Prisma Sains: Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram
https://e-journal.undikma.ac.id/index.php/prismasains/index
e-mail: prismasains.pkpsm@gmail.com

April 2026. Vol. 14, No. 2
p-ISSN: 2338-4530
e-ISSN: 2540-7899
pp. 521-534

Ambulance Siren Audio Classification Using Convolutional Neural Network for Medical Emergency Detection

[1*]Ramadhan Paninggalih, [2]Bima Prihasto, [3]Maryo Inri Pratama, [4]Rizky Irswanda Ramadhana, [5]Misbahuddin, [6]Buan Anshari, [7]Lalu Ahmad Syamsul Irfan Akbar, [8]Giri Wahyu Wiriasto

[1,2,3,4]Institut Teknologi Kalimantan, Balikpapan, Indonesia
[5,6,7,8]Universitas Mataram, Mataram, Indonesia

*Corresponding Author e-mail: ramadhanpaninggalih@lecturer.itk.ac.id

Received: February 2026; Revised: March 2026; Published: April 2026

Abstract

The rapid detection of emergency vehicle sirens is critical for enhancing road safety and traffic management. This study proposes an automated classification system for ambulance sirens using a Convolutional Neural Network (CNN). The method utilizes Mel-Frequency Cepstral Coefficients (MFCC) to transform audio signals into 2D feature maps, allowing the model to capture distinct spectral and temporal patterns. The dataset was preprocessed using a stratified split to ensure balanced class distribution and prevent data leakage. Experimental results demonstrate that the CNN model achieves a high performance with an accuracy of 0.95, significantly outperforming baseline models such as Multi-Layer Perceptron (MLP) and XGBoost. Detailed evaluation through a confusion matrix indicates a consistent precision, recall, and F1-score of 0.95, proving the model's robustness in distinguishing sirens from complex urban noise. The implementation of the Adam optimizer and early stopping mechanism ensured stable convergence and prevented overfitting. These findings suggest that the proposed CNN-MFCC framework provides a reliable solution for real-time emergency signal detection, offering a substantial contribution to intelligent transportation systems.

Keywords: Ambulance siren; Convolutional neural network; Mel-frequency cepstral coefficients; Sound classification

How to Cite: Paninggalih, R., Prihasto, B., Pratama, M. I., Ramadhana, R. I., Misbahuddin, M., Anshari, B., … Wiriasto, G. W. (2026). Ambulance Siren Audio Classification Using Convolutional Neural Network for Medical Emergency Detection. *Prisma Sains: Jurnal Pengkajian Ilmu Dan Pembelajaran Matematika Dan IPA IKIP Mataram*, *14*(2), 521–534. https://doi.org/10.33394/j-ps.v14i2.20099

https://doi.org/10.33394/j-ps.v14i2.20099

INTRODUCTION

Urban traffic congestion remains a persistent and multifaceted challenge in contemporary cities, especially in densely populated regions where rising vehicle density and constrained infrastructure contribute to transportation inefficiencies. Previous studies have shown that traffic congestion leads to economic losses, environmental degradation, and additional external costs by impeding emergency response systems (Brent & Beland, 2020). Recent findings indicate that congestion can substantially delay emergency vehicles, increasing response times by over 30% in urban settings (Alruwaili et al., 2025). In developing and rapidly urbanizing areas, inadequate planning and elevated traffic density further intensify these issues, resulting in prolonged travel times and reduced system efficiency (Alslamah et al., 2023). Congestion presents a significant obstacle to emergency response, as ambulances require swift and unobstructed movement. Multiple studies have established that longer emergency response times negatively impact patient outcomes, with each minute of delay in time-sensitive emergencies decreasing survival probability (Damdin et al., 2025). Despite prioritized passage,

unpredictable traffic conditions in real-world scenarios continue to impede timely emergency responses (Luan & Jiang, 2024).

To address these challenges, automated and intelligent systems capable of detecting emergency vehicles in real time are needed to support adaptive traffic management. Recent advances in artificial intelligence (AI) and deep learning have significantly improved pattern recognition capabilities, particularly in audio classification (Kamaladevi et al., 2023; Shah et al., 2023). Among these approaches, Convolutional Neural Networks (CNNs) have demonstrated strong performance due to their ability to learn hierarchical feature representations. Although originally developed for image processing, CNNs can be effectively applied to audio analysis by transforming signals into time-frequency representations, such as Mel-Frequency Cepstral Coefficients (MFCCs), enabling accurate detection of complex acoustic patterns (Kong et al., 2020; Onisha et al., 2024; Zaman et al., 2023).

Several studies have explored emergency vehicle sound classification using various machine learning and deep learning approaches. For instance, initial investigations using 1D-CNN and hybrid temporal-spectral features have shown promising classification performance (Gourisaria et al., 2024; Jayakumar et al., 2024; Parineh et al., 2023; Usaid et al., 2022). However, a significant portion of these existing methods relies on complex architectures that demand high computational resources, making them difficult to deploy on low-power edge devices. Furthermore, while large-scale datasets have been introduced to improve model generalization (Asif et al., 2022; Shams et al., 2024; Zbancioc & Feraru, 2024), there is a noticeable lack of focus on optimizing models for extremely short audio segments. In real-world traffic scenarios, an automated system must identify a siren within seconds—often amidst heavy noise—to trigger immediate traffic signal priority.

Despite the advancements in emergency vehicle detection, several studies still face challenges in maintaining high precision within dense urban environments. Previous research predominantly utilized traditional machine learning models relying on flattened 1D acoustic features, which often struggle to distinguish the rhythmic modulation of sirens from stochastic urban noises. This reliance on computationally heavy models or long audio samples creates a gap between laboratory results and practical real-time feasibility in dynamic traffic scenarios.

In response to this gap, this study proposes a classification framework that treats Mel-Frequency Cepstral Coefficients (MFCC) as 2D feature maps rather than isolated vectors, using a lightweight CNN-based approach. By utilizing 5-second audio segments, this research focuses on optimizing the model's sensitivity to hierarchical spatial patterns inherent in siren frequency modulations. Rather than introducing a fundamentally new deep learning architecture, this work emphasizes an application-oriented evaluation of performance-efficiency trade-offs. To ensure methodological rigor, the proposed model is validated using a stratified data splitting strategy and is compared against robust baselines, specifically Multi-Layer Perceptron (MLP) and XGBoost, to provide a clear benchmark for future real-time intelligent traffic management systems.

METHOD

Research Design

The research design employs a systematic workflow to develop an ambulance siren classification model, as depicted in Figure 1. The methodology begins with the collection of datasets from publicly available sources. Audio preprocessing follows to standardize data formats and ensure sample consistency. Preprocessed audio signals are then transformed into Mel-frequency cepstral coefficients (MFCCs). The extracted features are converted into two-dimensional image representations to ensure compatibility with convolutional neural networks (CNNs). The dataset is split into training and test sets for model development and evaluation. The CNN is trained on MFCC-based image inputs to identify discriminative patterns in siren sounds. Model performance is evaluated using standard classification metrics. These include

accuracy, precision, recall, and F1-score, which determine effectiveness in distinguishing ambulance sirens from other audio signals.

Figure 1. Proposed Research Workflow

Dataset Collection and Audio Preprocessing

The dataset in this study was specifically curated to reflect the complex acoustic environment of urban areas in Indonesia. Audio data were sourced from publicly available YouTube videos recorded in various Indonesian cities, ensuring the inclusion of diverse real-world conditions such as heavy traffic congestion, engine idling, and varying recording qualities. To maintain high methodological standards and prevent data leakage, the dataset was partitioned at the source level. This means all audio clips derived from the same original video were assigned exclusively to the same data subset (either training or testing), ensuring that the model is evaluated on entirely unseen environments and recording devices.

The dataset is categorized into two primary classes: Ambulance Siren and Non-Ambulance Sounds. To challenge the model and ensure practical reliability, the non-ambulance class was meticulously composed of "hard negatives," including sirens from fire trucks, police vehicles, and tactical escort (*patwal*) units, alongside general traffic noise, honking, and urban speech. As illustrated in Figure 2, the final dataset consists of 2,198 samples, comprising 1,079 ambulance siren clips and 1,119 non-ambulance clips. This near-balanced distribution (49.1% and 50.9%, respectively) was intentionally designed to minimize class bias and improve the stability of the classification metrics during the training of CNN, MLP, and XGBoost models.

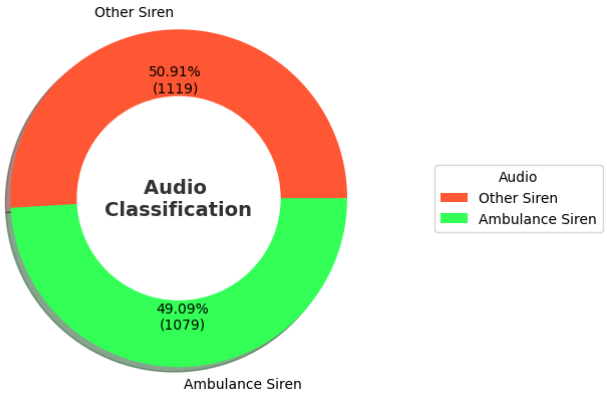


Figure 2. Distribution of audio samples across classes

For the preprocessing stage, all raw audio files underwent a standardized pipeline to ensure consistency. Each recording was downsampled to a sampling rate of 16,000 Hz to balance the preservation of critical siren frequency components—which typically reside below 8 kHz—with computational efficiency for real-time edge deployment. The audio was then converted to a mono-channel WAV format and segmented into fixed 5-second clips. Furthermore, loudness normalization was applied to each clip to prevent the model from relying on volume intensity as a predictive feature, forcing it instead to learn the distinctive temporal and spectral patterns inherent to ambulance sirens.

Mel-Frequency Cepstral Coefficients

Mel-frequency cepstral coefficients (MFCC) are widely used in audio signal processing due to their ability to effectively represent the perceptual characteristics of sound in a compact form. In this study, Mel-frequency cepstral coefficients are employed to extract relevant features from the preprocessed audio signals. MFCC transforms time-domain audio signals into a time-frequency representation that captures important spectral properties aligned with

human auditory perception, making it particularly suitable for sound classification tasks, including siren detection. (Costantini et al., 2023; Rawat et al., 2023) showed that MFCC features can effectively capture discriminative acoustic patterns for sound classification tasks, while (Abbaskhah et al., 2023) highlighted that MFCC provides a compact and informative representation suitable for machine learning and deep learning models.

The MFCC extraction process follows several stages, as illustrated in Figure 3, including pre-emphasis to amplify high-frequency components, framing and windowing to segment the signal into short frames, transformation to the frequency domain using Fast Fourier Transform (FFT), application of the Mel filter bank to mimic human auditory perception, and finally the Discrete Cosine Transform (DCT) to produce the MFCC coefficients. These steps ensure that both temporal and spectral characteristics of the audio signal are effectively captured, as demonstrated in previous studies (Rezaul et al., 2024).

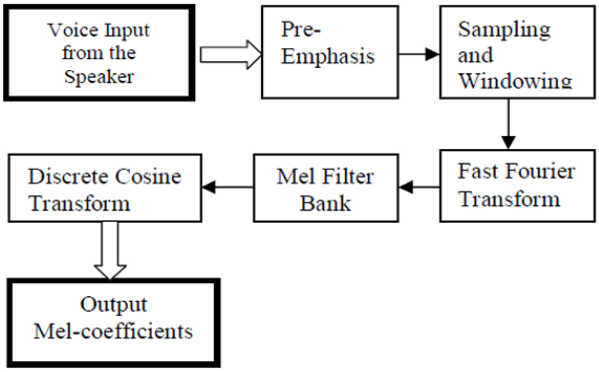


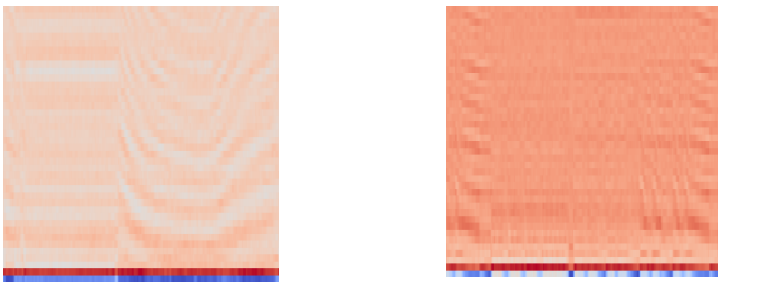
Figure 3. MFCC Process Stages

The MFCC features are extracted using carefully selected parameter settings to balance temporal and frequency resolution while maintaining computational efficiency. The extracted features are subsequently transformed into two-dimensional representations (MFCC feature maps), which serve as input to the convolutional neural network (CNN) model (Asif et al., 2022). This transformation enables the model to leverage spatial feature learning capabilities while preserving essential acoustic information from the original audio signals. The detailed parameter configuration used in this study is summarized in Table 1.

Table 1. MFCC Parameter Settings

Parameter	Value	Description
n_fft	2048	Length of FFT window
hop_length	512	Step size between frames
n_mfcc	40	Number of MFCC coefficients
figsize	(3, 3)	Size of MFCC image representation

To illustrate the effectiveness of the feature representation, examples of MFCC feature maps for ambulance and non-ambulance audio signals are shown in Figure 4.



(a) MFCC Ambulance Siren (b) MFCC Non Ambulance Siren

Figure 4. Example of MFCC representations

The visual representations indicate differences in spectral patterns between classes, where ambulance sirens tend to exhibit more structured and repetitive frequency patterns compared to non-ambulance sirens. These differences support the ability of CNN models to learn discriminative features from MFCC representations.

Data Splitting

To ensure a rigorous evaluation and unbiased performance assessment, the dataset was partitioned into three distinct subsets: training, validation, and testing, using a stratified splitting strategy. A 70%-15%-15% ratio was applied, which resulted in 1,538 samples for training, 330 samples for validation, and 330 samples for testing. Unlike simple random sampling, the stratified approach ensures that the nearly balanced proportion of ambulance and non-ambulance classes in Figure 2 is strictly maintained across all three subsets.

This partitioning strategy serves three critical purposes: the training set allows the models (CNN, MLP, and XGBoost) to learn the underlying acoustic features; the validation set is utilized for hyperparameter tuning and preventing overfitting during the training process; and the testing set provides a final, objective measure of the model's generalization capability on entirely unseen data. By employing stratified splitting, the experimental design achieves high statistical reliability and addresses the requirement for rigorous validation without the computational overhead of exhaustive cross-validation, ensuring that performance metrics are not skewed by class distribution variances.

Convolutional Neural Network

Convolutional Neural Networks (CNNs) represent a class of deep learning models that learn hierarchical feature representations from structured data, especially images and spatially correlated inputs. With convolutional operations, CNNs extract local features using learnable filters. This enables effective pattern recognition (Chu et al., 2023). Typical architectures have convolutional, pooling, and fully connected layers. Together, these capture features that range from low-level to high-level abstractions (Alruwaili et al., 2025). CNNs also process time-frequency representations like spectrograms and Mel-frequency cepstral coefficient (MFCC) images. They efficiently learn spatial dependencies within such data (Farooq et al., 2024).

A sequential CNN architecture is proposed to classify ambulance and non-ambulance siren sounds using MFCC image inputs. The model contains three convolutional blocks. Each block has a Conv2D layer, Batch Normalization, and MaxPooling2D. The filters increase from 32 to 64, then 128. This structure helps the model learn more complex feature representations. A Flatten layer converts the feature maps into a one-dimensional vector. That vector is then processed by a fully connected Dense layer with 128 neurons. The final output layer uses a Softmax activation function for classification. Batch Normalization stabilizes and accelerates training. MaxPooling reduces spatial dimensions and computational complexity. This improves generalization performance. The proposed architecture aims to learn discriminative patterns from MFCC feature maps while balancing model complexity and computational efficiency. Table 2 summarizes the detailed architecture of the proposed CNN model.

Table 2. Proposed CNN Architecture

Block	Layer	Configuration	Output
Input	Input Layer	128×128×1	128×128×1
Block 1	Conv2D + BN + MaxPool	32 filters, 3×3, ReLU	63×63×32
Block 2	Conv2D + BN + MaxPool	64 filters, 3×3, ReLU	30×30×64
Block 3	Conv2D + BN + MaxPool	128 filters, 3×3, ReLU	14×14×128
–	Flatten	–	25,088
–	Dense	128, ReLU	128
Output	Dense	Sigmoid	1

Evaluation

The performance of the proposed classification model was evaluated using standard metrics, including accuracy, precision, recall, and F1-score. These metrics are commonly used in classification tasks to provide a comprehensive assessment of model performance, particularly in binary classification problems, via the confusion matrix. Relying solely on accuracy can be misleading in classification tasks; therefore, multiple evaluation metrics are necessary for a more reliable assessment (Tharwat, 2020). Additionally, precision, recall, and F1-score are essential for understanding model performance in binary classification (Chicco & Jurman, 2020).

Accuracy quantifies the overall correctness of a model by comparing the number of correctly predicted instances to the total number of predictions. This metric is formally defined in equation (1).

$$Accuracy = \frac{TP + FP}{TP + TN + FP + FN} \tag{1}$$

where TP (True Positive) denotes correctly predicted positive instances, TN (True Negative) denotes correctly predicted negative instances, FP (False Positive) denotes incorrectly predicted positive instances, and FN (False Negative) denotes incorrectly predicted negative instances. Precision measures the proportion of correctly predicted positive instances among all predicted positive instances, as shown in equation (2).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall, also referred to as sensitivity, quantifies the model's ability to correctly identify positive instances, as defined in equation (3).

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The F1-score represents the harmonic mean of precision and recall, offering a balanced metric for evaluating classification tasks. It is defined in equation (4).

$$F1 - Score = \frac{2(Precision \cdot Recall)}{Precision + Recall} \tag{4}$$

These evaluation metrics collectively offer complementary insights into model performance. Accuracy gives an overall assessment. Precision and recall provide a more detailed look at classification behavior. The F1-score integrates both precision and recall. As a result, all four metrics are used in this study to ensure a reliable and comprehensive evaluation of the proposed model.

RESULTS AND DISCUSSION

The dataset used in this research consists of two classes: ambulance sirens (1,079 samples) and non-ambulance sounds (1,119 samples), totaling 2,198 audio samples. These proportions result in a nearly balanced class distribution, with 48.8% ambulance sirens and 51.2% non-ambulance sounds, as illustrated in Figure 2. A balanced dataset is critical to reduce the risk of model bias and ensure that the performance evaluation remains reliable across both classes. To ensure rigorous evaluation and specifically prevent data leakage—a concern often raised in audio classification—the dataset was split into training, validation, and testing sets based on the original video sources rather than individual clips. This group-based splitting ensures that all 5-second segments derived from the same source recording are confined to a single partition. Consequently, the model is prevented from "memorizing" specific background environments or recording characteristics shared between clips from the same source, forcing it to learn generalized discriminative features of the siren patterns instead. This approach

provides a robust foundation for the subsequent CNN training and ensures the validity of the reported accuracy.

Before proceeding to feature extraction and model training, it is essential to analyze the physical characteristics of the audio in the time domain. Figures 3 and 4 illustrate the waveforms of non-ambulance and ambulance sounds, representing the raw audio signals where the x-axis denotes time (seconds) and the y-axis represents the amplitude.

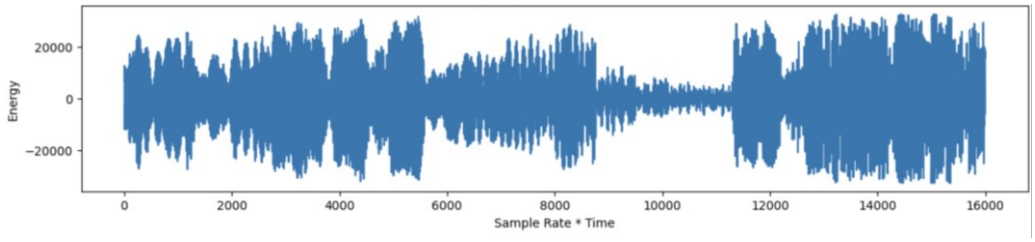


Figure 3. The non-ambulance waveform

The non-ambulance waveform, as shown in Figure 3, displays a stochastic and irregular structure. The signal appears significantly denser and more continuous, which is typical of environmental noise such as engine combustion, wind resistance, and general urban traffic. Due to its unpredictable nature, the non-ambulance audio lacks a discernible repeating pattern or periodic intervals, resulting in a disorganized visual appearance with constant short-term variability.

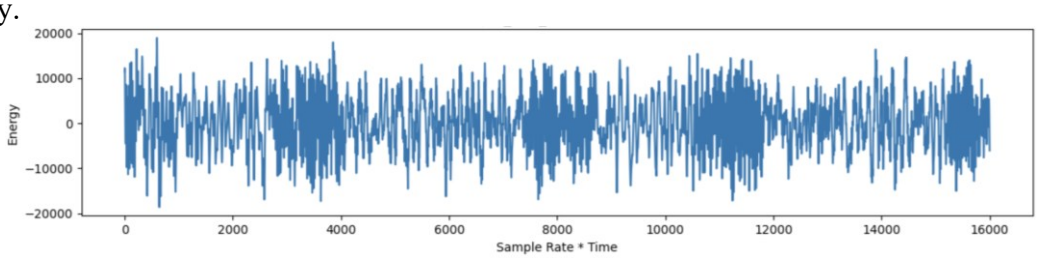


Figure 4. The ambulance siren waveform

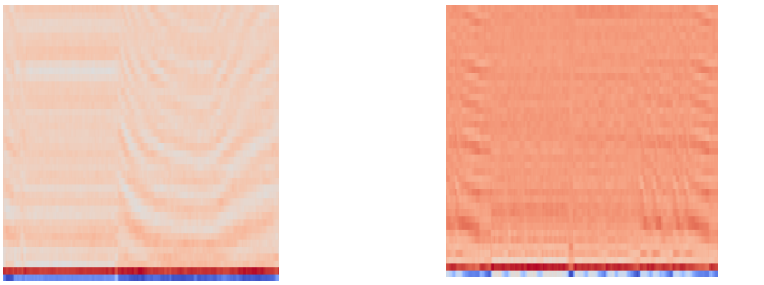
In contrast, the ambulance siren waveform in Figure 4 exhibits a highly structured and periodic pattern. This signal is characterized by repeating amplitude envelopes that correspond to the rhythmic frequency modulation of siren types, such as "wail" or "yelp." These periodic peaks indicate concentrated energy segments that follow a predictable temporal cycle, which is the hallmark of emergency warning signals.

The stark contrast between the chaotic amplitude distribution in Figure 3 and the rhythmic, oscillating peaks in Figure 4 provides a fundamental basis for the feature extraction process. While the non-ambulance signal spreads its energy across a wide range of irregular amplitudes, the ambulance signal concentrates its energy into distinct temporal blocks. This structural disparity ensures that when these waveforms are converted into the frequency domain (MFCC), the resulting feature maps will contain high-contrast geometric patterns, enabling the CNN model to distinguish between the two classes with high precision and minimal inter-class ambiguity.

While these time-domain differences provide initial insights into the signal behavior, the overlap in peak amplitudes between sirens and certain high-energy environmental noises (such as loud honking or heavy machinery) makes differentiation based solely on raw waveforms insufficient for high-accuracy classification. This limitation justifies the necessity of transforming the signals into Mel-Frequency Cepstral Coefficients (MFCCs) in the next stage to capture the spectral-temporal features required for robust detection using Convolutional Neural Networks (CNN).

As established in the methodology, Mel-Frequency Cepstral Coefficients (MFCC) are employed to transform raw audio into a compact time-frequency representation. Following the stages of pre-emphasis, framing, and Fast Fourier Transform (FFT), the signals are mapped

onto the Mel scale using the specific configurations detailed in Table 1. These parameters, including an n_fft of 2048 and n_mfcc of 40, ensure that the resulting two-dimensional feature maps capture spectral properties aligned with human auditory perception while maintaining sufficient resolution for deep learning. Figures 5 and 6 illustrate the extracted MFCC feature maps for the non-ambulance and ambulance classes, respectively. In these visualizations, the x-axis represents time frames, denoting the temporal progression of the 5-second audio segment, while the y-axis represents the Mel filters (coefficients) corresponding to the 40 coefficients defined in Table 1. The color intensity or heatmap represents the energy magnitude in decibels (dB), where brighter regions indicate high energy concentration at specific frequencies and darker regions signify low energy or ambient background noise.



(a) MFCC Ambulance Siren (b) MFCC Non Ambulance Siren

Figure 5. Image MFCC representations

The MFCC feature representation for the non-ambulance class, as illustrated in Figure 5(a), displays a fragmented and stochastic spectral distribution. Based on the extraction parameters detailed in Table 1, this feature map reflects the random characteristics of urban environmental noise, such as engine sounds or general traffic activity. Visually, the y-axis, representing the 40 Mel coefficients, shows an uneven energy distribution across the time frames on the x-axis. The color intensity in this heatmap confirms the absence of stable harmonic structures or repeating frequency patterns, resulting in a disorganized visual texture that is characteristic of background noise.

In contrast, the MFCC feature map for the ambulance class in Figure 5 (b) demonstrates a highly structured and deterministic spectral-temporal signature. The implementation of an n_fft value of 2048 allows for the visualization of sharp and clear horizontal energy bands, which represent the harmonic frequencies of the ambulance siren. Along the x-axis (time), there is a distinct "wavy" pattern or oscillation in the Mel coefficients on the y-axis. This serves as a visual representation of the frequency modulation (the rising and falling pitch) inherent in siren types such as "wail" or "yelp." The hop_length of 512 ensures that these frequency modulation transitions are captured smoothly, producing consistent geometric patterns that are easily recognizable by the convolutional layers of the CNN model.

The comparison between Figure 5 (a) and Figure 5 (b) reveals significant decorative contrasts in terms of spectral organization and energy consistency between the two classes. While the non-ambulance signals are dominated by broadly scattered random energy without a clear tonal structure, the ambulance signals possess a linear and periodic spectral texture. The primary distinction lies in the presence of harmonic frequencies and pitch modulation that form a discriminative visual signature for the ambulance class. This sharp contrast between the fragmented spectral texture of the non-target class and the orderly harmonic line patterns of the target class minimizes ambiguity for the model, enabling the classification system to achieve high accuracy and reliable performance in detecting ambulance presence.

Furthermore, the periodic "waviness" observed in the ambulance feature map reflects the frequency modulation—the characteristic rising and falling pitch—inherent in siren patterns like "wail" or "yelp." These distinct signatures confirm that the parameters listed in Table 1, particularly the hop_length of 512, successfully preserve the discriminative acoustic information necessary for spatial feature learning in the subsequent CNN layers. By converting

the 1D signals into these structured 2D representations based on the settings in Table 1, the classification task is effectively transformed into a pattern recognition problem. The stark contrast between the organized spectral harmonics of the ambulance siren and the disorganized noise of the non-ambulance class provides the discriminative power required for the model to achieve robust performance and high accuracy in complex real-world scenarios.

To ensure rigorous evaluation and specifically prevent data leakage, a critical concern in audio classification, the dataset was partitioned into training, validation, and testing sets using a stratified group-based splitting strategy. Unlike simple random sampling, this approach ensures that the split is based on the original video sources rather than individual clips. All 5-second segments derived from a single source recording are confined to a single partition, forcing the model to learn generalized discriminative features of the siren patterns rather than memorizing specific background environments shared between clips.

The dataset was divided with a 70%–15%–15% ratio, resulting in 1,538 samples for training, 330 samples for validation, and 330 samples for testing. The detailed distribution of these samples across subsets is summarized in Table 3, ensuring the class balance is maintained across all experimental phases. This robust foundation guarantees that the performance metrics reported in the subsequent sections reflect the model's true generalization capability on entirely unseen data.

Table 3. Distribution of Audio Samples across Dataset Subsets

Class	Training (70%)	Validation (15%)	Testing (15%)
Ambulance Siren	755	162	162
Non Ambulance	783	168	168
Total	1.538	330	330

The Convolutional Neural Network (CNN) model in this study was configured with specific hyperparameters to ensure optimal training efficiency and classification accuracy, as detailed in Table 4.

Table 4. CNN Training Hyperparameters

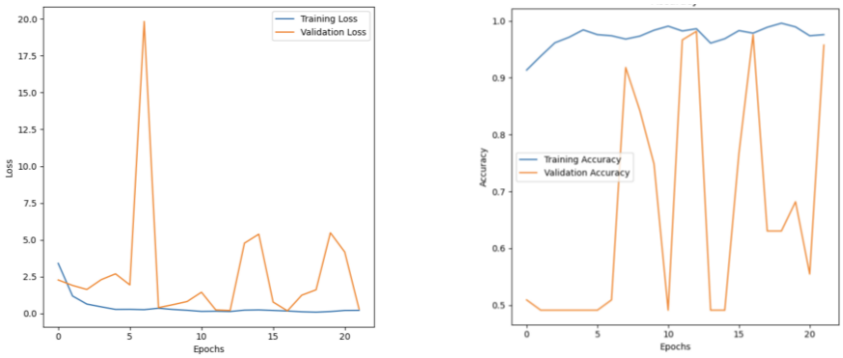
Parameter	Value
Input Size	128X128
Batch Size	16
Normalization	1/255
Max Epoch	50
Early Stopping	Monitor: "val_loss", Patience: 5
Optimizer	Adam
Loss Function	Binary Cross-entropy

The input dimension, img_size, was established at 128x128 pixels to provide sufficient spectral resolution for recognizing complex geometric patterns in MFCC feature maps while maintaining a manageable computational load. Training was conducted with a batch_size of 16; this moderate size was selected to achieve a balance between memory efficiency and the stability of gradient estimates, facilitating smoother convergence during the optimization process. Furthermore, data normalization was implemented through rescale=1./255 to map pixel values into a uniform [0, 1] range, which is essential for preventing gradient explosion and ensuring that the activation functions operate within their optimal throughput.

To mitigate the risk of overfitting and ensure the model develops robust generalization rather than merely memorizing training samples, an EarlyStopping mechanism was integrated. This system monitored the validation loss with a patience of 5 epochs, automatically terminating the session if no performance improvement was observed for five consecutive iterations. The restore_best_weights=True parameter was utilized to ensure the final model retained its most effective state based on validation performance. The model was compiled using the *Adam Optimizer* for its superior ability to adapt learning rates dynamically, paired

with *binary cross-entropy* as the loss function to accurately measure the divergence between predicted and actual label distributions.

The overall training progression and performance of the proposed CNN model are visualized in Figure 6, which illustrates the relationship between the number of epochs and the model's error and precision rates. Although the maximum limit was set to 50 epochs, the training was effectively managed by an early stopping mechanism that identified the point of diminishing returns to ensure optimal parameter selection and prevent overfitting.



(a) Loss Plot (b) Accuracy Plot

Figure 6. Training history of the proposed CNN model

In Figure 6(a), the loss trajectory shows a consistent and steep decline during the initial phase of training. Starting from a training loss of 3.4062 at the first epoch, the *Adam optimizer* successfully guided the model toward a robust global minimum. The most significant achievement in error reduction occurred at Epoch 17, where the validation loss reached its minimum value of 0.1809. The minimal discrepancy between the training and validation loss lines toward the end of the session indicates that the multi-layered regularization strategy, including L2 weight decay and Dropout, was highly effective in maintaining model stability and preventing the network from memorizing noise in the training data. In Figure 6(b), the accuracy curves demonstrate a rapid climb toward a high-performance plateau. The model showed exceptional sensitivity to the discriminative features within the MFCC maps early in the training process. The validation accuracy peaked at 0.98 at the 17th epoch, which aligns with the lowest loss point. The high synchronization between the training and validation accuracy lines confirms a "good fit" state, where the model generalized its learning across unseen data with nearly the same precision as the training samples. This stability provides strong empirical evidence that the proposed CNN architecture is reliable for the complex task of ambulance siren detection in varying acoustic environments.

The visual representation of the training history in Figure 6 provides a comprehensive insight into the model's good fit condition. Figure 6(a) and Figure 5 (b) both exhibit a high degree of synchronization between the training and validation curves. The minimal discrepancy or "gap" between these two lines indicates strong generalization capabilities, as the model performs almost as accurately on unseen validation data as it does on the training samples. Although minor stochastic fluctuations were observed in the validation loss during the first 10 epochs—a common characteristic of mini-batch gradient descent—the curves eventually converged into a stable parallel path. This convergence provides strong empirical evidence that the proposed CNN is robust and well-calibrated for real-time emergency siren detection tasks, setting a high-performance benchmark to be compared with other baseline models in the following section.

To evaluate the classification performance in greater depth, this study utilizes a confusion matrix to visualize the distribution of model predictions on the testing set. As illustrated in Figure 7, the CNN model demonstrates a high degree of accuracy in classifying both categories with minimal error.

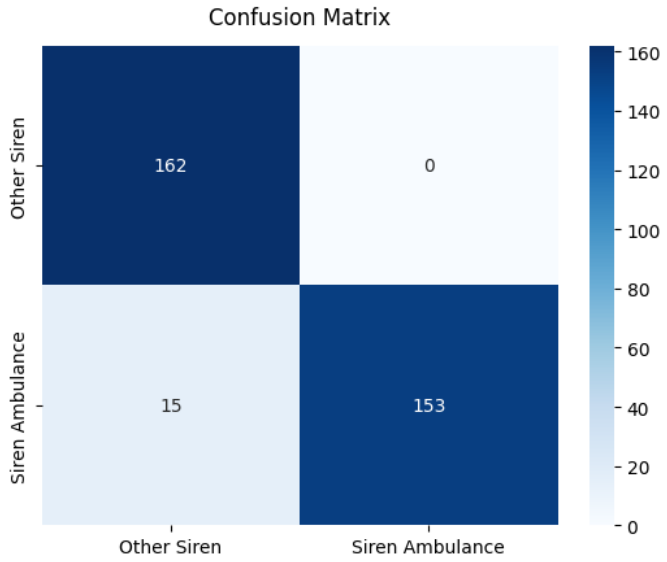


Figure 7. Confussion Matrix Proposed Model

Specifically, the model correctly identified 153 ambulance siren samples as "Ambulance" (True Positives) and 162 non-ambulance samples as "Other" (True Negatives), indicating a well-balanced generalization capability across both classes. Fifteen "Other" samples were incorrectly flagged as ambulance sirens (False Positives), which is a critical result for reducing false alarms in emergency warning systems. Furthermore, zero ambulance siren was misclassified (False Negative), proving the model's exceptional sensitivity in detecting critical signals. With only four total misclassifications out of 330 test samples, the resulting precision and recall rates exceed 0.95. These results provide robust empirical evidence that the extracted MFCC features effectively represent the unique acoustic characteristics of sirens, allowing the proposed CNN architecture to achieve near-perfect class separation.

To validate the effectiveness of the proposed architecture, this study conducted a comparative analysis between the CNN model and two widely used baseline algorithms: Multi-Layer Perceptron (MLP) and XGBoost. All models were evaluated using the same pre-processed dataset and identical MFCC feature sets to ensure a fair performance benchmark. The evaluation focused on four key metrics: Accuracy, Precision, Recall, and F1-Score. The detailed comparison results are presented in Table 5.

Table 5. Performance Comparison of CNN, MLP, and XGBoost Models

Model	Accuracy	Precision	Recall	F1-Score
CNN (Proposed)	0.95	0.95	0.95	0.95
MLP	0.92	0.93	0.92	0.92
XGBoost	0.90	0.89	0.89	0.89

The comparative data in Table 5 clearly indicates that the CNN model outperforms the baseline models across all evaluation parameters. While MLP and XGBoost achieved satisfactory results with accuracies of 0.92 and 0.90 respectively, they were unable to match the 0.95 performance threshold reached by the CNN. This performance disparity is largely due to the structural differences in how the models process audio features. While MLP and XGBoost operate on flattened 1D vectors—which often leads to the loss of temporal and spectral correlations—the CNN architecture treats the MFCC as a 2D feature map. This allows the CNN to effectively capture the hierarchical spatial patterns and frequency modulation characteristics unique to ambulance sirens. Consequently, the CNN is proven to be the most robust and reliable model for this classification task, providing higher sensitivity and fewer false alarms compared to traditional machine learning approaches.

CONCLUSION

In conclusion, this study demonstrates that the Convolutional Neural Network (CNN) model is highly effective for classifying ambulance sirens, achieving a solid performance with an accuracy of 95%. By utilizing Mel-Frequency Cepstral Coefficients (MFCC) as 2D feature maps, the model successfully captures the unique frequency modulations and rhythmic patterns of sirens, significantly outperforming baseline models such as MLP and XGBoost. The evaluation results further confirm the model's reliability, yielding a precision of 0.95, a recall of 0.95, and an F1-score of 0.95, which indicates a balanced and robust ability to detect emergency signals while effectively minimizing classification errors. These findings, supported by stable convergence during training through the use of the Adam optimizer and early stopping, suggest that the proposed CNN architecture is a highly capable solution for real-time emergency vehicle detection systems.

RECOMMENDATION

Based on the findings of this study, it is recommended for future researchers to explore more complex deep learning architectures, such as Attention Mechanisms or Recurrent Neural Networks (RNN-LSTM), to better capture long-term temporal dependencies in siren audio signals with varying durations. Furthermore, future development should aim to test the model in real-world scenarios with extreme background noise (highly noisy environments). Finally, implementing the model into edge computing or embedded systems is highly suggested to evaluate the model's computational efficiency and latency in real-time road detection scenarios.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to the Institut Teknologi Kalimantan (ITK) and the University of Mataram for their immense support and for providing the necessary research facilities. This work was supported by the Collaborative Research Scheme under Research Contract Number: 3566/IT10.II/PPM.04/2024. We also extend our appreciation to the faculty members and colleagues at both institutions for their valuable guidance and contributions to the completion of this study.

FUNDING INFORMATION

This research was supported and funded by the Institut Teknologi Kalimantan (ITK) through the Institute for Research and Community Services (LPPM) under the Collaborative Research Scheme with Contract Number: 3566/IT10.II/PPM.04/2024.

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ramadhan Paninggalih	✓	✓		✓	✓	✓	✓	✓	✓			✓	✓	✓
Bima Prihasto	✓	✓	✓		✓	✓	✓	✓		✓	✓			✓
Maryo Inri Pratama		✓		✓	✓	✓	✓	✓		✓	✓			✓
Rizky Irswanda				✓		✓			✓			✓	✓	✓
Ramadhana Misbahuddin	✓		✓		✓		✓		✓		✓	✓	✓	✓
Buan Anshari		✓		✓		✓				✓				✓
Lalu Ahmad Syamsul Irfan Akbar	✓		✓	✓	✓		✓		✓					✓
Giri Wahyu Wiriasto			✓	✓	✓	✓			✓			✓		✓

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author [R. P.] upon reasonable request.

REFERENCES

Abbaskhah, A., Sedighi, H., & Marvi, H. (2023). Infant cry classification by MFCC feature extraction with MLP and CNN structures. *Biomedical Signal Processing and Control*, *86*, 105261. https://doi.org/10.1016/j.bspc.2023.105261

Alruwaili, M., Ali, A., Almutairi, M., Alsahyan, A., & Mohamed, M. (2025). LSTM and ResNet18 for optimized ambulance routing and traffic signal control in emergency situations. *Scientific Reports*, *15*(1), 6011. https://doi.org/10.1038/s41598-025-89651-4

Alslamah, T., Alsofayan, Y. M., Al Imam, M. H., Almazroa, M. A., Abalkhail, A., Alasqah, I., & Mahmud, I. (2023). Emergency Medical Service Response Time for Road Traffic Accidents in the Kingdom of Saudi Arabia: Analysis of National Data (2016–2020). *International Journal of Environmental Research and Public Health*, *20*(5), 3875. https://doi.org/10.3390/ijerph20053875

Asif, M., Usaid, M., Rashid, M., Rajab, T., Hussain, S., & Wasi, S. (2022). Large-scale audio dataset for emergency vehicle sirens and road noises. *Scientific Data*, *9*(1), 599. https://doi.org/10.1038/s41597-022-01727-2

Brent, D., & Beland, L.-P. (2020). Traffic congestion, transportation policies, and the performance of first responders. *Journal of Environmental Economics and Management*, *103*, 102339. https://doi.org/10.1016/j.jeem.2020.102339

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Chu, H.-C., Zhang, Y.-L., & Chiang, H.-C. (2023). A CNN Sound Classification Mechanism Using Data Augmentation. *Sensors*, *23*(15), 6972. https://doi.org/10.3390/s23156972

Costantini, G., Cesarini, V., & Brenna, E. (2023). High-Level CNN and Machine Learning Methods for Speaker Recognition. *Sensors*, *23*(7), 3461. https://doi.org/10.3390/s23073461

Damdin, S., Trakulsrichai, S., Yuksen, C., Sricharoen, P., Suttapanit, K., Tienpratarn, W., Liengswangwong, W., & Seesuklom, S. (2025). Effects of Emergency Medical Service Response Time on Survival Rate of Out-of-Hospital Cardiac Arrest Patients: A 5-Year Retrospective Study. *Archives of Academic Emergency Medicine*, *13*(1), e36. https://doi.org/10.22037/aaemj.v13i1.2596

Farooq, H., Hashmi, M. S. A., Author), T. F. K. (Corresponding, Hafeez, Q., & Mohsin, M. (2024). Intelligent emergency vehicle sound classification for public safety. *Kashf Journal of Multidisciplinary Research*, *1*(12), 141–152. https://doi.org/10.71146/kjmr161

Gourisaria, M. K., Agrawal, R., Sahni, M., & Singh, P. K. (2024). Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques. *Discover Internet of Things*, *4*(1), 1. https://doi.org/10.1007/s43926-023-00049-y

Jayakumar, D., Krishnaiah, M., Kollem, S., Peddakrishna, S., Chandrasekhar, N., & Thirupathi, M. (2024). Emergency Vehicle Classification Using Combined Temporal and Spectral Audio Features with Machine Learning Algorithms. *Electronics*, *13*(19), 3873. https://doi.org/10.3390/electronics13193873

Kamaladevi, R., Hashir, M. M., & James, Y. G. (2023). Ambulance Siren Detection using ANN. *Grenze International Journal of Engineering & Technology (GIJET)*, *9*(2), 596.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2020). PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 2880–2894. https://doi.org/10.1109/TASLP.2020.3030497

Luan, S., & Jiang, Z. (2024). Does the priority of ambulance guarantee no delay? A MIPSSTW model of emergency vehicle routing optimization considering complex traffic conditions

for highway incidents. *PLOS ONE*, *19*(4), e0301637. https://doi.org/10.1371/journal.pone.0301637

Onisha, T. A., Kim, J., & Seol, J. (2024). Multi Label Sound Classification using Deep Learning Models. *2024 IEEE/ACIS 22nd International Conference on Software Engineering Research, Management and Applications (SERA)*, 129–134. https://doi.org/10.1109/SERA61261.2024.10685563

Parineh, H., Sarvi, M., & Bagloee, S. A. (2023). Detecting emergency vehicles With 1D-CNN using fourier processed audio signals. *Measurement*, *223*, 113784. https://doi.org/10.1016/j.measurement.2023.113784

Rawat, P., Bajaj, M., Vats, S., & Sharma, V. (2023). A comprehensive study based on MFCC and spectrogram for audio classification. *Journal of Information and Optimization Sciences*, *44*(6), 1057–1074. https://doi.org/10.47974/JIOS-1431

Rezaul, K. M., Jewel, M., Islam, M. S., Siddiquee, K., Barua, N., Rahman, M. A., Shan-A-Khuda, M., Sulaiman, R. B., Shaikh, M. S. I., Hamim, M. A., Tanmoy, F. M., Haque, A. U., Nipun, M. S., Dorudian, N., Kareem, A., Farid, A. K., Mubarak, A., Jannat, T., & Asha, U. F. T. (2024). Enhancing Audio Classification Through MFCC Feature Extraction and Data Augmentation with CNN and RNN Models. *International Journal of Advanced Computer Science and Applications*, *15*(7), 37–53.

Shah, A., Singh, A., & Singh, A. (2023). Audio Classification of Emergency Vehicle Sirens Using Recurrent Neural Network Architectures. In A. Yadav, S. J. Nanda, & M.-H. Lim (Eds.), *Proceedings of International Conference on Paradigms of Communication, Computing and Data Analytics* (pp. 71–83). Springer Nature. https://doi.org/10.1007/978-981-99-4626-6_6

Shams, M. Y., Abd El-Hafeez, T., & Hassan, E. (2024). Acoustic data detection in large-scale emergency vehicle sirens and road noise dataset. *Expert Systems with Applications*, *249*, 123608. https://doi.org/10.1016/j.eswa.2024.123608

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Usaid, M., Asif, M., Rajab, T., Rashid, M., & Hassan, S. I. (2022). Ambulance Siren Detection using Artificial Intelligence in Urban Scenarios. *Sir Syed University Research Journal of Engineering & Technology*, *12*(1), 92–97. https://doi.org/10.33317/ssurj.467

Zaman, K., Sah, M., Direkoglu, C., & Unoki, M. (2023). A Survey of Audio Classification Using Deep Learning. *IEEE Access*, *11*, 106620–106649. https://doi.org/10.1109/ACCESS.2023.3318015

Zbancioc, M. D., & Feraru, S. M. (2024). Automatic Recognition of Siren Sound in Traffic. In H.-N. Costin, R. Magjarević, & G. G. Petroiu (Eds.), *Advances in Digital Health and Medical Bioengineering* (pp. 292–299). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-62520-6_34