A NOVEL DEEP LEARNING MODELS FOR EMERGENCY VEHICLE DETECTION

HAIQING KI LIU

ABSTRACT

The essence of sound events, manifested in their temporal and spectral structure within the time-frequency domain, forms the core of an evolving field focused on analyzing and categorizing acoustic environments through recorded sound. By employing convolutional layers, this study efficiently extracts high-level features that remain invariant to shifts in this domain. Our investigation centers on the detection of emergency vehicles. We explore three distinct deep neural network (DNN) architectures – dense layer, Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN) – with varying configurations and parameters. Subsequently, we devise an ensemble model by meticulously selecting optimal models through experimental testing across various configurations, coupled with hyper-parameter tuning. This ensemble model achieves the pinnacle accuracy of 98.7%, surpassing the standalone RNN model, which achieves 94.5% accuracy. Additionally, we conduct a thorough performance analysis, comparing the deep learning models with a spectrum of machine learning algorithms, including Perceptron, Support Vector Machine (SVM), and decision trees.

Index Terms *Audio recognition; CNN DNN; emergency vehicle detection; MFCC; RNN; siren sound*

Reference *to this paper should be made as follows**:** HAIQING KI LIU, (2023), "A Novel Deep Learning Models For Emergency Vehicle Detection" Int. J. Electronics Engineering and Applications, Vol. 11, Issue II, pp. 1-11.*

Biographical notes:

HAIQING KI LIU *is currently associate professor at the School of Control and Computer Engineering, South China, Electric Power University.*

Published by: International Associations of Professionals and Technical Teachers (IAPATT)

1. INTRODUCTION

Emergency Vehicle Detection (EVD) systems are pivotal for ensuring prompt emergency responses amidst traffic congestion and roadway crises. These systems rely on identifying emergency vehicles through their unique siren sounds, which are categorized into signals for ambulances, fire trucks, and police cars. Each signal type adheres to specific regulations outlined by the International Organization for Standardization (ISO).

This study proposes the utilization of an ensemble of deep learning models for audio recognition, leveraging features extracted from both the time and frequency domains of siren sounds. By integrating EVD systems into intelligent transportation infrastructure, traffic controllers can prioritize emergency vehicle routes by dynamically adjusting traffic signals, thereby facilitating their swift passage through intersections.

The contributions of this research are manifold. It addresses the urgent need for efficient EVD systems, bolstering emergency response capabilities in urban settings. Additionally, it emphasizes compliance with ISO guidelines for standardized siren signals, ensuring uniformity and compatibility. Furthermore, the adoption of deep learning models underscores the potential of advanced technologies in addressing real-world challenges in traffic management and emergency response.

In essence, the integration of EVD systems with intelligent transportation infrastructure marks a significant leap forward in enhancing traffic safety and streamlining emergency services coordination. By harnessing deep learning methodologies and upholding international standards, this study contributes to enhancing the efficiency and efficacy of emergency response operations.

I. The initial phase involves gathering data, preprocessing it, and extracting features from audio files sourced from an open-access library.
II. Following that, we proceed with the implementation of fully connected neural network (NN) and Convolutional Neural Network (CNN) models, each configured with various layers and parameters.
III. Subsequently, we deploy the Recurrent Neural Network (RNN) model and fine-tune hyperparameters by exploring different configurations.
IV. To enhance classification accuracy, an ensemble model is crafted by combining three deep learning models: fully connected (FCNet), CNN model (CNN_Net), and RNN model (RNN_Net), specifically designed for siren sound classification.

The paper is structured as follows: In Section II, a comprehensive review of literature pertaining to acoustic-based emergency vehicle detection is provided. Section III delves into the analysis and evaluation of different models for siren sound classification. Experiment results are discussed in detail in Section IV, with concluding remarks presented in Section V.

2. RELATED WORK

Research in the field of siren sound recognition has been relatively sparse, as indicated by existing literature spanning references [2–13]. For instance, J. Liaw et al. [2] proposed a methodology for identifying ambulance sirens using Longest Common Subsequence (LCS) in Taiwan, achieving an accuracy rate of 85%. Another study [3] introduced Mel-Frequency Cepstral Coefficient (MFCC)-based speech recognition technology combined with multilayer neural networks and majority voting techniques for siren sound detection. Despite boasting low computational complexity, this approach encountered difficulties in effectively analyzing noisy and diverse datasets, prompting the use of a reproduction technique to augment training and testing data. Additionally, [4] explored two methodologies for siren identification: a multilayer neural network (MNN) system adapted from speech recognition, and a sinusoidal model system aimed at extracting signals from background noise to minimize interference. Both methods underwent evaluation on limited datasets, yielding comparable accuracy rates. Moreover, [5] introduced part-based models (PBMs), initially utilized in computer vision, for detecting sirens amidst noisy traffic environments, leveraging spectro-temporal domain analysis. PBMs trained on MFCC or log-mel attributes outperformed hidden Markov models (HMMs), albeit with accuracy rates below 90%.

In the domain of intelligent vehicle systems, L. Marchegiani et al. [6] proposed a two-stage detector customized for audio-based detection. The initial phase aimed at identifying irregular sounds, while the subsequent stage focused on mitigating noise and conducting classification. Inspired by image processing methodologies, each incoming signal's spectrogram was treated akin to an image, employing segmentation techniques to isolate and discern the target signal. Following noise reduction, the K-Nearest Neighbor (KNN) algorithm was applied, resulting in an accuracy rate of 83%. Meanwhile, in [7], the detection of sirens involved analyzing audio signals using digital signal processing techniques, such as estimating frequency components within specific frequency ranges. Additionally, [8] utilized Support Vector Machine (SVM) with feature selection methods for alarm sound detection, achieving an accuracy of over 90% on a limited dataset. However, a notable drawback of this approach was the substantial time investment required for feature engineering.

Various investigations have explored the utilization of microcontrollers [9, 10, and 11] and hardware design [12, 13] for alarm sound detection. In [9], ambulance detection based on siren sound involved conducting Fast Fourier Transform (FFT) twice on a microcontroller. Despite its ability to accommodate the Doppler Effect, this method's high computational cost posed limitations. F. Meucci et al. [10] developed a microcontroller-based system that utilized the frequency and periodic repetition characteristics of siren sounds for emergency vehicle detection. However, this model's evaluation was confined to sirens with frequencies of 392 Hz and 660 Hz. In [11], Durbin's recursive method was employed alongside a linear prediction model to assist drivers with hearing impairments. Concurrently, R. Dobre et al. [12, 13] engineered an analog electronics circuit-based system with low computational demands, evaluated using the SPICE simulator. Although this system exhibited satisfactory accuracy on small datasets, its performance diminished when applied to larger datasets. Additionally, Fatimah et al. [25] proposed an emergency vehicle detection model that utilized bandpass filters for signal processing. This model incorporated two types of features and compared various machine learning algorithms, including KNN, SVM, and ensemble methods, to identify the most effective approach.

CNN has gained widespread recognition for its effectiveness in various audio recognition applications, such as music tagging, automatic speech recognition (ASR) [18, 19], and environmental sound classification [14–17]. In the realm of environmental sound classification, V. Boddapati et al. [14] investigated popular image recognition networks like GoogleNet and AlexNet, training them using spectrogram and Mel-Frequency Cepstral Coefficients (MFCC) as input data, achieving accuracies of up to 90%. Similarly, K. Piczak [16] and J. Salamon et al. [15] proposed CNN-based models for environmental sound classification, utilizing log-mel spectrogram data for training. However, the models presented in [15] and [16] achieved accuracies of less than 80%. Additionally, Baghel et al. [26] employed the YOLO model with two phases for emergency vehicle recognition, with one phase generating bounding boxes and the other handling classification. Tran et al. [27] proposed a hybrid audio and vision-based model for emergency vehicle recognition, incorporating YOLO for image processing and WaveResNet for audio processing.

The preceding efforts in Emergency Vehicle Detection (EVD) have been hindered by several significant limitations: a scarcity of experimental data, dependence on hardware-based systems and shallow algorithms, and reliance on manually crafted features extracted solely from either the time or frequency domain for model training. To overcome these challenges, advancements in EVD have been achieved through the acquisition of datasets from an open-source repository, particularly the

Audio Set ontology, and the adoption of advanced deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).

3. PROPOSED METHODOLOGY

This study introduces and evaluates a novel approach for the classification of emergency vehicle siren sounds, employing both a recurrent neural network (RNN) and an ensemble model composed of three distinct deep learning architectures. The ensemble model comprises a fully connected neural network (FCNet), a convolutional neural network (CNN_Net), and a recurrent neural network (RNN_Net), each contributing unique strengths to the classification task.

The FCNet architecture is characterized by dense layers exclusively, omitting convolutional layers. It undergoes thorough exploration with varying configurations, encompassing up to 8 fully connected layers, while different parameter settings are examined to identify the most effective model. Meanwhile, the CNN_Net incorporates a flexible number of 2D convolutional layers, ranging up to 6, with adjustable filter counts and a fixed 4x4 kernel size. To counteract overfitting, a max-pooling layer follows the convolutional layers, supplemented by a dropout layer with a parameter set to 0.25 after the dense layer. On the other hand, the RNN_Net leverages a recurrent neural network (RNN) structure, integrating various combinations of long short-term memory (LSTM) layers with differing neuron counts.

Following an exhaustive evaluation of different configurations of the three base models, the ensemble model is constructed via majority voting. The FCNet component comprises eight layers, CNN_Net features 3 layers, and RNN_Net is composed of five layers. The ensemble's architecture is visually depicted in Figure 1, elucidating the interplay among the constituent models. Additionally, Algorithm 1 elucidates the operational mechanism of the proposed system, offering a step-by-step guide to its functioning.

This holistic approach harnesses the unique capabilities of each base model to enhance the overall classification performance, thereby providing a comprehensive solution for accurately identifying emergency vehicle sirens. By leveraging the strengths of both individual architectures and the collective wisdom of the ensemble, this method promises improved accuracy and reliability in siren sound classification, addressing a crucial need in emergency vehicle detection systems.

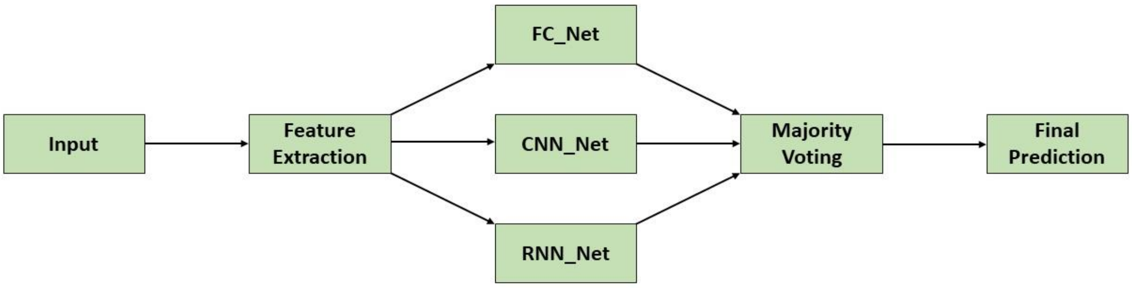

Figure 1: The proposed ensemble model's architecture

Algorithm 1: Proposed Methodology

Input: An audio file

Output: The anticipated category of rescue vehicle.

START:

Step-1: Use the MFCC method to extract features from the provided audio file.

Step-2: Give the three basic models—FCNet, CNN_Net, and RNN_Net—the extracted features, and then store the predictions in the appropriate variables, y_FCNet, y_CNN_Net, and y_RNN_Net.

Step-3: After applying majority voting to the generated predictions, return the final forecast for mode (y_FCC_Net, y_CNN_Net, y_RNN_Net).

STOP

4. EXPERIMENTAL RESULTS and DISCUSSION

a. Data Collection

The experimental dataset employed in this research is sourced from the Google Audioset Ontology [24], which organizes sound events hierarchically. This ontology encompasses a wide range of sounds, spanning from animals and humans to environmental sounds, music, and various miscellaneous categories. Specifically, it includes recordings of siren sounds from four distinct types of vehicles: Police Car, Ambulance, Fire Engine, and Civil Defence Siren, all presented in video format. Detailed information about each video is stored in a CSV file, including essential details such as the YouTube link, start and end times, and corresponding labels. To obtain relevant data, videos featuring sirens from Ambulance, Police cars, and Fire trucks are downloaded using the "pafy" and "youtube_dl" Python libraries. Subsequently, using the "moviepy" library, the audio files are processed to extract only the siren sound from the downloaded videos, discarding unnecessary audio segments.

b. Feature Extraction

In this research, Mel Frequency Cepstral Coefficient (MFCC) serves as the chosen method for feature extraction from audio data. It yields 39 distinct features from the dataset, with the first feature representing audio pitch, while the subsequent 12 features relate to frequency amplitude. The process of feature extraction is visually outlined in Figure 2. Utilizing the "Librosa" library [20], relevant information is extracted from the audio files. The resulting shape of the feature vector and target is (259169, 40) and (1,301), respectively. Moreover, Figure 3 visually presents the waveform depicting the audio of a police car siren sound.

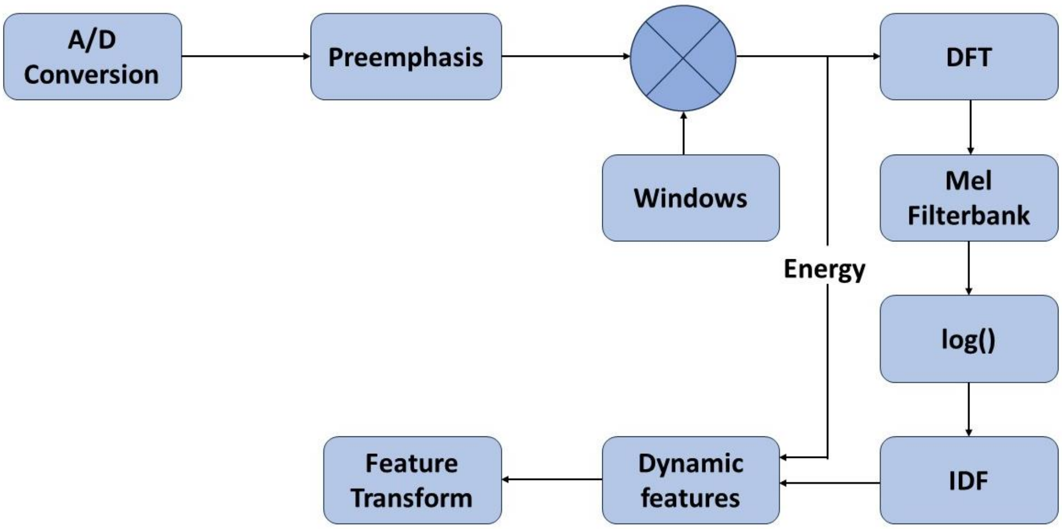

Figure 2: Feature extraction flow chart

Figure 3: Police vehicle siren waveform

c. Hyperparameter Tuning

In this study, thorough exploration and analysis were conducted on three different types of deep learning models. Optimal configuration of layers and parameters is crucial for maximizing the performance of deep learning networks, thus extensive experimentation was carried out to understand the impact of varying configurations on all three models. Subsequently, models demonstrating superior performance in acoustic-based Emergency Vehicle Detection (EVD) were selected from each configuration.

Tables 1, 2, and 3 detail the training and testing accuracies corresponding to different layer configurations and parameters in the FCNet, CNN_Net, and RNN_Net models, respectively. Implementation of these deep learning architectures was facilitated using the "TensorFlow" framework, an open-source library developed by Google, tailored for complex mathematical operations, particularly in machine learning and neural networks. Model training was conducted on Google Colaboratory, offering free GPU support to the public, expediting the training process.

Throughout all configurations, the Rectified Linear Unit (ReLU) activation function was applied to the hidden layers, while the Softmax activation function was employed at the output layer. Furthermore, model training utilized the Adam optimizer [22] with a learning rate of 0.001 and decay set at 0.0001, alongside categorical cross-entropy loss. This rigorous methodology ensured a comprehensive exploration of configurations, ultimately enhancing the accuracy and effectiveness of the acoustic-based EVD system.

Table 1: Quantity of fully connected layers and parameters within a multilayer fully connected neural network

Layer	FC Layer-2	FC Layer-3	FC Layer-4	FC Layer-5	FC Layer-6	FC Layer-7	FC Layer-8
Input	0	0	0	0	0	0	**0**
FC-1024	35267584	35267584	35267584	35267584	35267584	35267584	**35267584**
FC-512	524800	524800	524800	524800	524800	524800	**524800**
FC-512	—-	262656	262656	262656	262656	262656	**262656**
FC-256	—-	—-	—-	65792	65792	65792	**65792**
FC-128	—-	—-	—-	—-	32896	32896	**32896**
FC-64	—-	—-	—-	—-	—-	8256	**8256**
FC-32	—-	—-	—-	—-	—-	—-	**2080**
Output-3	1539	1539	771	771	387	195	**99**
Total Parameters	36,056,579	36,319,235	36,449,795	36,515,587	36,548,099	36,556,163	**36,558,147**
Training Accuracy %	100	99.58	100	99.58	100	100	**100**
Testing Accuracy %	60	70	75	84	92	94.6	**96.4**

Table 2: Quantity of 2D convolutional layers and parameters within a convolutional neural network

Layer	Conv_Layers-2	Conv_Layers-3	Conv_Layers-4	Conv_Layers-5	Conv_Layers-6
Input	0	**0**	0	0	0
Conv 4X4 – 32	16416	**16416**	16416	16416	16416
Conv 4X4 – 64	—-	**32832**	32832	32832	32832
Conv 4X4 – 128	—-	**—-**	—-	—-	262272
FC – 512	564,265,472	**281805312**	70451712	35062272	6947328
FC- 64	32832	**16416**	32832	32832	32832
Output 3	195	**99**	195	195	195
Total	564315459	**281871619**	70600131	35341891	7489219

Training Accuracy %	100	**99.58**	93.3	95	95
Testing Accuracy %	61	**92.4**	85.3	88.6	84.4

Table 3: Distinct long short term memory (LSTM) layers in anRNN, along with their corresponding parameters

Layer	**LSTM_Layer-2**	**LSTM_Layer-3**	**LSTM_Layer-4**	**LSTM_Layer-5**	**LSTM_Layer-6**
Input	0	0	0	**0**	0
LSTM 32	9344	9344	9344	**9344**	9344
LSTM 32	8320	8320	8320	**8320**	8320
LSTM 128	—-	—-	—-	**—-**	131584
FC 128	4224	8320	8320	**16512**	16512
Output 3	387	387	387	**387**	387
Total	22275	51203	84227	**191235**	322819
Training Accuracy %	84.07	89.6	92.2	**98.7**	90.4
Testing Accuracy %	61.29	75.7	85.2	**94.5**	84.1

d. Results and Discussion

The comparison between the proposed model and various deep learning architectures is illustrated in Figure 4, where four distinct models are examined. Firstly, the FC_Net model, comprising dense layers exclusively, achieved an accuracy of 96.4% with a remarkably low inference time of 0.061 seconds. Following that, the CNN_Net model delivered an accuracy of 92.4%, while the RNN_Net and Ensemble models exhibited accuracies of 94.5% and 98.7%, respectively. Table 4 supplements these findings by presenting inference times, revealing that RNN_Net and FC_Net share similar processing times, while CNN_Net necessitates a longer duration for processing. Remarkably, the Ensemble model displays the lengthiest response time, nearly 1.5 seconds.

Furthermore, the study conducts evaluations on various machine learning models using the collected dataset and compares their results against the proposed deep learning models. Table 5 illustrates that while decision trees and random forests may achieve higher training accuracy, their testing accuracy falls significantly lower, indicative of overfitting. Conversely, the proposed deep learning models demonstrate superior accuracy and are deemed acceptable models for emergency vehicle detection tasks.

Given the limited scope of accuracy evaluations on siren detection systems based on microcontrollers and circuit design, the study predominantly focuses on comparing machine learning (ML) and deep learning (DL) methods. Table 6 juxtaposes the proposed models with previous findings [2, 6, 8, 21, and 23] in terms of methodology, functionality, and prediction accuracy. Notably, the classification accuracy of the CNN model proposed by L. Marchegiani [21] aligns closely with the proposed model employing RNN. Machine learning models such as KNN [6], HMM [8], and part-based models [8] exhibit accuracies below 90%. Conversely, CNN models developed by Tran [23] achieved accuracies up to 98.24%. However, the proposed ensemble model surpasses these results, boasting an accuracy of 98.7%, the highest among all the works compared. This remarkable achievement elevates the proposed ensemble model as the top performer, surpassing both the proposed RNN_Net (94.5%) and other related works [2] 85%, [6] 83%, [8] 86%, [21] 94%, and [23] 98.24%.

In conclusion, the comparative analysis underscores the superior performance of the proposed ensemble model in terms of accuracy compared to both deep learning and machine learning models. The ensemble model not only outperforms individual deep learning models but also surpasses the accuracy of previous works, demonstrating its potential for effective emergency vehicle detection. An in-depth summary of the proposed model's performance is presented in Table 6, further emphasizing its superiority compared to existing methods.

Table 4: Evaluation of Various Models in Comparison

Model	Accuracy	Inference Time (s)
RNN_Net	94.5	0.0651
Ensemble	97.7	1.7
CNN_Net	92.8	0.141
FCNet	95.4	0.066

Table 5: A Comparison of the Suggested Model and Current Approaches

Model/Technique	Accuracy(%)
[2]	86
[6]	84
[8]	87
[21]	93
[23]	97.24
RNN_Net (Proposed)	94.9
Ensemble (Proposed)	98.6

5. CONCLUSION and FUTURE WORKS

This paper presents a novel ensemble approach to emergency vehicle detection, utilizing deep learning models based on siren sounds. The ensemble model comprises three components: a fully connected model, a CNN model, and an RNN model, all trained on MFCC features extracted from collected data. Through extensive experimentation, the study showcases the superior performance of the proposed ensemble model compared to existing alternatives. Notably, the ensemble model achieves an impressive accuracy of 98.7%, outperforming the individual RNN model, which achieves an accuracy of 94.5%. Moreover, the study conducts a comprehensive performance analysis, evaluating deep learning models against various machine learning algorithms such as Perceptron, SVM, and decision trees.

One significant advantage of acoustic-based models over image-based counterparts is emphasized. While capturing images of fast-moving emergency vehicles poses challenges, detecting and processing their siren sounds from long distances is feasible, enabling early warnings and efficient processing. This highlights the suitability of acoustic-based approaches for emergency vehicle detection, particularly in real-time applications such as at intersections where prompt detection is critical for prioritizing vehicle movement and reducing waiting times.

Despite the promising results achieved by the proposed ensemble model, the study acknowledges the need for further enhancements to meet the stringent requirements of a reliable and user-friendly emergency vehicle detection system. Specifically, future research could explore techniques for localizing siren sounds to determine the direction of approaching emergency vehicles, thereby augmenting the system's capabilities and enhancing overall efficiency. Continued innovation and refinement in this domain hold significant potential for improving safety and response times in emergency scenarios.

REFERENCES

1. Schmidhuber J. Deep Learning in Neural Networks: An Overview. *Neural Netw.* 2015;61:85–117.

2. Hatcher W.G., Yu W. A Survey of Deep Learning: Platforms, Applications and Emerging Research Trends. *IEEE Access.* 2018;6:24411–24432.

3. Celesti F., Celesti A., Wan J., Villari M. Why Deep Learning Is Changing the Way to Approach NGS Data Processing: A Review. *IEEE Rev. Biomed. Eng.* 2018;11:68–76.

4. Lee D., Yang J., Kim S. Learning the Histone Codes with Large Genomic Windows and Three-Dimensional Chromatin Interactions Using Transformer. *Nat. Commun.* 2022;13:6678.

5. Schroder J., Goetze S., Grutzmacher V., Anem uller Jorn, Automatic acoustic siren detection in traffic noise by Part-based Models, in: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

6. Marchegiani L., Posner I., Leveraging the urban soundscape: Auditory perception for smart vehicles, in: *IEEE Int. Conf. on Robotics and Automation* (ICRA), , Singapore, 2017, pp. 6547–6554.

7. Khan A, Sohail A, Zahoora U, Qureshi AS (2020) A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev 53(8):5455–5516.

8. Lin T-Y et al (2014) Microsoft COCO: common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) ECCV 2014, vol 8693. LNCS. Springer, Cham, pp 740–755.

9. Sen, S.; Sugiarto, D.; Rochman, A. Komparasi Metode Multilayer Perceptron (MLP) Dan Long Short Term Memory (LSTM) Dalam Peramalan Harga Beras. *Ultimatics* **2020**, *XII*, 35–41.

10. Mittal, U.; Chawla, P. Acoustic Based Emergency Vehicle Detection Using Ensemble of Deep Learning Models. *Procedia Comput. Sci.* **2023**, *218*, 227–234.

11. T. Miyazaki, Y. Kitazono and M. Shimakawa, "Ambulance siren detector using FFT on dsPIC", *Proc. 1st IEEE/IIAE Int. Conf. Intell. Syst. Image Process.*, pp. 266-269, Sep. 2013.

12. Penzel T., Moody G.B., Mark R.G., Goldberger A.L., Peter J.H. The apnea-ECG database; Proceedings of the Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163); Cambridge, MA, USA. 24–27 September 2000; pp. 255–258.

13. Cortes C., Vapnik V. Support-vector networks. *Mach. Learn.* 1995;**20**:273–297. doi: 10.1007/BF00994018.

14. K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, A. K. Nandi, Credit card fraud detection using adaboost and majority voting, IEEE Access 6 (2018) 14277–14284. [214]

15. A. Mart´ın, R. Lara-Cabrera, D. Camacho, Android malware detection through hybrid features fusion and ensemble classifiers: The andropytool framework and the omnidroid dataset, Information Fusion 52 (2019) 128–142.

16. F. Meucci, L. Pierucci, E. Del Re, L. Lastrucci and P. Desii, "A real-time siren detector to improve safety of guide in traffic environment", *Proc. 16th Eur. Signal Process. Conf. (EUSIPCO)*, pp. 25-29, 2008.