INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Design of Audio-Based Accident and Crime Detection and Its Optimization

Afis Asryullah Pratama [a,*], Sritrusta Sukaridhoto [a], Mauridhi Hery Purnomo [b], Vita Lystianingrum [c], Rizqi Putri Nourma Budiarti [d]

[a] *Department of Electronic Engineering, Politeknik Elektronika Negeri Surabaya, Surabaya, Indonesia*
[b] *Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[c] *Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[d] *Engineering Department, Universitas Nahdlatul Ulama Surabaya, Surabaya, Indonesia*

*Corresponding author: *afisarsy@gmail.com*

***Abstract*—The development of transportation technology is increasing every day; it impacts the number of transportation and their users. The increase positively impacts the economy's growth but also has a negative impact, such as accidents and crime on the highway. In 2018, the number of accidents in Indonesia reached 109,215 cases, with a death rate of 29,472 people, which was mostly caused by the late treatment of the casualties. On the other hand, in the same year, there were 8,423 mugs, and 90,757 snitches cases in Indonesia, with only 23.99% of cases reported. This low reporting rate is mostly caused by the lack of awareness and knowledge about where to report. Therefore, a quick response surveillance system is needed. In this study, an audio-based accident and crime detection system was built using a neural network. To improve the system's robustness, we enhance our dataset by mixing it with certain noises which likely to occur on the road. The system was tested with several parameters of segment duration, bandpass filter cut-off frequency, feature extraction, architecture, and threshold values to obtain optimal accuracy and performance. Based on the test, the best accuracy was obtained by convolutional neural network architecture using 200ms segment duration, 0.5 overlap ratio, 100Hz and 12000Hz as bandpass cut-off frequency, and a threshold value of 0.9. By using mentioned parameters, our system gives 93.337% accuracy. In the future, we hope to implement this system in a real environment.**

***Keywords*— Audio recognition; dataset manipulation; optimization; neural networks; surveillance system.**

Manuscript received 11 Jan. 2022; revised 7 Mar. 2022; accepted 29 Apr. 2022. Date of publication 31 Mar. 2023.

I. INTRODUCTION

Emergency situation happens rarely and unpredictably, but the chance never is 0. Emergency situations cause individuals or groups to shift their focus to handle the situation [1]. The most challenging part during an emergency is keeping calm and responding with a fast and fitting act [2]. In this research, we particularly focused on accident and crime situations. The advancement of transportation technology affects the number of vehicles and their passengers. In 2018, 146,858,759 vehicles were recorded in Indonesia, classified as passenger cars, buses, freight cars, and motorcycles [3], [4]. The increment of vehicles also affects the number of accidents that happen. In Indonesia, 109,215 accidents and 29,472 deaths were recorded in 2018 [5]. Most of the deaths were caused by the late treatment of the casualties [6].

The other emergency situation we focused on is a crime. The data by the Central Bureau of Statistics reported that in 2018, 8,423 mugs and 90,757 snitches cases occurred in Indonesia, of which 23.99% were reported. This low reporting rate happened due to a lack of awareness and knowledge about where to report [7]. For that reason, a reliable system which able to detect accidents and crimes was needed. By referencing other research that uses audio recognition mostly focuses on accidents or impulsive sound detection that withstands environmental noises [8]–[18].

In this research, we propose an audio-based accident and crime detection system and tune several parameters in the overall process to obtain the optimal result. To increase its accuracy and robustness, we enhance the used dataset by mixing the raw audio with several noises related to the real environment.

II. Materials and Method

Our proposed method is divided into two major processes: dataset creation and inference. The dataset creation process aims to create a dataset with various noise mixed to improve the inference accuracy and robustness, while the inference process mainly aims to recognize and decide whether it is normal, accident, or crime based on the audio.

A. Dataset Creation

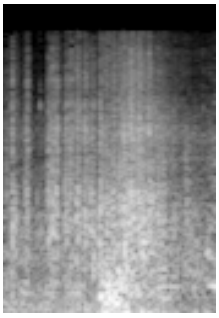
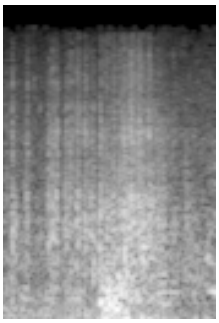
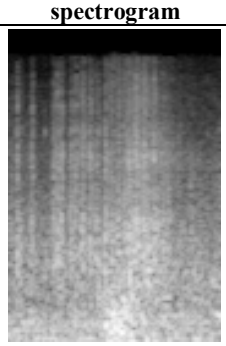
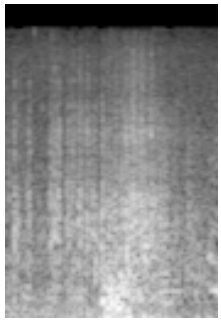
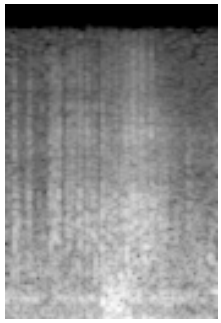
We collect audio data labeled as a car crash, engine idling, gunshot, rain, road traffic, scream, thunderstorm, and wind from various resources, including other publications [8], [19]–[21] and YouTube, the chosen labels represent the normal, accident, and crime condition. Collected data were then resampled to 44100Hz and enhanced by mixing it with environmental noises such as rain, road traffic, thunderstorm, and wind. The enhancement process was done using the following rules.

TABLE I
MIXING RULES

label	noises				
	none	rain	road traffic	thunder storm	wind
car crash	√	√	√	√	√
engine idling	√	√	√	√	√
gunshot	√	√	√	√	√
rain	-	-	-	-	-
road traffic	√	√	-	√	√
scream	√	√	√	√	√
thunderstorm	-	-	-	-	-
wind	-	-	-	-	-

The mixing process changes the sound of the data based on the used noise. These changes provide wider data coverage, which were benefit the real scene [22].

TABLE II
MIXING RESULT

noise type	spectrogram
none	
rain	
road traffic	
thunderstorm	
wind	

This process gives a total of 5352 audio data divided into eight labels. Twenty data from each label were randomly excluded as the data test, and the rest were divided into train and validation data with a 7:3 ratio. As a result, we used 3,635 train data, 1,557 validation data, and 160 test data.

B. Inference

The inference method consists of segmentation, Bandpass filter, Short Time Fourier Transform (STFT), Mel spectrogram, neural network, and thresholding, as shown in Fig. 1.

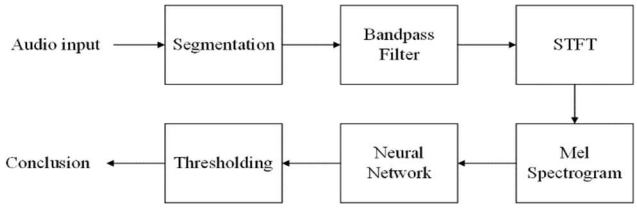


Fig. 1 Inference workflow

1) Segmentation: This process slices the audio into smaller segments to reduce the processing load and fasten the response for each segment. A good segmentation process is required due to the important information of audio, mostly not at the same part of a segment [23]. Thus, this research conducts two different parameters: segment duration and overlap ratio. We also use overlapped segmentation process,

which gives a higher accuracy than the non-overlapped segmentation for the recognition system because it has less correlation with its adjacent segments [24].

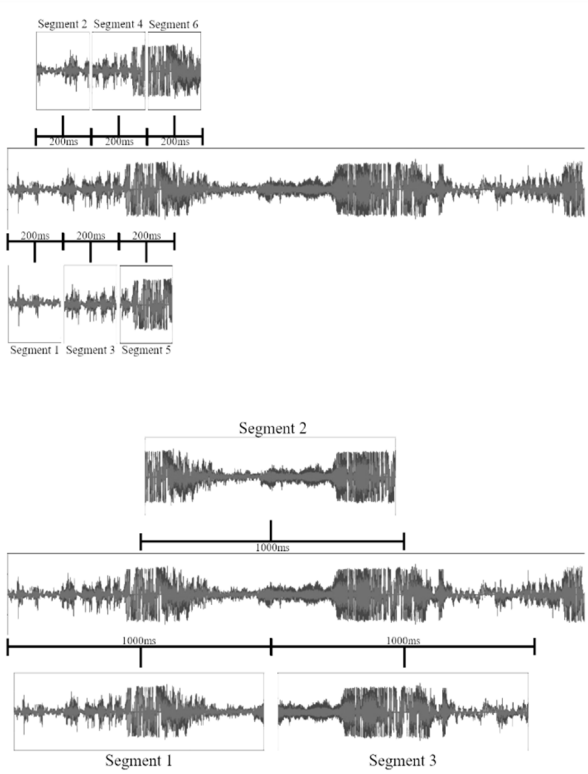
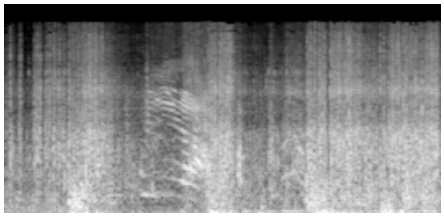
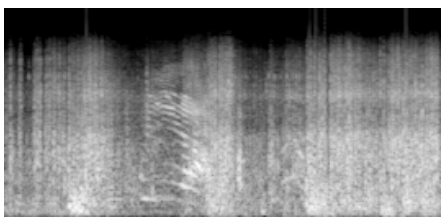
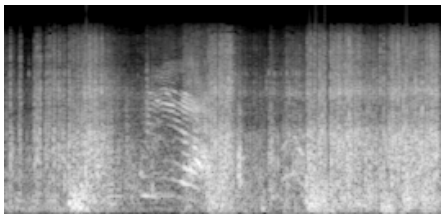
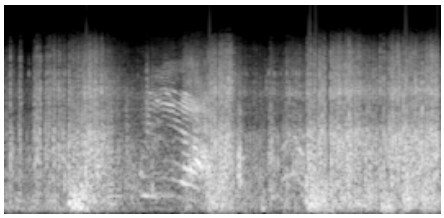
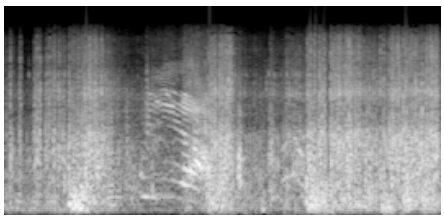


Fig. 2 Audio segmentation

As shown in Fig. 2 each segment overlapping to its adjacent. In this research, we test our system with various segmentation parameters TABLE IV.

2) Bandpass filter: This process occurred to reduce noises based on their frequency. Bandpass filter could generate fine samples, increasing the system's robustness [25]. The key point of this process is the cut-off frequency used. In this research, we use a 4th-order bandpass filter with various combinations of cut-off frequencies. The impact of the bandpass filter on the audio data is shown in the spectrogram in TABLE III.

TABLE III
FILTERED AUDIO

cut-off frequency (Hz)	spectrogram
none	
80, 8000	
80, 12000	
100, 8000	
100, 12000	

3) STFT: This process is an improvement of the Fast Fourier Transform that calculates the Fourier transform coefficients in a smaller time fraction [26]. STFT was chosen due to its speed and no repetition data. STFT is a key component for signal processing systems with a wide application range, such as medicine, industrial measurement and control, and audio signals analysis [27]. STFT step consists of three subprocesses as follows.

Fig. 3 STFT process

Framing is a process of capturing a smaller piece of the segment. In this research, we frame each segment into smaller frames with a frame width of 1,764 samples and a hop length of 441 samples. Windowing is a process to avoid spectral leakage by reducing spikes at the start and end of the frame. One of the windowing methods is Hann window (1). We used Hann window with window width equal to frame width.

$$w(k) = 0.5 \cdot \left(1 - \cos\left(\frac{2\pi k}{K-1}\right)\right), k = 1 \dots K \qquad (1)$$

Fast Fourier Transform (FFT) is a faster process to calculate Fourier transform. We use FFT with Fourier width same as window and frame width. The FFT equation is shown in (2).

$$\hat{x}\left(\frac{k}{N}\right) = \sum_{n=0}^{N-1} x(n) \cdot e^{-2i\pi n \frac{k}{N}}, k = 0 \dots N-1 \qquad (2)$$

This process converts time-domain audio data into a spectrogram based on user parameters.

4) Mel Spectrogram: Human hearing perception of frequencies is logarithmic, which means that human hearing has a higher resolution at high frequencies. In order to utilize our hearing system, we convert the spectrogram into a Mel scale. Mel Spectrogram is a data form made of a combination between the Mel scale and spectrogram to represent frequency and amplitude by the time domain [28]. We use a total of 128 Mel bands for each spectrogram.

5) Neural Network: Most used architectures for audio recognition are Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Deep Neural Network (DNN) architectures as the classification system, with the best result mostly performed by RNN and DNN [29]. In this research, we test and compare all the mentioned architectures as the classification system. We used the Mel spectrogram as input, and eight output refers to 8 labels from the enhanced dataset. The same training parameters were applied to all tested architectures.

TABLE IV
MODELS PARAMETERS

model	architecture	segment duration (ms)	overlap ratio	lower cut-off (Hz)	upper cut-off (Hz)
C_1	CNN	200	0.5	80	8000
C_2	CNN	200	0.5	80	12000
C_3	CNN	200	0.5	100	8000
C_4	CNN	200	0.5	100	12000
C_5	CNN	1000	0.5	80	8000
C_6	CNN	1000	0.5	80	12000
C_7	CNN	1000	0.5	100	8000
C_8	CNN	1000	0.5	100	12000
R_1	RNN	200	0.5	80	8000
R_2	RNN	200	0.5	80	12000
R_3	RNN	200	0.5	100	8000
R_4	RNN	200	0.5	100	12000
R_5	RNN	1000	0.5	80	8000
R_6	RNN	1000	0.5	80	12000
R_7	RNN	1000	0.5	100	8000
R_8	RNN	1000	0.5	100	12000
D_1	DNN	200	0.5	80	8000
D_2	DNN	200	0.5	80	12000
D_3	DNN	200	0.5	100	8000
D_4	DNN	200	0.5	100	12000
D_5	DNN	1000	0.5	80	8000
D_6	DNN	1000	0.5	80	12000
D_7	DNN	1000	0.5	100	8000
D_8	DNN	1000	0.5	100	12000

By combining different parameters from the mentioned process, we created twenty-four models to find the optimal parameter and performance of the system.

6) Thresholding: Thresholding was used to reduce the false positive result from the Neural Network. The Neural Network's output was mapped into three types of output with mapping rules shown in TABLE V and thresholding based on its confidence value. The three types of output represent the outcome of accident and crime detection.

TABLE V
LABEL TO OUTPUT MAP

label	output
engine idling	normal
rain	normal
road traffic	normal
thunderstorm	normal
wind	normal
car crash	accident
gunshot	crime
scream	crime

To obtain optimal accuracy, the threshold value was tuned by a trial-and-error process [30].

III. RESULTS AND DISCUSSION

In this section, we describe the result of experiments from our proposed method. The experiments were divided into neural network classification, thresholding, and comparison section. The neural network classification section contains model, architecture, segmentation and filter analysis, and the Thresholding section contains thresholding analysis. Moreover, the comparison section provides a brief comparison with another related research.

A. Neural Network Classification

We test our models with 160 data tests consisting of 20 audio data from each label we prepared before. The test result is shown in the following table.

TABLE VI
CLASSIFICATION ACCURACY

model	accuracy (%)
C_1	70.98
C_2	70.43
C_3	70.85
C_4	72.59
C_5	70.19
C_6	71.77
C_7	70.88
C_8	72.46
R_1	66.12
R_2	70.01
R_3	64.03
R_4	68.93
R_5	71.5
R_6	74.38
R_7	75.21
R_8	77.54
D_1	67.76
D_2	72.07
D_3	67.28
D_4	68.65
D_5	76.3
D_6	73.28
D_7	73.56
D_8	79.81

The result shows that each model gives various accuracy starting from model R_3, with the lowest accuracy at 64.03%, and model D_8, with the best classification result with an accuracy of 79.81%. The next step is to find the optimal architecture type. We analyzed the average accuracy from tested architecture: CNN, RNN, and DNN. The result is shown in TABLE VII.

TABLE VII
ARCHITECTURE ANALYSIS

architecture	average accuracy (%)
CNN	71.27
RNN	70.97
DNN	72.34

Based on TABLE VII, RNN architecture gives the lowest average accuracy at 70.97%, and DNN gives the best performance with an average accuracy of 72.34%. Each architecture only gives a slightly different average accuracy from others. In TABLE VIII, we analyzed the impact of segmentation parameters on system accuracy to conclude the optimal parameter value of the system.

TABLE VIII
SEGMENTATION PARAMETERS ANALYSIS

segment duration (ms)	overlap ratio	average accuracy (%)
200	0.5	69.14
1000	0.5	73.91

TABLE VIII shows that models with 1000ms segment duration give a better average accuracy than models with 200ms segment duration. A shorter segment duration means fewer data to be processed, and the total data in 200ms segment duration is mostly insufficient to analyze properly. Behavior analysis of bandpass filter parameters was done to obtain the optimal cut-off frequency range for accident and crime detection systems.

TABLE IX
BANDPASS FILTER PARAMETERS ANALYSIS

lower cut-off frequency (Hz)	upper cut-off frequency (Hz)	average accuracy (%)
80	8000	70.48
80	12000	71.99
100	8000	70.3
100	12000	73.34

Based on the analysis, we found that the best result was obtained from models with 100Hz and 12000Hz cut-off frequencies, which gained an average accuracy of 73.34%. This means the accident and crime audio mostly occurred at a frequency between 100-12000Hz.

B. Thresholding

We apply the thresholding process to the best model of each architecture, which are C_4, R_8, and D_8. Various threshold value was used to find the optimum performance for each selected model. Then we compare the accuracy of the selected model with and without thresholding.

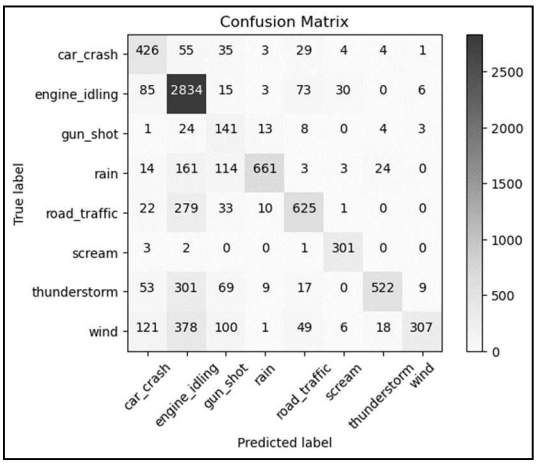


Fig. 4 Confusion matrix of model C_4 confusion matrix.

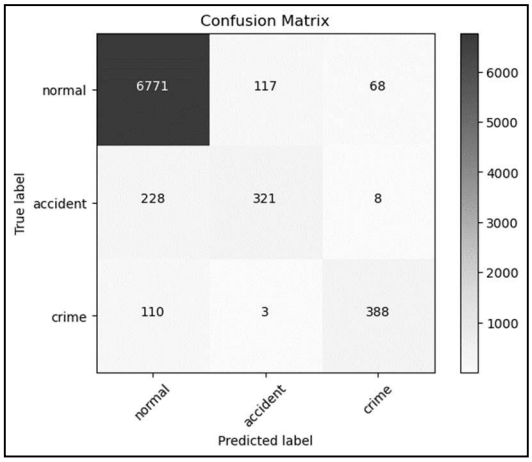


Fig. 5 Confusion matrix of model C_4 with Thresholding at 0.9

The accuracy of model C_4 increases from 72.59% to 93.34%. Recorded a 20.75% accuracy improvement.

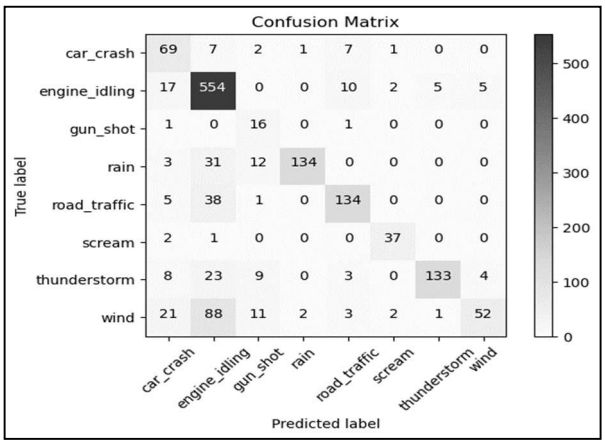


Fig. 6 Confusion matrix of model R_8

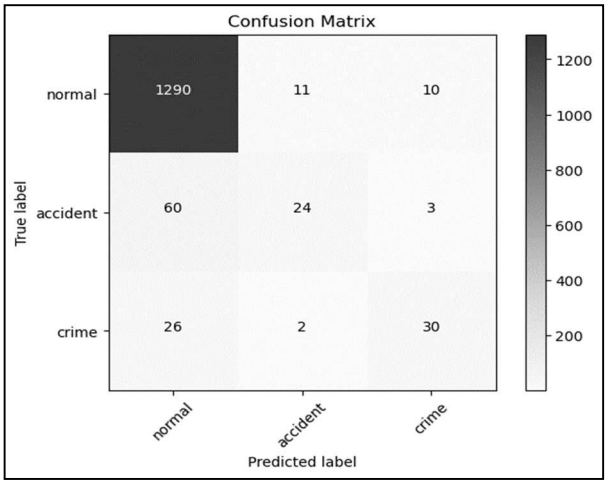


Fig. 7 Confusion matrix of model R_8 with thresholding at 0.9

The accuracy of model R_8 increases from 77.54% to 92.31%. Fig 7 shows a 14.77% accuracy improvement.

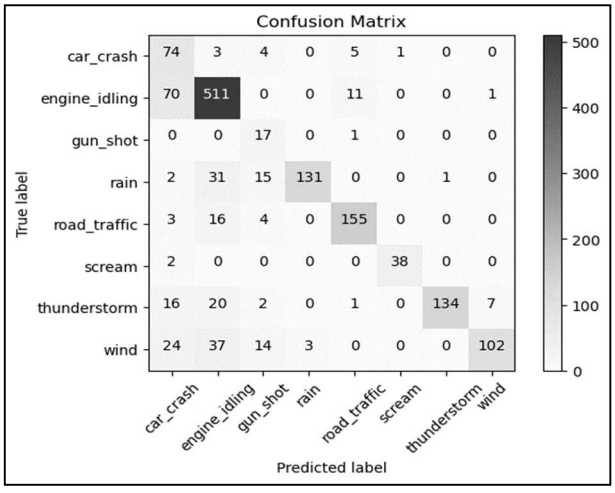


Fig. 8 Confusion matrix of model D_8

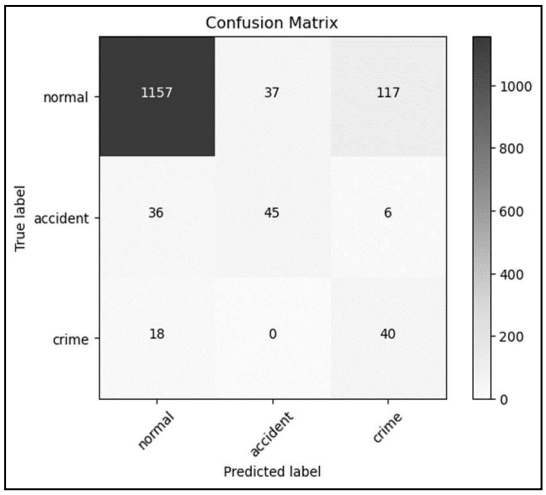


Fig. 9 Confusion matrix of model D_8 with thresholding at 0.95

The accuracy of model D_8 increases from 79.81% to 85.3%, giving a 5.49% accuracy improvement. Overall, model C_4 gives the best accuracy improvement due to its prediction error mostly in the same output label due to the mapping table in TABLE V.

C. Comparison

We compare our proposed method with methods from Sammarco et al. [8], Gatto et al. [9], and Arslan et al. [11] in terms of accuracy. The comparison details are shown in Table X.

TABLE X
ACCURACY COMPARISON

method	accuracy (%)
our proposed method	85.3-93.34
Sammarco et al.	65 – 82
Gatto et al.	93
Arslan et al.	98.4

Our proposed method performs better than the method from Sammarco et al. [8] and Gatto et al. [9] but is unable to beat the methods presented by Arslan et al. [11].

IV. CONCLUSION

From the experiments, we can conclude that our proposed method can recognize accidents and crimes using audio data with an accuracy of 85.3-93.34%. The thresholding process could improve the accuracy. The optimal parameters are CNN architecture, 200ms segment duration, 0.5 overlap ratio, 100Hz and 12000Hz as bandpass cut-off frequency, and a threshold value of 0.9. In the future, we hope to improve our method with more dataset and implement it into an embedded system to test its accuracy and robustness in the real environment.

NOMENCLATURE

w	window coefficients
k	sample index (discrete)
K	window width sample
$\hat{x}$	discrete Fourier series
N	Fourier width
x	input samples
n	sample index (continue)

ACKNOWLEDGMENT

This research was registered in World Class Professor program with number 082/E4.1/AK.04.PT/2021. Special gratitude to the World Class Professor program, which funded this research. Also, the authors are grateful to the Design of Audio-based Accident and Crime Detection member and the optimization team.

REFERENCES

[1] B. van de Walle and M. Turoff, "Decision Support for Emergency Situations," *Handb. Decis. Support Syst. 2*, pp. 39–63, 2008, doi: 10.1007/978-3-540-48716-6_3.
[2] J. Radianti, S. G. Martinez, B. E. Munkvold, and M. Konnestad, "Co-Designing a Virtual Training Tool for Emergency Management," no. May, 2018.
[3] F. L. Munthe, N. Sinaga, and B. Yunianto, "Perancangan dan Pembuatan Sistem Akuisisi Data Dinamometer Sasis Sepeda Motor Berbasis Labview serta Pengujiannya pada Sepeda Motor Honda Beat FI 110 CC," *J. Tek. MESIN*, vol. 10, no. 1, pp. 69–78, 2018.
[4] A. Mahfuzhon and G. E. Setyawan, "Rancang Bangun Alat Pendeteksi Kecelakaan Mobil Menggunakan Sensor Akselerometer dan Sensor 801s Vibration," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, 2018.
[5] Y. Ryandi and N. L. P. S. E. Setyarini, "Evaluasi Ruas Jalan Gatot Subroto Menggunakan Metode Irap Untuk Mencapai Star Rating 4 Dan 5," *JMTS J. Mitra Tek. Sipil*, vol. 4, no. 3, p. 777, 2021, doi: 10.24912/jmts.v0i0.12649.
[6] N. Kattukkaran, A. George, and T. P. M. Haridas, "Intelligent accident detection and alert system for emergency medical assistance," 2017, doi: 10.1109/ICCCI.2017.8117791.
[7] Badan Pusat Statistik, *Statistik Kriminal 2019*. 2019.
[8] M. Sammarco and M. Detyniecki, "Crashzam: Sound-based car crash detection," 2018, doi: 10.5220/0006629200270035.
[9] R. C. Gatto and C. H. Q. Forster, "Audio-Based Machine Learning Model for Traffic Congestion Detection," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–8, 2020, doi: 10.1109/tits.2020.3003111.
[10] R. K. Kodali and S. Sahu, "MQTT based vehicle accident detection and alert system," *Proc. 2017 3rd Int. Conf. Appl. Theor. Comput. Commun. Technol. iCATccT 2017*, pp. 186–189, 2018, doi: 10.1109/ICATCCT.2017.8389130.
[11] Y. Arslan and H. Canbolat, "Performance of deep neural networks in audio surveillance," 2018, doi: 10.1109/CEIT.2018.8751822.
[12] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE Trans. Intell. Transp. Syst.*, 2016, doi: 10.1109/TITS.2015.2470216.
[13] R. Leiba, F. Ollivier, R. Marchiano, N. Misdariis, J. Marchal, and P. Challande, "Acoustical classification of the urban road traffic with large arrays of microphones," *Acta Acust. united with Acust.*, vol. 105, no. 6, 2019, doi: 10.3813/AAA.919387.
[14] S. Chandrakala and S. L. Jayalakshmi, "Environmental Audio Scene and Sound Event Recognition for Autonomous Surveillance," *ACM Comput. Surv.*, vol. 52, no. 3, pp. 1–34, 2020, doi: 10.1145/3322240.
[15] N. Almaadeed, M. Asim, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "Automatic detection and classification of audio events for road surveillance applications," *Sensors (Switzerland)*, 2018, doi: 10.3390/s18061858.
[16] H. H. Pour *et al.*, "A Machine Learning Framework for Automated Accident Detection Based on Multimodal Sensors in Cars," *Sensors*, vol. 22, no. 10, pp. 1–21, 2022, doi: 10.3390/s22103634.
[17] A. Bonyar *et al.*, "A review on current eCall systems for autonomous car accident detection," *Proc. Int. Spring Semin. Electron. Technol.*, 2017, doi: 10.1109/ISSE.2017.8000985.
[18] M. Muthuvel, M. Marimuthu, S. Nivetha, and K. Sirushti, "Accident Detection and Reporting System using Internet of Things Biometrics View project Internet of Things View project Accident Detection and Reporting System using Internet of Things," *Res. J. Sci. Eng. Syst.*, vol. 3, no. 2, pp. 121–130, 2018, [Online]. Available: www.rjsces.com.
[19] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," 2014, doi: 10.1145/2647868.2655045.
[20] K. J. Piczak, "ESC: Dataset for environmental sound classification," *MM 2015 - Proc. 2015 ACM Multimed. Conf.*, pp. 1015–1018, Oct. 2015, doi: 10.1145/2733373.2806390.

[21] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "The NIGENS General Sound Events Database," Feb. 2019, doi: 10.5281/zenodo.2535878.
[22] Q. Zhou *et al.*, "Cough Recognition Based on Mel-Spectrogram and Convolutional Neural Network," *Front. Robot. AI*, vol. 8, no. May, pp. 1–7, 2021, doi: 10.3389/frobt.2021.580080.
[23] I. Tsiamas, G. I. Gállego, J. A. R. Fonollosa, and M. R. Costa-Jussà, "SHAS: Approaching optimal Segmentation for End-to-End Speech Translation," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2022-September, pp. 106–110, 2022, doi: 10.21437/Interspeech.2022-59.
[24] S. Sahoo, P. Kumar, B. Raman, and P. P. Roy, "A Segment Level Approach to Speech Emotion Recognition Using Transfer Learning," 2020, doi: 10.1007/978-3-030-41299-9_34.
[25] H. Yakura and J. Sakuma, "Robust audio adversarial example for a physical attack," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2019-August, pp. 5334–5341, 2019, doi: 10.24963/ijcai.2019/741.
[26] J. B. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," *IEEE Trans. Acoust.*, vol. 25, no. 3, pp. 235–238, 1977, doi: 10.1109/TASSP.1977.1162950.
[27] G. Rybak and K. Strzecha, "Short-time fourier transform based on metaprogramming and the stockham optimization method," *Sensors*, vol. 21, no. 12, 2021, doi: 10.3390/s21124123.
[28] T. Tran and J. Lundgren, "Drill fault diagnosis based on the scalogram and MEL spectrogram of sound signals using artificial intelligence," *IEEE Access*, vol. 8, pp. 203655–203666, 2020, doi: 10.1109/ACCESS.2020.3036769.
[29] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.
[30] L. H. Iksan, M. I. Awal, R. Z. Fhamy, A. A. Pratama, D. K. Basuki, and S. Sukaridhoto, "Implementation of Cloud Based Action Recognition Backend Platform," *AIMS 2021 - Int. Conf. Artif. Intell. Mechatronics Syst.*, 2021, doi: 10.1109/AIMS52415.2021.9466068.