Advances in Artificial Intelligence and Machine Learning; Research 4 (2) 2189-2200 Received 12-02-2024; Accepted 12-04-2024; Published 19-04-2024

Acoustic Spectral Analysis for Emergency Vehicle Detection: Symbolic and CNN Approaches

Alberto Pacheco alberto.pg@chihuahua.tecnm.mx
TecNM campus Chihuahua, Chihuahua, Chih., Mexico 31310

Mariano Rivera mrivera@cimat.mx
Centro de Investigacion en Matematicas, Guanajuato, Gto., Mexico 36023

Raymundo Torres r.torreset@gmail.com
TecNM campus Chihuahua, Chihuahua, Chih., Mexico 31310

Corresponding Author: Alberto Pacheco

Abstract

During emergencies, ambulances on city streets face delays due to traffic obstacles. This paper addresses two efficient emergency vehicle detection (EVD) methods for restricted hardware implementation considering noisy conditions: a symbolic processing-based algorithm and a convolutional neural network (CNN) model, both of which utilize Mel spectrogram representations of Hi-Lo siren audio records. The symbolic method employs regular expressions and acceptance criteria to process text-pattern features extracted from spectrograms, offering a self-explanatory, easily tunable, and resource-efficient solution suitable for low-cost hardware platforms. On the other hand, the CNN model directly processes spectrogram representations, leveraging spatial correlation for classification with a streamlined architecture consisting of very few layers. The experimental results demonstrate that both approaches achieve high accuracy (97-98%) in classifying Hi-Lo sirens, with the CNN model exhibiting slightly better performance. Challenges such as signal noise and harmonics are addressed through iterative algorithms and signal reconstruction considerations. Future directions include identifying additional siren effects and conducting performance measurements on constrained hardware devices. Overall, this study presents viable EVD solutions suitable for real-time implementation and underscores the importance of adaptable and explainable AI methods in enhancing road safety.

Keywords: Emergency vehicle detection (EVD), Signal symbolization, Spectral signal processing, Convolutional neural network, Sound classification.

Citation: A. Pacheco, et al. Acoustic Spectral Analysis for Emergency Vehicle Detection: Symbolic and CNN Approaches. Advances in Artificial Intelligence and Machine Learning, 2024;4(2):125.

1. INTRODUCTION

Sirens, such as police cars, fire trucks, or ambulances, are devices that emit alerts or warning sounds as part of an emergency vehicle (EV) operation. In emergency situations, these sirens draw attention and warn nearby drivers and pedestrians about the presence of an EV that requires immediate right-of-way clearance. The latest advancements in automotive technology have increased the availability of advanced driver assistance systems (ADASs) as part of modern cognitive cars [1]. Consequently, there is a compelling opportunity to explore the development of an emergency vehicle detection (EVD) module to aid drivers by automatically adjusting the volume of their music player, activating hazard lights, and/or alerting them to cooperate more effectively in promptly clearing the path for emergency vehicles . Such innovation facilitates the swift and safe passage of emergency vehicles to their destinations. Various EV siren effects convey distinct codes and levels of urgency, contributing to effective emergency signaling. Among the most prevalent siren effects are Hi-Lo, *yelp*, and *wail*. The choice of siren sound depends on the type of emergency vehicle and protocol, local regulations, and the intended purpose of the alarm. Two-tone or Hi-Lo sirens are commonly used in emergency vehicles worldwide, but no universal regulation exists. ISO 7010/7731 [2], established an emergency frequency range recommendation from 500 Hz to 2,500 Hz. Other specifications and standards for emergency vehicle sirens may vary between countries and regions, reflecting diverse regulatory frameworks and safety considerations, such as the American SAE J1849, European Norm EN 4713-2, Canadian CSA-D250, Japanese MLIT regulations, and Mexican NOM-034-SSA3-2013.

This paper introduces two emergency vehicle detection (EVD) methods that were designed to meet key requirements and challenges: a) providing an interpretable and portable model; b) offering a lower memory footprint than CNN models; and c) enabling stand-alone on-device execution. These requirements are crucial for supporting constrained, low-cost hardware–embedded devices, *i.e.* edge computing. Although benchmark results have been obtained, they are currently limited solely to classification accuracy metrics and do not include any performance metrics.

2. RELATED WORKS

A feature extraction algorithm was presented in [3], which detects emergency vehicles through the identification of their dominant tones (pitch detection) by applying the module difference function (MDF) technique and searching for peaks of each tone to detect Hi-Lo siren patterns over time. Several deep neural network EVD models have been proposed [4–10], including the fully connected network model (FCN), convolutional network (CNN), and recurrent network (RNN), which directly process raw audio signals; others, such as [11], which use Mel frequency cepstral coefficients (MFCCs). Spectral analysis using fast Fourier transform (FFT) was utilized in [12], to identify sirens and attempt to estimate the distance to an EV. In [13], a comparison was made between the CNN models AlexNet and GoogleNet using three different representations for input audio: a frequency spectrogram, an MFCC, and cross-recurrence plots (CRPs). Additionally, time series analysis involving cross-correlation of two signal sequences can be employed with moderate success in finding their similarity by measuring their relative displacement [14]. Finally, symbolic analysis of time series to identify features of interest was applied in [15], to provide an explainable model

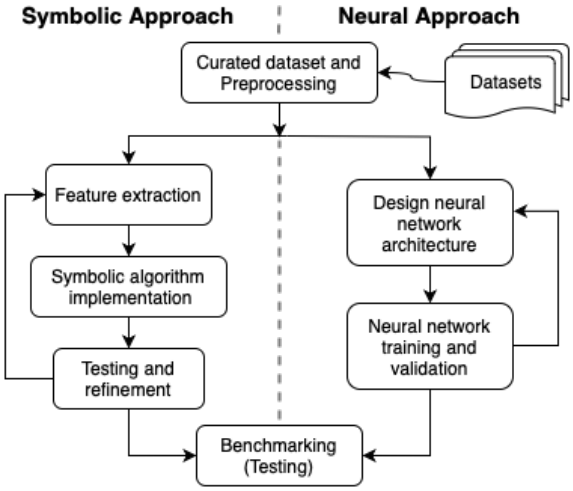

Figure 1: Method of developing the EVD symbolic and neural models.

[16], using probabilistic finite state automata and through the SAX technique [17, 18], in which signal segments are mapped as symbols based on their normal distributions.

3. METHOD

This section introduces the proposal for detecting two–tone sirens. The method, as shown in Figure 1, starts with an introduction to the dataset preparation and preprocessing, followed by the presentation of both classification (detection) methods and its comparative study.

3.1 Dataset Preparation and Preprocessing

The dataset was compiled by selecting audio from: a) ESC-50, a dataset with 50 different urban sound classes [19]; b) the UrbanSound8k (US8K) dataset [20]; and c) the Large Scale Audio Dataset (LSAD) [21]. The curated audio collection only includes Hi-Lo sirens because the detectors are designed for this specific siren type. Based on the literature analyzed in [4–15], empirical testing, and the feasibility of feature analysis, it has been determined that using a 3-second audio duration with a uniform sampling rate and resolution are suitable audio parameter settings. Therefore, all the audio files were preprocessed to ensure that these attributes were met. A total of 381 Hi-Lo siren audios and 840 urban sounds met these criteria and were selected for the training, validation, and testing of both the symbolic and the machine learning audio classifiers (TABLE 1).

Table 1: Training, validation and testing datasets.

Classes	Train	Validation	Test	Total
Hi-Lo siren	300	29	52	381
Urban sound	510	50	280	840
Subtotal	814	79	332	1,221

The sound dataset undergoes preprocessing by converting the records to the lowest common dataset audio settings, which is a single-channel format with a 44.1 kHz sampling rate, 16-bit resolution, uncompressed raw audio format, and a duration of 3 seconds that corresponds to signals $x(n)$ for $n = 0, 1, \ldots, N-1$; with $N = 15,360$. Subsequently, the audio data representation was transformed into spectrograms, transitioning from the time domain, denoted as t, to the frequency domain, represented by ω. A collection of subsignal spectrograms was obtained, each covering a time window of size $T = 1024$. A time shift of $P = 320$ samples was applied between consecutive subsignals to obtain a log-Mel spectrogram, defined as the magnitude of the Discrete Fourier Transform (DFT) of each subsignal resulting from shifting the window function over the signal [22]:

$$M(\omega, t) = \log \|\mathrm{DFT}\{x(n) \odot h(tP - n)\}\|, \tag{1}$$

where $\odot$ denotes the pointwise product; $h(n)$ is a smooth window function with a support interval (the region where it takes nonzero values) of length T and with $h(0) = h(T-1) \to 0$. The following Hann window function was selected:

$$h(n) = \begin{cases} \sin^2\left(\frac{\pi n}{T}\right) & 0 \le n < T, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Since human perception of frequencies is not linear (pitch perception), it is preferred to change from the frequency scale from Hertz ω to Mels f with:

$$f = 1125 \log[1 + \omega/700]. \tag{3}$$

For the experiments, according to the Nyquist rate $> 2(f_a + f_{R_5}^{hi})$, a mel-sampling rate of 5120 Hz and 64 mel-bands were chosen for the Librosa routine [23]. FIGURE 2 depicts the log-Mel spectrogram of two Hi-Lo siren audios, where the y-axis corresponds to the windowed DFT log-magnitude (dB). Note that the temporal maxima of the spectrogram exhibit a clear correspondence with the high and low tones of the sirens. The spectrogram on the right depicts harmonics that may correspond to square-shaped waves. This visual analysis of the data led us to implement two detection methods, one based on symbolic processing of the spectrogram response and another based on a convolutional neural network (CNN) for classifying spectrogram images, as detailed below.

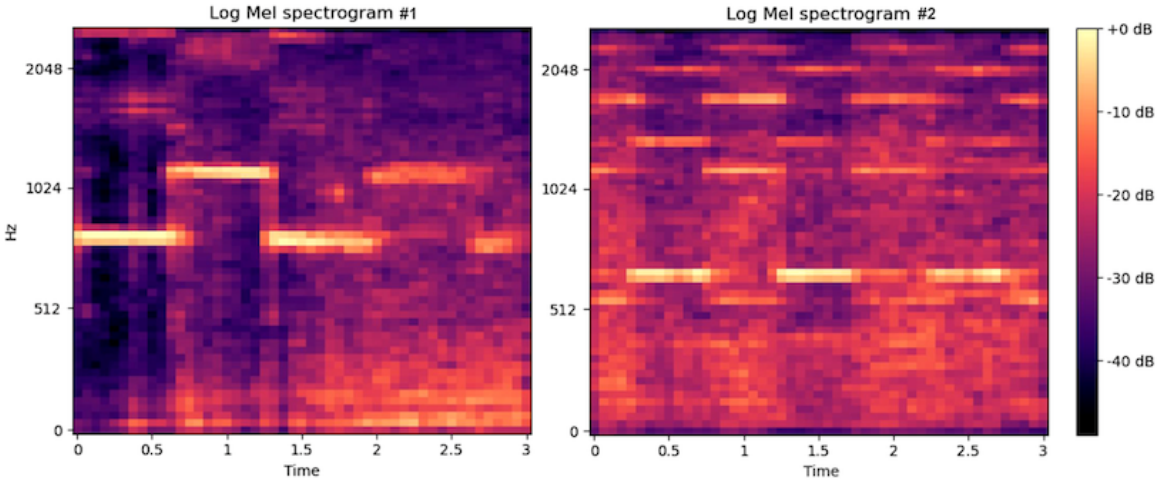

Figure 2: Hi-Lo siren log-Mel spectrograms: a) faded signal; b) high-harmonic spectrum.

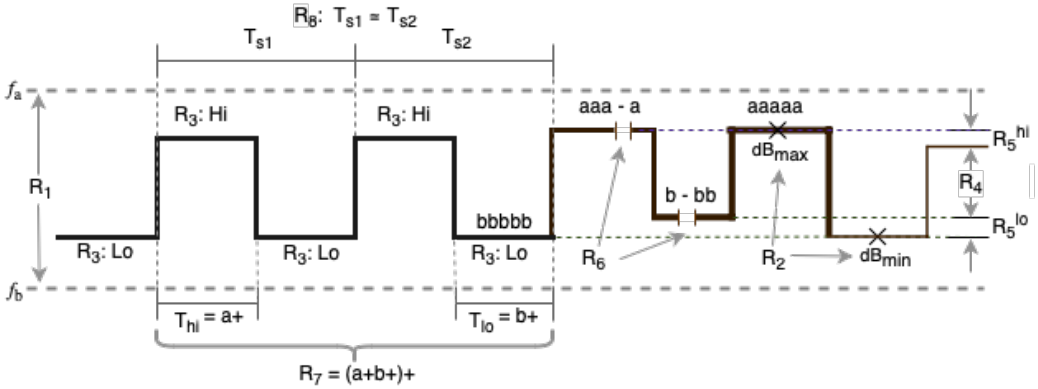

Figure 3: Hi-Lo siren audio features R_i.

3.2 Symbolic Processing-Based Method

Feature Analysis. Through iterative algorithm refinement and empirical analysis, as summarized in FIGURE 3, and further detailed in [24], the following siren signal features were defined:

- R_0: minimum zero RMS rate, fast signal-to-noise ratio measure (SNR)
- R_1: frequency range $f_a - f_b$
- R_2: dominant frequency threshold
- R_3: presence of Hi-Lo tones
- R_4: tonal gap $f_a - f_b$
- R_5: variability in Hi-Lo tone frequency $f^{hi}_{R_5}$, $f^{lo}_{R_5}$
- R_6: maximum discontinuity for Hi-Lo tones
- R_7: minimum expected periodicity
- R_8: expected pattern of periodic regularity

Symbolic spectral analysis of the EVD algorithm. The signal features R_i can be understood as a set of features computed from the input signal. Then, we can define a Boolean subcriterion C_i associated with each feature R_i as follows: $C_i = \{\text{False} : if R_i = \emptyset;\ \text{True} : \text{otherwise}\}$ to define a simpler and more straightforward model using only the following single detection criterion D:

$$D = \bigwedge_{i=0}^{8} C_i. \tag{4}$$

This approach offers the following advantages: high efficiency (stops upon any subcriterion failure) and robust parallelism (simultaneous subcriterion execution). Unfortunately, this model is impractical due to signal noise and harmonics, and a more complex iterative EVD algorithm, which is based on obtaining and processing a sequence of symbolic feature candidates, is needed, as represented in pseudo-code below:

1. Read audio signal and computation of the log-Mel spectrogram.
2. Test C_0, if it fails, exits (unable to detect, SNR below R_0).
3. For C_1, a $f_a - f_b$ bandpass filter is applied according to R_1.
4. $cEncoder_1$: Binarized filtered spectrogram based on C_2 verifying R_2.

5. Find the dominant frequency candidates for each spectrogram frame.
6. *candidates* = list of the dominant frequencies.
7. End and return false if there are no Hi-Lo tones (C_3).
8. Detection best candidates loop (up to four harmonic-finder cycles):
   8.1. Select Hi-Lo tones with higher counts.
   8.2. Satisfy criterion C_4; if it fails, the most frequent tone and cycle are removed.
   8.3. $cEncoder_2$: encodes the respecting criterion C_5.
   8.4. Reconstruct signal using C_4 and C_6; signal unchanged if the previous criteria fail.
   8.5. If the C_5 and C_7 criteria fail, the higher harmonics are removed, and the cycle is repeated.
   8.6. $cEncoder_3$: C_6, C_7 and C_8 are applied, if they fail, cycle. If met, stop returning true.
9. After the end of the cycle, the algorithm returns false.

The most meaningful signal transformations are explained below using some examples. After processing the audio to derive the log-Mel spectrogram (FIGURE 2), given the highest (f_a) and lowest (f_b) tone frequencies, the $f_a - f_b$ frequency band is filtered using the expression:

$$C_1 = \{M(f,t) \mid f_a \leq f \leq f_b\} \tag{5}$$

The first signal symbolization, via $cEncoder_1$, transforms the audio intensity levels (dB) inside the frequency range $f_a - f_b$ (Hz) into binary digits (0,1), according to:

$$cEncoder_1 = \begin{cases} 1, & \text{if } |f| > \text{dB}_{min} \text{ where } f \in C_1 \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The next symbolization transformation is performed by using $cEncoder_2$, where the following symbol Σ alphabet is considered:

$$\sigma = \{\texttt{a},\texttt{b},\texttt{-}\} \tag{7}$$

The terminal symbol `'a'` represents a high tone, the symbol `'b'` represents a low tone, and the symbol `'-'` encompasses any value outside each tonal range of interest (R_5), as shown in FIGURE 3, where $cEncoder_2$ applies the following expression:

$$cEncoder_2 = \begin{cases} a, & \text{if } f \in f_{R_5}^{hi} \\ b, & \text{if } f \in f_{R_5}^{lo} \\ -, & \text{otherwise.} \end{cases} \tag{8}$$

After identifying the histogram with dominant frequencies (R_2), selecting candidates for Hi-Lo tones (R_3), and confirming that these tones meet R_4, a sequence of tones is obtained (*candidates*), e.g. `'aabbaabb'` sequence may indicate the presence of a Hi-Lo siren consisting of two full cycles. However, due to sampling errors such as noise, latency, and conversion issues, it may be necessary at times to reconstruct the signal considering the following discontinuity cases (R_6):

$$C_6 = \texttt{a-a|a--a|b-b|b--b} \tag{9}$$

When reconstructing the signal (step 8.4), for example, a given tone discontinuity such as `'aaa-a'`, or `'b-bb'`, as shown in FIGURE 3, may be reconstructed as `'aaaaabbbb'`, that is, a full siren cycle; thus, the ability to detect a Hi-Lo siren requires verifying the minimum periodicity (R_7) established by the following regular expression:

$$C_7 = (\mathtt{a}^+\mathtt{b}^+)^+. \tag{10}$$

However, this expression does not guarantee compliance with the periodic regularity criterion (R_8); that is, both tones must be adjusted to a certain periodic range ($T_{s1} \cong T_{s2}$ in FIGURE 3) since there may be slight variations due to noise, precision, temperature or incomplete signal cycles (clipping). To carry out this analysis, it is necessary to symbolize the signal again using $cEncoder_3$ according to the following format:

$$cEncoder_3 = \mathtt{[ab-]\backslash d}^+. \tag{11}$$

For example, at the beginning of step 8.6, given the sequence `'aaabbaabbaab'`, the output of $cEncoder_3$'s will be `'a3b2a4b3a4b3'`. After this step, the tone occurrences are sorted, that is, $L_{hi} = [4, 4, 3]$ and $L_{lo} = [3, 3, 2]$. Then the median μ of each sorted list is calculated using:

$$L'_i = [x : x = \mu_i, x \in L_i] \text{ and } i \in \{hi, lo\} \tag{12}$$

In the example, this corresponds to $L'_{hi} = [4, 4]$ and $L'_{lo} = [3, 3]$. Finally, the acceptance criterion for periodic regularity R_8, from experimental testing, requires at least two cycles, i.e. `a4b3a4b3`:

$$C_8 = (\sharp L'_{lo} \geq 2) \wedge (\sharp L'_{hi} \geq 2). \tag{13}$$

where $\sharp$ denotes length or cardinality, L'_{hi} are symbol `a` occurrences, and L'_{lo} symbol `b` occurrences.

In summary, after all the (C_i) criteria are satisfied, the detection algorithm considers that identifying a Hi-Lo siren was successful; otherwise, the identification algorithm stops being executed at any given point at which a criterion is not met.

Adjusting and validating the symbolic processing algorithm. For this symbolic EVD algorithm, several parameter adjustments were performed after a series of successive refinements over the corresponding benchmark test siren audio and urban sound datasets (TABLE 1). The following operating parameters were established [24]: $R_0 : 40\%; R_1 : f_a = 1,500Hz, f_b = 700Hz; R_2 : dB_{min} = 20dB; R_4 : gap = 122Hz; f^{hi}_{R_5} = f^{lo}_{R_5} = 31Hz; R_6 : 2cycles \ @ \ dur3secs; R_7 : 2frames.$

3.3 Convolutional Neural Network for Siren Detection

The proposed approach uses a convolutional neural network (CNN) as the classifier model. The network must be simple so that it can be implemented on limited hardware. From the visual examination of the spectrogram, where a periodic pattern associated with the Hi-Lo siren was revealed, it was assumed that a two-dimensional convolutional network could effectively identify Hi-Lo sirens.

The convolutional stage of the network is organized as a sequence of three convolutional blocks designed to process an input spectrogram resized to dimensions of (49×49). The first block consists of a *two-dimensional convolutional* layer employing ReLU activation, no padding, and a stride=1. Subsequently, a 2×2 *MaxPooling* layer is applied. To mitigate overfitting, a *Dropout* layer with a 25% probability of masking activations is added. Following the convolutional stage, a *Flatten*

layer is introduced, followed by two *Dense* layers serving as decision layers, with an intermediate *Dropout* layer ($p = 0.25$). The architecture of our model is summarized in TABLE 2.

Table 2: EVD convolutional neural network architecture with 20,385 parameters.

#	Layer	Type	Description	Parameters	Output Size
1	Conv1	Convolution2D	64 filters of 3x3 (ReLU)	640	49x49x1
2	Pool1	MaxPooling	2x2, stride 2	0	23x23x64
3	Drop1	Dropout	p=0.25	0	23x23x64
4	Conv2	Convolution2D	24 filters of 3x3 (ReLU)	13,848	21x21x24
5	Pool2	Max Pooling	2x2, stride 2	0	10x10x24
6	Drop2	Dropout	p=0.25	0	10x10x24
7	Conv3	Convolution2D	8 filters of 3x3 (ReLU)	1,736	8x8x8
8	Pool3	Max Pooling	2x2, stride 2	0	4x4x8
9	Drop3	Dropout	p=0.25	0	4x4x8
10	Flat1	Flatten	—	0	128
11	FC1	Dense	32 neurons (ReLU)	4,128	32
12	Drop4	Dropout	p=0.25	0	32
13	FC2	Dense	1 neuron (Sigmoid)	33	1

As previously mentioned, all the audio clips were converted to log-Mel spectrograms with dimensions $(49, 49)$ to train and validate the proposed model. The CNN model training compromises 329 Hi-Lo sirens and 560 urban noise samples, as outlined in TABLE 1, using a batch size of 16 samples, the ADAM optimizer algorithm and *binary cross–entropy* (BCE) as the loss function. The model achieved satisfactory accuracy and a loss after 20 epochs.

4. RESULTS

During the benchmark analysis, both models were tested using identical test datasets, with a) the predicted positive class comprising 52 Hi-Lo siren audio recordings and b) the false class comprising 280 urban sound recordings. For classification benchmarking purposes, the prediction categories were as follows: nonsiren sounds were designed as true negatives (TNs), and false positives (FPs) were those inaccurately classified as siren sounds. Similarly, each instance where a Hi-Lo siren audio recording was correctly identified as a siren, it was classified as a true positive (TP). Conversely, instances where such recordings were erroneously classified as non-siren sounds were categorized as false negatives (FN). Once the respective confusion matrices were obtained (TABLE 4), the following predictive performance metrics were selected: overall accuracy, sensitivity (recall), specificity, precision, and F1 score.

The confusion matrices for the neural and symbolic models are shown in TABLE 3. The predictive performance metrics are summarized in TABLE 4, which shows that the neural network model achieved an accuracy of 98%, whereas the symbolic approach model achieved an accuracy of 97%. However, the machine learning model had one false negative (FN), meaning that it correctly detected 51 siren sounds (TPs). On the other hand, the symbolic model failed to detect two of the 52 siren sounds (FNs) and correctly identified the remaining 50 sounds. As a result, the CNN model exhibited a sensitivity of 98%, whereas the symbolic approach model had a sensitivity of 96%.

Table 3: Neural/symbolic confusion matrices.

Dataset classes (actual)	Siren Neural/Symbolic (predicted)	No-Siren Neural/Symbolic (predicted)
52 Sirens (positive class)	51 / 50 (TP)	1 / 2 (FN)
280 No-Siren (negative class)	3 / 8 (FP)	277 / 272 (TN)

Table 4: Predictive performance metrics.

Metric	Neural	Symbolic
Accuracy	0.9879	0.9698
Sensitivity	0.9807	0.9615
Specificity	0.9892	0.9714
Precision	0.9444	0.8620
F1 Score	0.9622	0.9090

In terms of specificity, which refers to the proportion of nonsiren audios correctly classified out of all the negative predictions it made, the CNN model performed 0.9892 better than the 0.9714 for the symbolic model. The CNN model incorrectly detected 3 nonsiren sounds as sirens, while the symbolic model generated 8 false positives. Similarly, the precision, which refers to the proportion of the Hi-Lo siren audio correctly classified as siren audio out of all its positive predictions, was greater in the CNN model (0.9444) than in the symbolic model (0.862). Finally, the F1 score was 0.9622 for the neural model and 0.909 for the symbolic model. In summary, despite exhibiting a slightly lower overall classification accuracy, the symbolic model can serve as a viable auto-explainable algorithm that is well-suited for low-cost edge computing devices.

5. DISCUSSION

The proposed EVD algorithm uses a symbolic representation at a higher level, which differs from techniques based on Fourier transforms, *wavelets*, *eigenwaves*, and polynomial models. This symbolization technique provides an adaptive representation of signal features as text patterns, enabling regular expressions for pattern detection, unlike other data science methods, such as clustering, classification, indexing, summarizing, trees, and anomaly detection [17]. The proposed symbolic algorithm is self-explanatory and highly efficient, unlike black-box neural network models [4–11], which always execute the same vast number of operations for each inference, unlike the symbolic algorithm, which is halted if a specific criterion fails. Additionally, the algorithm provides a more extensive feature set than any previous related works [3, 12, 13, 15, 25–27], it is relatively easy to adjust and extend, and its regular expressions are highly portable to many different platforms [28]. Some challenging classification scenarios can also be discussed. If the siren signal weakens, due to the Doppler effects of vehicles approaching and moving away, in the presence of elevated background noise (first image in 2), then the algorithm's signal reconstruction module demonstrated significant efficacy and accuracy. However, the criterion R_0 will preclude any detection analysis for extremely loud noise, i.e. very low SNR. In the second case shown in FIGURE 2, the presence of siren harmonics posed challenges in identifying the fundamental frequencies of the siren tones. Therefore, an iterative harmonic search was conducted specifically for this case (step 8). The quantity and quality of dataset samples for urban sounds and sirens significantly influenced the CNN model training. Therefore, it is advisable for future endeavors to incorporate a larger volume of higher-quality Hi-Lo siren sounds, considering a well-defined audio recording protocol. Multiple iterations were conducted throughout each symbolic algorithm improvement across a dozen versions to fine-tune features, criteria, and parameters, e.g. the first four versions, including $R_1 - R_3$, have unacceptable precision ($prec$); ver. 7 $(R_1 - R_4)$ $prec 0.5$; ver. 9 $(R_1 - R_5)$ $prec = 0.64$. Due to

its symbolic criteria, this model is more adaptable and comprehensible than the alternative neural network model. This adaptability enables parametric adjustments to enhance EVD capabilities, enabling customization for specific regions. Incorporating the detection of additional siren effects, such as *yelp*, is a relatively straightforward task that involves collecting enough *yelp* audio records, adding a new class, and retraining the CNN model. However, a completely new feature engineering process will be required to characterize a *yelp* signal and develop a new detection algorithm based on signal symbolization. Another significant avenue and challenge for EVD is to explore a neuro-symbolic approach [29], and explainable AI methods, especially for the feature extraction stage [16]. Additionally, there is a need to conduct performance measurements of each model under constrained hardware devices to validate its feasibility and efficiency .

6. CONCLUSIONS

Two Hi-Lo siren detection methods using the log-Mel spectrogram representation of audio records have been presented. The first method is based on a symbolic procedure that analyzes spectrograms, extracting features represented as text patterns and processing them with regular expressions and a set of acceptance criteria. This method is self–explanatory, easy to tune, and requires minimal computing resources. Thus, a lower–cost hardware platform can execute this symbolic-based model to identify Hi-Lo sirens in real–time. The second proposed model is a convolutional neural network (CNN) that directly processes the spectrogram representation of the audio. This approach is directly trained using the spectrogram by taking advantage of the spatial correlation of the representation. The proposed CNN has fewer parameters and layers (20K parameters, 80 Kb, 99% accuracy) than the state-of-the-art reported CNNs, such as MLNet and SirenNet (7-27M parameters, 400-700 Kb, 96-98% accuracy) [4]. The performance of the methods was demonstrated through experiments, and it was confirmed that both approaches are suitable for EVD.

7. Acknowledgments

Work supported in part by TecNM, Mexico, Grant 17458.23-P (AP); Conahcyt, Mexico, Grant CB-A1-43858 (MR); and Conachcyt MEng Scolarship Grant (RT).

References

[1] Li L, Wen D, Zheng N, Shen L. Cognitive Cars: A New Frontier for Adas Research. IEEE Trans Intell Transp Syst. 2012;13:395-407.

[2] International Organization for Standardization, “Auditory danger signals,” ISO-7731, 2013.

[3] Meucci F, Pierucci L, Del Re E, Lastrucci L, Desii P, et al. A Real Time Siren Detector to Improve Safety of Guide in Traffic Environment. In: the 16th Europ. Signal. Proceedings of the conference. 2008:1-5.

[4] Tran V, Tsai W. Acoustic-Based Emergency Vehicle Detection Using Convolutional Neural Networks. IEEE Access. 2020;8:75702-75713.

[5] Huzaifah M. Comparison of Time-Frequency Representations for Environmental Sound Classification Using Convolutional Neural Networks. 2017. Arxiv Preprint: https://arxiv.org/pdf/1706.07156.pdf 2017

[6] Palanisamy K, Singhania D, Yao A. Rethinking CNN Models for Audio Classification. 2021. Arxiv Preprint: https://arxiv.org/abs/2007.11154

[7] Nanni L, Maguolo G, Brahnam S, Paci M. An Ensemble of Convolutional Neural Networks for Audio Classification. Appl Sci. 2021;11:5796.

[8] Zinemanas P, Rocamora M, Miron M, Font F, Serra X, et al. An Interpretable Deep Learning Model for Automatic Sound Classification. Electronics. 2021;10:850.

[9] Tsalera E, Papadakis A, Samarakou M. Comparison of Pre-trained Cnns for Audio Classification Using Transfer Learning. J Sens Actuator Netw. 2021;10:72.

[10] Shin S, Kim J, Yu Y, Lee S, Lee K, et al. Self-Supervised Transfer Learning From Natural Images for Sound Classification. Appl Sci. 2021;11:3043.

[11] Mittal U, Chawla P. Acoustic Based Emergency Vehicle Detection Using Ensemble of Deep Learning Models. Procedia Comput Sci. 2023;218:227-234.

[12] Ebizuka Y, Kato S, Itami M. Detecting Approach of Emergency Vehicles Using Siren Sound Processing. In: the IEEE Intell Transp Syst Conf. (ITSC). 2019:4431-4436.

[13] Boddapati V, Petef A, Rasmusson J, Lundberg L. Classifying Environmental Sounds Using Image Recognition Networks. Procedia Comput Sci. 2017;112:2048-2056.

[14] Vio R, Wamsteker W. Limits of the Cross Correlation Function in the Analysis of Short Time Series. Astronomical Society of the Pacific. 2001;113:86-97.

[15] Li Y, Ray A. Unsupervised Symbolization of Signal Time Series for Extraction of the Embedded Information. Entropy. 2017;19:148.

[16] Belle V, Papantonis I. Principles and Practice of Explainable Machine Learning. Front Big Data. 2021;4:688969.

[17] Lin J, Keogh E, Wei L, Lonardi S. Experiencing Sax: A Novel Symbolic Representation of Time Series. Data Min Knowl Disc. 2007;15:107-144.

[18] Lin J, Keogh E, Lonardi S, Chiu B. A Symbolic Representation of Time Series, With Implications for Streaming Algorithms. In: Proceedings of the ACM SIGMOD DMKD. 2003:2-11.

[19] Piczak KJ. Esc: Dataset for Environmental Sound Classification. In: Proceedings of the 23rd ACM international conference multimedia. 2015:1015-1018.

[20] Salamon J, Jacoby C, Bello JP. A Dataset and Taxonomy for Urban Sound Research. In: Proceedings of the 22nd ACM international conference multimedia. 2014:1041-1044.

[21] Asif M, Usaid M, Rashid M, Rajab T, Hussain S, et al. Large-Scale Audio Dataset for Emergency Vehicle Sirens and Road Noises. Sci Data. 2022;9:599.

[22] Meng H, Yan T, Yuan F, Wei H. Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network. IEEE Access. 2019;7:125868- 125881.

[23] McFee B, Raffel C, Liang D, Ellis D, McVicar M, et al. Librosa: Audio and Music Signal Analysis in Python. Proceedings of the Python in Science Conference. Proceedings of the 14th python sci conf. 2015:18-24.

[24] SOzdemir S. Feature Engineering Bookcamp. Simon Schuster. 2022.

[25] Sherratt RS, Guy CG, Townsend DM. Cancellation of Siren Noise From Two-Way Voice Communication Inside Emergency Vehicles. Comput Control Eng J. 2002;13:5-10.

[26] Schroder J, Goetze S, Grützmacher V, Anemüller J, et al. Automatic Acoustic Siren Detection in Traffic Noise by Part-Based Models. In: the ICASSP international conference on Acoustics, speech, and signal processing. 2013:493-497.

[27] Haitsma J, Kalker T. A Highly Robust Audio Fingerprinting System. In: the Int. Society for Music Info Retrieval conference. 2002.

[28] Davis J, Michael IV LG, Coghlan CA, Servant F, Lee D. et al. Why Aren't Regular Expressions a Lingua Franca? An Empirical Study on the Reuse and Portability of Regular Expressions. In: Proceedings of the ACM ESEC/FSE. 2019:443-454.

[29] Hitzler P, Sarker M. Neuro-Symbolic Artificial Intelligence: The State of the Art. In: Frontiers in ai and applications. 2021;342.