Data for Constrained Devices," 2019, pp. 98–105, doi: 10.1109/dcoss.2019.00036.

[30] J. Salamon, J. P. Bello, A. Farnsworth, and S. Kelling, "Fusing shallow and deep learning for bioacoustic bird species classification," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2017, pp. 141–145, doi: 10.1109/ICASSP.2017.7952134.

[31] A. Glowacz, "Acoustic based fault diagnosis of three-phase induction motor," Appl. Acoust., vol. 137, pp. 82–89, Aug. 2018, doi: 10.1016/j.apacoust.2018.03.010.

[32] B. Kim and B. Pardo, "A human-in-the-loop system for sound event detection and annotation," ACM Trans. Interact. Intell. Syst., vol. 8, no. 2, Jul. 2018, doi: 10.1145/3214366.

[33] Z. H. Janjua, M. Vecchio, M. Antonini, and F. Antonelli, "IRESE: An intelligent rare-event detection system using unsupervised learning on the IoT edge," Eng. Appl. Artif. Intell., vol. 84, pp. 41–50, Sep. 2019, doi: 10.1016/j.engappai.2019.05.011.

[34] D. Carmel, A. Yeshurun, and Y. Moshe, "Detection of alarm sounds in noisy environments," in Signal Processing Conference (EUSIPCO), 2017 25th European, 2017, vol. 2017-Janua, pp. 1839–1843, doi: 10.23919/EUSIPCO.2017.8081527.