Journal of Network Intelligence ©2025 ISSN 2414-8105 (Online)
Taiwan Ubiquitous Information Volume **10**, Number **1**, February **2025**

TFANet: Vehicle Sound-based Classification Utilizing Time-Frequency Hybrid Attention

Bo Jiang*, Hong-Bin Li, Yue-Lin Xu, Kai Ding
Science and Technology on Near-Surface Detection Laboratory, Wuxi 214035, China
jb668899@163.com, lhb198510@163.com, 345179186@qq.com, dingkai66683@163.com

Xu-Dong Zhao
Jiangsu Puxu Software Information Technology Co., Ltd, Nanjing 210001, China
303423430@qq.com

*Corresponding author: Bo Jiang

Received February 2, 2024, revised May 16, 2024, accepted June 13, 2024.

Abstract. *Vehicle sound-based classification is essential for preserving the natural environment and intelligent transportation, helping monitor device operation, maintain environmental safety, and regulate vehicle traffic. Nevertheless, the intricate nature of outdoor situations, combined with elements like wind noise, brings substantial extraneous or disruptive information into the data, which has a detrimental impact on classification performance. To tackle this issue, this study introduces a vehicle classification network called the Time-Frequency Hybrid Attention Mechanism Classification Network(TFANet). Firstly this network is built on the mature ResNet18 architecture and utilizes Mel Frequency Cepstral Coefficients (MFCC) features to analyze sound signals, significantly improving the accuracy and robustness of the classification model. In addition, it designs a time-frequency hybrid attention mechanism to assign more importance to the semantic relevant time frames and key frequency bands within the MFCCs spectrogram, helping to minimize the impact of irrelevant information and ensure the accuracy of vehicle classification. Finally, we collected and organized the Field Vehicle Sound Dataset (FVSD), and conducted relevant experiments on this dataset, demonstrating that the approach achieved an F1 score of 0.950. Meanwhile, additional experiments were performed using the IDMT-Traffic open-access dataset. The TFANet showed an improved F1 score in each category compared to the baseline model of this dataset. Notably, the Truck category, which had the poorest classification performance, significantly improved by 0.16 in the F1 score.*

Keywords: MFCC, Attention mechanism, Signal processing, Vehicle classification

1. **Introduction.** Due to social and economic development, motor vehicles are rapidly spreading. At the same time, the rapid increase in the number of vehicles on the road has also made the traffic situation more complicated, and there will be many traffic problems, such as traffic accidents and road congestion [1]. Vehicle classification and recognition technologies play a crucial role in the effectiveness of intelligent transportation systems. They are essential for monitoring equipment operations, maintaining environmental safety, and regulating vehicular traffic [2]. Categorization of vehicles in natural field areas is essential for detecting possible dangers and plays a critical role in strengthening data-driven support for environmental conservation actions [3]. This paper explores the complexities of these technologies, emphasizing their significance in modern urban and environmental planning. Through an analysis of the implementation and effects of these systems, we offer

a thorough summary of their role in promoting sustainable transportation infrastructure and environmental conservation.

In terrestrial transportation systems, the detection of vehicles has traditionally been anchored in video sensors and image processing methodologies [4] [5]. Nevertheless, these systems that rely on images require the accurate positioning of cameras along the roads, along with the requirement of a clear and unobstructed vision. Moreover, the identification of vehicle types beyond the line of sight presents considerable challenges when employing this approach [6]. Instead, the use of acoustic communication presents itself as an appealing solution, eliminating the need for extra equipment at both the sending and receiving ends [7]. This work explores the field of vehicle model recognition using acoustic signals. The method involves placing sound collection devices strategically along roadways to catch the sounds emitted by passing automobiles at specific sampling frequencies. The analytical procedure involves the utilization of acoustic analysis, pattern recognition techniques, and a systematic approach that includes reducing data volume, handling processing requirements, and ensuring economic feasibility. These characteristics lead to its increasing adoption and acknowledgment in both national and international contexts [8] [9].

The conventional approach to in-vehicle sound classification and identification generally involves three main steps: capturing vehicle noises, extracting significant characteristics from these sounds, and then classifying these characteristics. Traditionally, research in this field has been focused on using a single characteristic signal, such as a particular vibration or sound, to identify a target vehicle. This computationally efficient method is most suitable for circumstances where vehicle sound data is largely free of noise and can be described by different dataset features [10]. Sharma et al. [11] utilized wavelet transform and spectral statistics to examine and categorize vehicle vibration signals in their study. Aljaafreh et al. [12] utilized the short-time Fourier transform in conjunction with power spectral energy analysis on time-frequency domain data to extract characteristics from vehicle acoustic sounds. They then employed a support vector machine for the classification process. Yang et al. [13] utilized discrete spectrum analysis to extract features from vehicle sounds and devised an algorithm based on wireless sensor network protocols for vehicle classification and recognition. Nevertheless, these strategies for extracting features manually are usually limited by their shallow level of information processing, which lacks the profound abstraction capabilities required to successfully distinguish unique vehicle characteristics [14].

The progress made in deep learning has led to a significant change in sound classification methods. This is evident from the increase in academic research that utilizes neural network models for various acoustic classification problems. These encompass a variety of tasks, such as classifying environmental sounds [15] [16], linguistic emotion discernment [17], and classifying bioacoustic signals [18] [19], among others. Simultaneously, there have been notable advancements in the field of vehicle categorization. Mohine et al. [20] were the first to develop a hybrid model that combines a Convolutional Neural Network with a Bidirectional Long Short-Term Memory Model (CNN-BiLSTM). This proficient model excels in extracting distinctive characteristics from audio signals and classifying them into five unique categories (two-wheelers, low, medium, heavy vehicles, and noise), hence improving the accuracy of vehicle classification. Ashhad et al. [21] proposed a convolutional neural network (CNN) framework to preprocess acoustic data characteristics. They utilized MFCC to obtain a fourfold classification on the IDMT-Traffic dataset. Chen et al. [22] presented a new hybrid neural network classifier that combines Long Short-Term Memory (LSTM) with CNN layers. The integration of MFCC, Pitch

Class Profile (PCP), and Short-Term Energy (STE) properties serves as inputs for advanced modeling. Sun et al. [23] devised an intra-frame network and fusion technique that utilizes AlexNet in conjunction with LSTM to extract feature vectors from signals for extensive vehicle classification tasks. Mohine et al. [24] developed an architecture for vehicle classification using the Fast Fourier Transform (FFT). The architecture identifies strong features that are appropriate for mobile vehicle recognition systems based on acoustic modality. In their study, Luo et al. [25] introduced a novel approach called Sound Convolutional Recurrent Neural Network (S-CRNN), which combines Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) techniques. In addition, Wang et al. [26] applied spectral augmentation techniques to Mel spectrograms before extracting MFCC, thereby enhancing the resilience of the system. The utilization of deeply divisible convolution in the CNN network enhanced the efficiency of the vehicle detection system by enabling model compression. Ashhad et al. [27] developed an advanced approach for recognizing vehicle sounds. This approach combines global and local audio characteristics in a multi-input neural network structure, facilitating sound-based vehicle classification tasks.

While the previously mentioned methodologies have exhibited certain advancements in vehicle sound classification, they often overlook the substantial impact of silent segments and noise elements within sound signals on classification accuracy. These silent segments and noise elements are notably prevalent in sound data acquired from natural environments, particularly in scenarios characterized by complex conditions such as heightened wind and noise interferences, where their prominence is exacerbated. These factors present formidable challenges in the processing and feature extraction of sound signals, potentially obfuscating genuine vehicle characteristics and thereby compromising classifier efficacy. Consequently, the effective handling and filtration of these silent segments and noise elements emerge as pivotal issues confronting contemporary research in vehicle sound classification.

Eliminating extraneous signals poses a multifaceted challenge often disregarded in conventional methodologies [28]. To surmount this obstacle, rigorous field data collection was conducted. This meticulous endeavor encompassed the comprehensive gathering of data spanning prominent vehicle categories to ensure dataset completeness and representativeness. Subsequently, a meticulous organization and in-depth analysis of the acquired data culminated in the establishment of the Field Vehicle Sound Dataset (FVSD). The inception of FVSD facilitated advanced vehicle classification experiments in more demanding natural settings, thereby validating the feasibility and efficacy of our approach. To counteract the deleterious impact of irrelevant signals on classification accuracy, we devised a novel vehicle classification framework named TFANet. Leveraging a time-frequency attention mechanism, TFANet adeptly discerns auditory data, selectively amplifying features bearing pivotal information while concurrently attenuating the significance of less pertinent features. Through this strategy, TFANet achieves enhanced precision in identifying and classifying vehicle sounds, thereby augmenting the performance and resilience of the classification system.

Contribution. The notable contributions of this study are summarized as follows:

(1) This paper introduces the TFANet, a new network designed specifically for classifying vehicle sounds. TFANet utilizes MFCC as its primary feature, due to its strong immunity to interference. By doing a thorough investigation and careful comparison of several classifiers, the method incorporates ResNet18 as its classification model. This combination notably enhances the accuracy and robustness of the sound-based vehicle type classification.

(2) TFANet uses a novel time-frequency hybrid attention mechanism to reduce the negative impact of non-essential or interfering data in audio on categorization performance. This module uses the attention mechanism in the time domain drinking frequency domain direction of MFCCs spectrogram respectively, adaptively assigns more weights to the effective features with significant information, and reduces the interference of irrelevant information. As a result, it greatly enhances the network's ability to accurately identify and distinguish the unique sound patterns associated with different types of vehicles.

(3) We established the FVSD dataset after thorough organization and analysis of the collected data in a real-world setting. To verify the efficacy of TFANet, we conducted a rigorous experimental evaluation of the FVSD dataset. The results were compelling, with TFANet achieving a performance level exceeding 0.95 for the metrics of the F1 score. Furthermore, experimental evaluations were conducted using the publicly available IDMT-Traffic dataset, The TFANet model showed improved F1 score in each category compared to the baseline model. This confirms without any doubt that our method is highly effective in practice.

1.1. **Organization.** The rest of this paper is structured as follows. We present the datasets, reparations, and the methodology proposed in this paper in Section 2. In Section 3 a variety of experiments are designed to evaluate the effectiveness of the methods in this paper. In Section 4, we summarize this paper.

2. Materials and Method.

2.1. Datasets.

2.1.1. *Field Vehicle Sound Dataset.* A comprehensive experimental data collection campaign was carried out in a natural field setting, as shown in Figure 1. This entailed vehicles traversing a specifically designated 500-meter straight section in the field. The data collection process was carried out using an advanced vibroacoustic signal collection device. This device had an array aperture of one meter and a height of 30 centimeters. It was designed to simultaneously capture both auditory and vibrational signals. As shown in Table 1. The resulting dataset includes a wide range of vehicle types, such as jeep, coupe, truck, and off-road vehicles. The FVSD is a comprehensive dataset consisting of 203 individual data entries. Each entry has an average duration of 67.9 seconds. The data structure is partitioned into 6 channels per entry, where channels 1 to 4 are specifically designated for sound signal data, and channels 5 and 6 are assigned for vibration signal data. All signals are captured at a high-fidelity sampling rate of 10240Hz.

For real-time vehicle recognition in natural field conditions, the dataset was divided into intervals of five seconds, considering the parameters of data volume and sampling rate. The intervals were later used as the training dataset.

Table 1. The summary of the FVSD self-produced dataset.

Vehicle types	Jeep	Coupe	Truck	Off-Road Vehicles
Quantities	24	70	30	76
Sampling rate	10240Hz			
Road situation	Dirt/Asphalt roads			

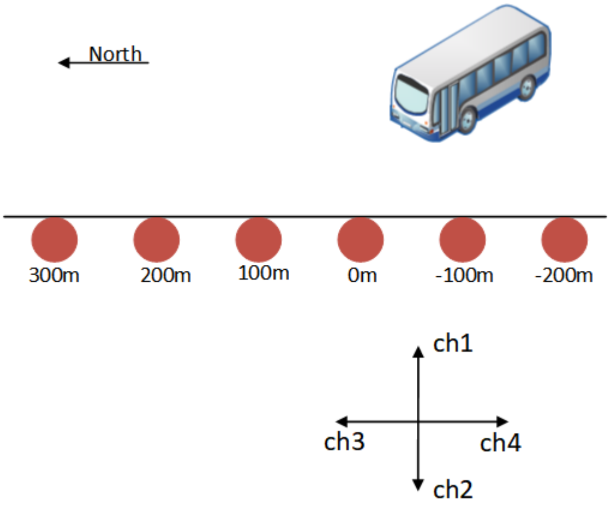

Figure 1. Schematic diagram of the vehicle information collection site

2.1.2. *IDMT-Traffic Dataset.* The IDMT-Traffic dataset [29] was introduced in 2021 as an innovative open benchmark repository designed specifically for acoustic traffic monitoring. This dataset combines a range of recordings from four different locations that are far apart geographically. It includes recordings of urban traffic situations from Ilmenau, Germany, and the surrounding urban areas, as well as recordings of a different rural road environment. The collection includes four sessions of precisely synchronized stereo recordings of vehicle movement. The recordings were made with great accuracy, using high-grade sE8 microphones in combination with cost-effective Micro-Electro-Mechanical Systems (MEMS) microphones. The dataset is extensive, encompassing a wide range of vehicular dynamics. As shown in Table 2. It includes 3903 car-related events, 511 instances of trucks, 251 occurrences of motorcycles, and additional background acoustic data recorded during periods without any vehicles present.

To enhance processing efficiency and greatly enhance analytical accuracy, the data obtained during the experimental phase was methodically divided into intervals of two seconds. This segmentation strategy was carefully implemented to guarantee a more detailed and accurate analysis.

Table 2. The summary of the IDMT-Traffic dataset.

Vehicle types	Car	Truck	Motorcycle	No vehicle
Quantities	3903	511	251	251
Sampling rate	22050Hz			
Road situation	Dry/Wet roads			

2.2. **Feature Processing.** In recent decades, numerous feature extraction methods have been extensively studied to determine their effectiveness in analyzing acoustic properties. These methods encompass a wide range of domains, such as time, frequency, short spectral, wavelet, and time-frequency domains [30]. The MFCC is a commonly used technique in sound recognition because it accurately simulates the human auditory system [31]. The core of MFCC lies in its initial conversion of the linear spectrum into the Mel nonlinear spectrum, which corresponds to auditory perception, and subsequently transforming it into the cepstral spectrum. MFCC is a versatile and signal-independent method that is not limited by any preconceived constraints or assumptions about the input sound signal, exhibiting high recognition success rates and stable performance in the field of sound recognition [32]. The MFCC model exhibits exceptional resilience and immunity to noise, which greatly enhances its performance in recognition tasks when compared to other models. This has been confirmed through numerous studies and experiments that specifically examine auditory models.

The sequential procedures required for converting input data using MFCC are illustrated in Figure 2. The process begins by applying pre-emphasis to the sound signal, which involves compensating and amplifying the high-frequency components. The process is quantitatively represented by Equation (1), demonstrating the systematic method of improving the signal for further analysis.

$$f'(n) = f(n) - \alpha \cdot f(n-1) \tag{1}$$

where $\alpha = 0.97$, $f(n)$ is the original acoustic signal and $f'(n)$ is the output signal.

After pre-emphasis, the signal is divided into discrete short-time segments through frame splitting. In this research, every individual frame is assigned a window duration of 23 milliseconds. To maintain a significant amount of overlap for the purpose of consistency and precision in analysis, the step size of the window is established as half of its length, resulting in a 50% overlap with the neighboring frame. The utilization of this overlapping technique improves the dependability of the time-window analysis.

After the segmentation process, each short-time frame is subjected to spectral analysis using the Fast Fourier transform (FFT). This essential step is pivotal in converting the time domain signal into its corresponding frequency domain spectrum. Afterwards, the Mel filter bank is used. This step is crucial as it transforms the acquired spectrograms into the Mel frequency domain, which closely emulates the way the human ear perceives different sound frequencies. The process of mapping based on auditory perception is systematically calculated according to Equation (2), thereby guaranteeing that the frequency representation corresponds to the characteristics of human auditory characteristics.

$$Mel(f) = 2595 \cdot log10(1 + f/700) \tag{2}$$

After applying the Mel filter bank, the energy contained within it undergoes logarithmic computation. This step is crucial as it primarily emphasizes the information found in the lower frequency ranges, which are typically more important in auditory analysis. Employing a logarithmic method helps to highlight nuanced yet essential elements of the sound that could otherwise be overpowered by higher-frequency components. The subsequent crucial stage entails the utilization of the Discrete Cosine Transform (DCT). This transformation is strategically utilized to decrease the correlation among the features, thus guaranteeing the preservation and highlighting of crucial feature information. Its purpose is to enhance the feature set by concentrating on the most pertinent aspects for analysis, thus optimizing the data for subsequent processing stages. The final MFCC features are obtained as the result of this complex process. These features constitute a thorough and refined dataset, prepared to be used as input for the classification network. Next, the network performs high-dimensional feature extraction, which is an essential step that sets the foundation for the subsequent classification tasks. The meticulous MFCC process greatly improves the network's capacity to precisely and effectively classify the auditory data, leading to a high level of accuracy in achieving the desired classification goals.

2.3. **Time-Frequency Hybrid Attention Mechanism.** Obtaining vehicle sound signals in real-world outdoor settings is often hindered by the existence of wind noise and

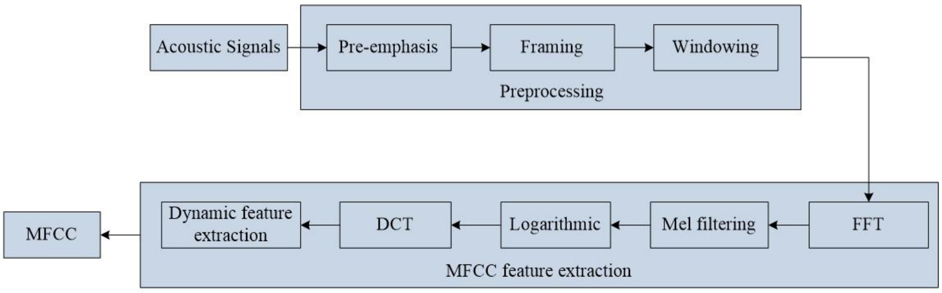

Figure 2. MFCC feature processing flow

other interference from the environment. These disruptions frequently result in the obstruction or alteration of sounds emanating from crucial vehicle elements such as engines and tires. Consequently, the sound signals' integrity is compromised, which presents substantial difficulties in categorizing vehicle sounds. To overcome these challenges, our study presents a new method called the time-frequency hybrid attention mechanism [33]. This innovative approach combines attention mechanisms in both the time and frequency domains.

In the context of time, the temporal attention mechanism is important for effectively reducing the impact of background noise on sound signal analysis by suppressing noisy or silent frames. Simultaneously, in the dimension of frequency, the frequency attention mechanism is skilled at allocating greater importance to frequency bands that contain significant discriminative information. Conversely, it diminishes the focus on less significant frequency ranges that contain restricted informational worth. The novel time-frequency hybrid attention mechanism module dynamically prioritizes significant time-frequency structures crucial for vehicle sound classification. This strategic focus enables the network to prioritize specific time periods and frequency bands that are semantically important, thereby enhancing the emphasis on valuable information while reducing the network's sensitivity to irrelevant data or noise.

Due to this two-dimensional focus, the network acquires an enhanced capability to differentiate and categorize unique sound characteristics specific to different types of vehicles. Figure 3 illustrates the essential operational steps of the time-frequency hybrid attention mechanism, offering a visual depiction of this complex process.

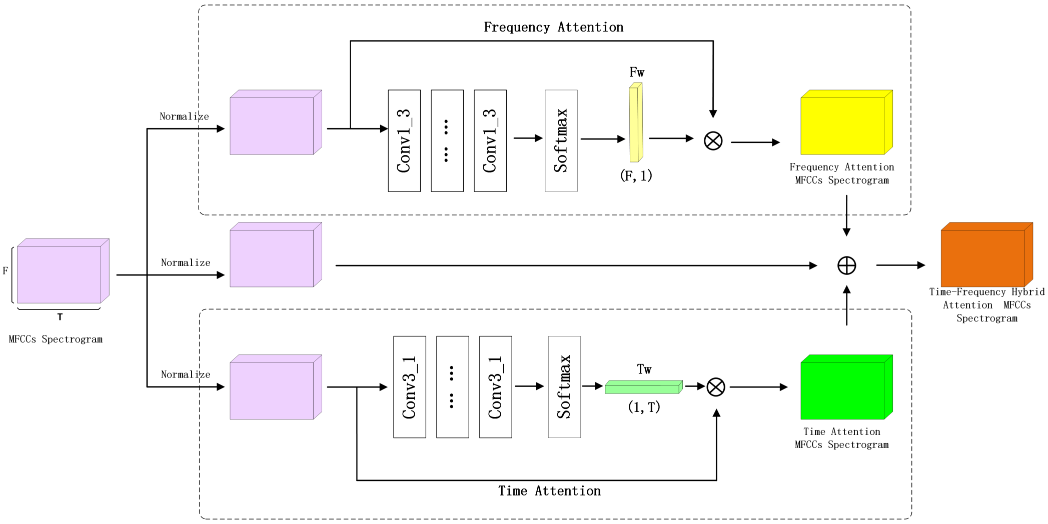

Figure 3. The generation process of Time-Frequency hybrid attention mechanisms.

1. Normalization of MFCCs spectrogram $x(f,t)$, as shown in Equation (3).

$$X(f,t) = Normalize(x(f,t)), 1 \leq f \leq F, 1 \leq t \leq T \tag{3}$$

2. The process of generating the time attention MFCC spectrogram $M_T(t)$ is depicted in the lower section of Figure 3.

(1) Conducting a convolution operation on the regularized MFCC features is accomplished through the utilization of a 3×1 convolution kernel, as illustrated in Equation (4), which iterative procedure persists until the MFCC is diminished to 1 within the frequency dimension. This process extracts the nonlinear features in the time dimension, resulting in a one-dimensional matrix of size $(1,T)$.

$$A_T = Conv3 \times 1(X(f,t)), 1 \leq t \leq T \tag{4}$$

(2) The weight of time attention is calculated utilizing Equation (5).

$$T_w(t) = \frac{exp(A_T(1,t))}{\sum_{i=1}^{T} exp(A_T(1,i))}, 1 \leq t \leq T \tag{5}$$

(3)The regularized MFCCs spectrogram is multiplied element-wise with the obtained attentional weight matrix along the temporal direction to yield a temporal attention spectrogram $M_T(t)$. The calculation method is depicted in Equation (6).

$$M_T(t) = X(f,T) * T_w, 1 \leq f \leq F \tag{6}$$

3. The upper portion of Figure 3 delineates the procedure for generating the frequency attention MFCC spectrogram $M_F(f)$.

(1) Conducting a convolution operation on the regularized MFCC features is accomplished through the utilization of a 1×3 convolution kernel, as illustrated in Equation (7), which iterative procedure persists until the MFCC is diminished to 1 within the time dimension. This process extracts the nonlinear features in the frequency dimension, resulting in a one-dimensional matrix of size $(F,1)$;

$$A_F = Conv1 \times 3(X(f,t)), 1 \leq f \leq F, \tag{7}$$

(2) As shown in Equation (8), the Softmax function is employed to calculate the temporal attention weights;

$$F_w(t) = \frac{exp(A_F(f,1))}{\sum_{j=1}^{F} exp(A_F(j,1))}, 1 \leq f \leq F \tag{8}$$

(3) The MFCCs spectrogram was multiplied element-wise in the frequency direction with the obtained attention weight matrix to yield a frequency attention MFCCs spectrogram $M\_F(f)$, as depicted in Equation (9);

$$M_F(f) = X(F,t) * F_w, 1 \leq t \leq T \tag{9}$$

4. Generation of time-frequency hybrid attention MFCCs spectrogram M_{T-F}. The attentional spectrograms are skillfully combined with the original MFCCs spectrograms, as described in Equation (10), resulting in the formation of sophisticated time-frequency hybrid attentional spectrograms. This novel method greatly enhances the ability to distinguish different characteristics of the acoustic signal in the MFCCs spectrogram analysis.

$$M_{T-F} = M_T + M_F + X(f,t) \tag{10}$$

2.4. **Classification Network.** In this paper, we utilize the Residual Network (ResNet) [34] as our classification network. The specific parameters for each layer in the ResNet framework are outlined in Table 3 with detailed specifications. The model's architecture consists of four two-layer residual modules, with each module utilizing a kernel size of (3 × 3). The modules are configured with channel numbers 64, 128, 256, and 512, respectively. Our implementation incorporates the residual structure, which accelerates the learning process and improves the network's trainability. This design has demonstrated its effectiveness in tackling prevalent obstacles in deep learning, such as the issues of gradient explosion and vanishing. Each residual block in the architecture is meticulously crafted to address potential challenges, such as diminishing the size of the feature map or modifying the number of channels in the main path. If these changes are not dealt with, they could result in the loss of information or a decrease in the performance of the model. In order to reduce these risks, each residual block includes a skip connection that directly connects its input and output. As shown in Figure 4, these skip connections facilitate the seamless integration of input data onto the output of specific layers within the network. This approach guarantees a smoother transmission of feature information across the network. Importantly, it maintains the complex details and semantic depth present in the original input data, making it easier to pass on to subsequent layers without losing important information. In this study, we employ an 18-layer ResNet network, which provides an ideal trade-off between computational complexity and classification accuracy. The network is configured to process the dataset using MFCC features.

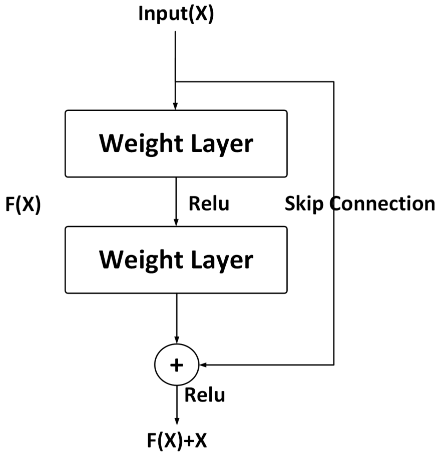

Figure 4. Skip connection in the residual network

2.5. **Our Method.** The TFANet is introduced in this paper. The overall architecture of TFANet is depicted in Figure 5. It comprises three main components: a section dedicated to feature processing, a section implementing a time-frequency hybrid attention mechanism, and a section functioning as a classifier. As vehicle sound signals in natural field environments are significantly impacted by noise, MFCC was chosen as the base feature during the feature processing phase on account of its comparatively low sensitivity to noise interference. However, noise interference cannot be eliminated by relying solely on MFCC feature extraction, given the variety of noise sources present in field vehicle sounds. A time-frequency hybrid attention mechanism has been implemented to address this issue. By utilizing this mechanism, critical information in the time and frequency domains is effectively identified and highlighted during the vehicle sound classification procedure. Decreasing the feature weights of extraneous or disruptive data effectively mitigates the

TABLE 3. Description of the different layers of residual network used in the proposed method.

Layer Name	Description
Input Layer	H=Height, W=Width,C=Channels
Convolutional Layer_1	Kernel Size=7×7, Stride=2,Channel=64
Max Pooling Layer	Kernel Size=3×3,Stride=2
Convolutional Layer_2 ×2	Kernel Size=3×3, Channel=64
Convolutional Layer_3 ×2	Kernel Size=3×3, Stride=2, Channel=128
Convolutional Layer_4 ×2	Kernel Size=3×3, Stride=2, Channel=256
Convolutional Layer_5 ×2	Kernel Size=3×3, Stride=2, Channel=512
Average pooling	Kernel Size=7×7
Fully Connected Layer	
Softmax Layer	

influence of noise on the accuracy of classification and improves the network's overall performance. ResNet18 is chosen as the foundational architecture for classification purposes, owing to the capability of its residual block to capture comprehensive contextual information present in the feature map. This functionality guarantees the retrieval of exhaustive and representative characteristics to classify vehicle sounds. The TFANet architecture significantly improves the accuracy of vehicle type classification under intricate environmental circumstances.

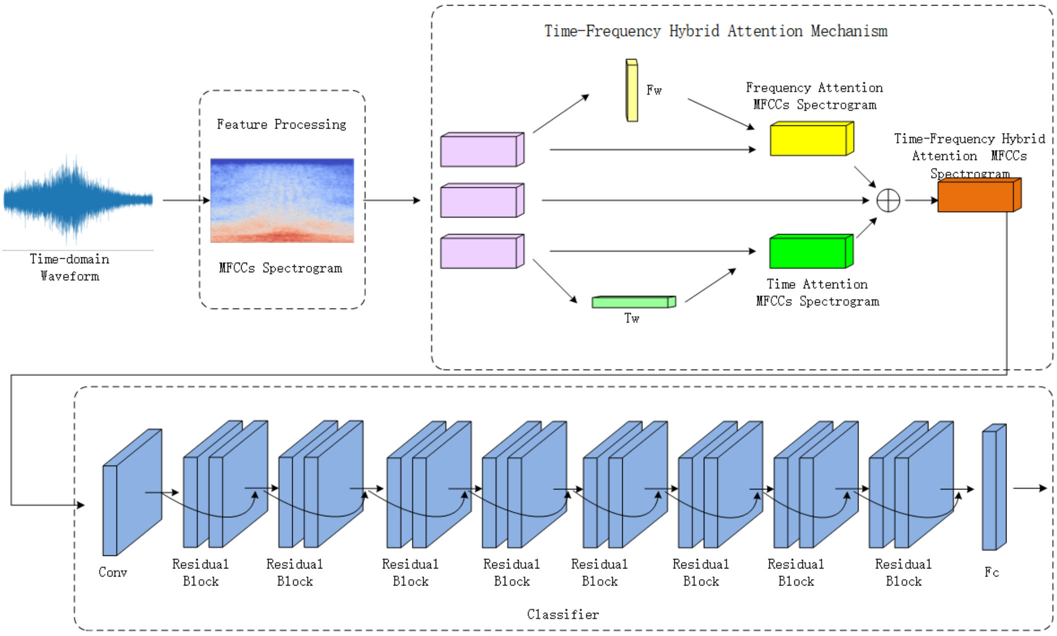

FIGURE 5. The overall architecture of the proposed TFANet classification model.

2.6. **Performance Evaluation.** The validation performance is assessed in terms of F1-score, Accuracy, Precision, and Recall. The quantitative definitions of each of these metrics are given in Equations (11) through (14), where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{13}$$

$$F1\ score = 2 \times (\frac{Precision \times Recall}{Precision + Recall}) \times 100\% \tag{14}$$

The accuracy metric is of utmost importance as it represents the overall effectiveness of the model in accurately categorizing samples as either positive or negative. Accuracy alone may not provide a comprehensive evaluation of the model's performance in situations involving class imbalance; therefore, for a more nuanced assessment, Precision, Recall, and F1 score should be employed in addition to Accuracy. Precision indicates the model's ability to discriminate between negative samples; recall indicates the model's ability to identify positive samples (and hence the higher the Precision, the stronger the model's capacity to discriminate between negative samples); and F1 score is a combination of the two; the higher the F1 score, the more robust the model.

3. Result.

In order to verify the efficacy of TFANet, experiments are carried out in this section using both the FVSD collected in the laboratory field and the publicly available IDMT-Traffic datasets.

3.1. Comparison of Different Classifiers.

Due to the high occurrence of significant background noise that accompanies vehicle sounds in real-world environments, we chose to use the MFCC as the input for our classification network. Subsequently, a sequence of carefully planned experiments was carried out to assess and contrast the effectiveness of different classifiers in these circumstances. The chosen classifiers for evaluation comprise Deep Neural Networks (DNN), 1-Dimensional Convolutional Neural Networks (CNN-1D), Long Short Term Memory Networks (LSTM) [35], AlexNet [36], and ResNet. Each of these classifiers has previously shown substantial effectiveness in their respective domains. The variations in performance when processing sound data can be attributed to the differences in architecture and processing mechanisms of the individual classifiers. The CNN-1D is believed to be highly skilled at extracting local time-frequency features, whereas the LSTM is anticipated to excel in capturing the inherent temporal dynamics in sound data. Similarly, AlexNet and ResNet, being sophisticated models for image classification, may exhibit benefits in feature acquisition and generalization. The results of the experiment, which involved classifying vehicle types based on sound using various classifiers, are presented in Table 4.

Table 4. The results of different classifiers.

Classifier	FVSD			IDMT-Traffic		
	Precision	Recall	F1 score	Precision	Recall	F1 score
DNN	0.814	0.809	0.810	0.875	0.932	0.901
CNN-1D	0.856	0.872	0.863	0.944	0.945	0.933
LSTM	0.854	0.841	0.844	0.801	0.882	0.837
AlexNet	0.839	0.825	0.830	0.874	0.931	0.800
ResNet18	0.930	0.930	0.929	0.947	0.950	0.941

We thoroughly evaluated various classifiers for vehicle sound detection using Precision, Recall, and F1 score as evaluation metrics. DNN showed the least impressive performance in FVSD due to its simple structure, limiting its capability to handle complex features.

CNN-1D and LSTM performed comparably well in handling temporal data but struggled with utilizing frequency domain information from MFCC features effectively. In contrast, ResNet18 outperformed other models with Precision, Recall, and F1 scores exceeding 0.90, attributed to its deeper architecture and distinctive skip connections, enhancing accuracy and generalization, particularly in noisy environments. In IDMT-Traffic, ResNet18 achieved the highest F1 score of 0.941, affirming its effectiveness in accurately detecting vehicle types based on sound. We attempted to enhance the classification performance by augmenting the number of layers in the ResNet. The outcomes of the experiment are presented in Table 5.

Table 5. The results of different layers of ResNet.

Classifier	FVSD			IDMT-Traffic		
	Precision	Recall	F1 score	Precision	Recall	F1 score
ResNet18	0.930	0.930	0.929	0.947	0.950	0.941
ResNet34	0.933	0.931	0.927	0.941	0.945	0.932
ResNet50	0.919	0.916	0.917	0.947	0.944	0.930

The results in Table 5 reveal an unexpected trend, the F1 score did not exhibit the anticipated improvement when the model's number of layers increased from 18 to 50, for both datasets. The potential cause for this lack of satisfactory performance could be attributed to overfitting or difficulties in optimizing the network due to the increased complexity of the model. In addition, the 18-layer network is computationally more efficient due to its lower complexity, which is a result of having fewer parameters. This aspect is especially vital during the training process. Therefore, considering these results, ResNet18 has been chosen as the main classification network for this research.

3.2. **Comparison of Attention Mechanism.** This section performed a series of comparative experiments to determine the impact of the recently introduced time-frequency hybrid attention mechanism on vehicle-type classification tasks. The analyses were formulated by the methodology and results outlined in the previous section. The ResNet18 classifier was used, with MFCC as input features. The results of these experiments are methodically displayed in Table 6 T-Attention, where the attention mechanism is exclusively applied in the time dimension; F-Attention, indicating the inclusion of the attention mechanism in the frequency domain; and T-F Attention, representing the implementation of the attention mechanism across both time and frequency dimensions.

Table 6. The results of the attention mechanism.

Strategy	FVSD			IDMT-Traffic		
	Precision	Recall	F1 score	Precision	Recall	F1 score
No attention	0.930	0.930	0.929	0.947	0.950	0.941
T-Attention	0.937	0.942	0.933	0.947	0.951	0.946
F-Attention	0.940	0.949	0.936	0.946	0.951	0.946
T-F Attention	0.951	0.951	0.950	0.954	0.957	0.955

The FVSD achieved a notable baseline performance using ResNet18 and MFCC features, attaining an F1 score of 0.929 without employing any attention mechanism. Incorporating the temporal attention mechanism yielded a slight performance enhancement, resulting in a 0.004 increase in F1 score. This suggests that by prioritizing temporal aspects in sound signals, the model becomes more adept at identifying critical moments in the sound. Moreover, the introduction of the frequency attention mechanism led to further performance improvement, indicating the model's enhanced ability to distinguish vehicle sounds from background noise with greater precision. Specifically, it effectively attenuates noise frequency bands while focusing attention on relevant sound frequency bands. Integration of the time-frequency attention mechanism significantly boosted the model's performance, with Precision, Recall, and F1 scores surpassing 0.950. This enhancement underscores the mechanism's effectiveness in enhancing the precision of vehicle sound classification by emphasizing relevant time frames and key frequency bands in the spectrogram. In complex auditory environments, this mechanism substantially enhances the model's ability to accurately identify and differentiate vehicle sounds. Similarly, the IDMT-Traffic model exhibited strong performance with the time-frequency attention mechanism, with a 0.014 increase in F1 score compared to its absence, highlighting the mechanism's efficacy in accurately classifying field vehicles. Figure 6 illustrates the impact of the T-F attention mechanism on classifying each class in the two datasets. Notably, the incorporation of the T-F attention mechanism significantly enhanced classification efficacy across different vehicle classes in the FVSD, enabling the model to accurately identify distinct sound patterns associated with them. The application of the time-frequency hybrid attention mechanism notably improved classification results for the Truck category in the IDMT-Traffic datasets, which initially exhibited poor performance. By thoroughly examining the time and frequency dimensions aided by the attention mechanism, subtle distinctions can be identified, thereby enhancing classification accuracy. Figure 6 illustrates the impact of the T-F attention mechanism on the classification of each class in the two datasets. Within the FVSD, a varying degree of f1 score enhancement is observed across different vehicle classes. The incorporation of the T-F Attention mechanism has significantly enhanced the efficacy of classification. This enables the model to accurately identify the distinct sound patterns associated with them. The application of the time-frequency hybrid attention mechanism significantly improved the classification results for the Truck category in the IDMT-Traffic datasets, which initially had the poorest performance. While Trucks may share some sound characteristics with other vehicle types, a thorough examination of the time and frequency dimensions, aided by the attention mechanism, allows for the identification of subtle distinctions, thereby enhancing the accuracy of classification.

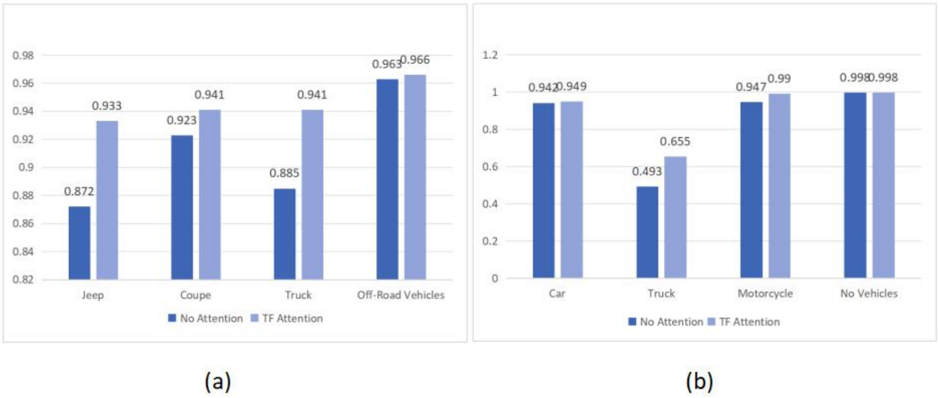

Figure 6. Impact of the attention mechanism on F1 scores across various vehicles Categories. (**a**) Result of FVSD dataset. (**b**) Result of IDMT-Traffic dataset.

3.3. **Comparison of Different Methods.** In order to further substantiate the efficacy of the techniques described in this paper, we conducted a comparative analysis of multiple established cutting-edge methodologies using two distinct datasets. A detailed description of these methods is given below: Wang et al. [26]: This is a compact deep neural network architecture that leverages MFCC as a primary feature and employs deep separable convolution to design a new classifier. Its objective is to enhance the effectiveness and precision of vehicle detection in intelligent sensor systems. The product's lightweight design allows it to be used in environments with limited computational resources, while still delivering strong performance. Abeßer et al. [29]: This approach serves as a fundamental model for the IDMT-Traffic dataset. It employs Mel spectrograms as the input and utilizes deep neural networks for classification. The primary benefit of this approach lies in its capacity to fully exploit the Mel spectrogram's capability to depict the attributes of audio signals, in conjunction with deep learning methodologies, for efficient sound classification. Ashhad et al. [27] proposed a technique that utilizes Gamma Frequency Cepstrum Coefficients (GFCC) and a set of static features to build a multi-input neural network. This model aims to achieve precise vehicle classification based on sound by combining both global and local audio features. This method is distinguished by its capacity to analyze and exploit multi-level information in the audio signal, resulting in more comprehensive classification outcomes. The experimental outcomes of the various techniques on the FVSD dataset are presented in Table 7, while the results on the IDMT-Traffic dataset are provided in Table 8.

Table 7. The results of different methods on the FVSD dataset.

Method	Feature Name	Precision	Recall	F1 score
Wang et al.	MFCC	0.818	0.812	0.814
Abeßer et al.	Log-Mel	0.848	0.835	0.840
Ashhad et al.	GFCC +Statistical	0.856	0.872	0.863
TFANet	MFCC	0.951	0.951	0.950

Table 8. Class-wise F1 score of different methods on the IDMT-Traffic dataset.

Method	Feature Name	Car	Truck	Motorcycle	No vehicle
Abeßer et al.	Log-Mel	0.94	0.50	0.96	1.00
Ashhad et al.	GFCC +Statistical	0.96	0.57	0.98	1.00
TFANet	MFCC	0.95	0.66	1.00	1.00

The FVSD dataset was evaluated using Wang et al's model, which is known for its lightweight nature. The model achieved a Precision of 0.818, a Recall of 0.812, and an F1 score of 0.814. Despite its fast training and inference capabilities, the model showed limitations in accurately classifying certain sounds. This could be attributed to the constraints of its deep separable convolution in handling intricate sound features. Ashhad et al's model enhances the classification outcomes by achieving a Precision of 0.856, a Recall of 0.872, and an F1 score of 0.863, owing to its utilization of various audio features in combination. Nevertheless, the most notable enhancement in performance is evident in our model, with all metrics surpassing 0.950. This notable enhancement can be credited to the utilization of the time-frequency attention mechanism. This mechanism allows the model to concentrate more precisely on the crucial temporal and spectral characteristics of the sound signals. Consequently, it enables the model to explore the inherent feature representation of the vehicle sound signals more effectively and filter out any unwanted noise interference.

TFANet was evaluated using the IDMT-Traffic dataset, and the F1 scores across various categories were compared with the experimental results presented in [27]. Significantly, our approach demonstrated strong and consistent performance even when classifying difficult categories, such as Trucks, which are typically less accurately classified by alternative methods. This result emphasizes the superiority of our method in precisely differentiating sound categories that are difficult to distinguish. Our model showcases both a consistently high level of accuracy and notable resilience, particularly when faced with categories that present more difficult classification tasks.

3.4. **Visualization.** Figure 7 visually represents the outcomes of these experiments. Figure 7(a) displays the results of the FVSD dataset. The graph on the left displays the fluctuation in the loss value throughout the training process. The blue curves represent the training loss per iteration, while the red curves represent the validation loss. Both the training and validation sets demonstrate a comparable decline in the loss values, suggesting successful learning. The Intermediate graph depicts the variations in accuracy, with the blue curve representing the accuracy of the training set and the red curve indicating the accuracy of the validation set. The network's learning process reaches a stable state after around 80 epochs, indicating that the network has impressive learning abilities. The figure's right side exhibits a confusion matrix, showcasing the network's classification performance across all four categories. Furthermore, the utilization of TFANet on the IDMT-Traffic dataset resulted in outstanding performance. The results not only confirm the effectiveness of the suggested approach but also showcase its suitability and resilience across various datasets.

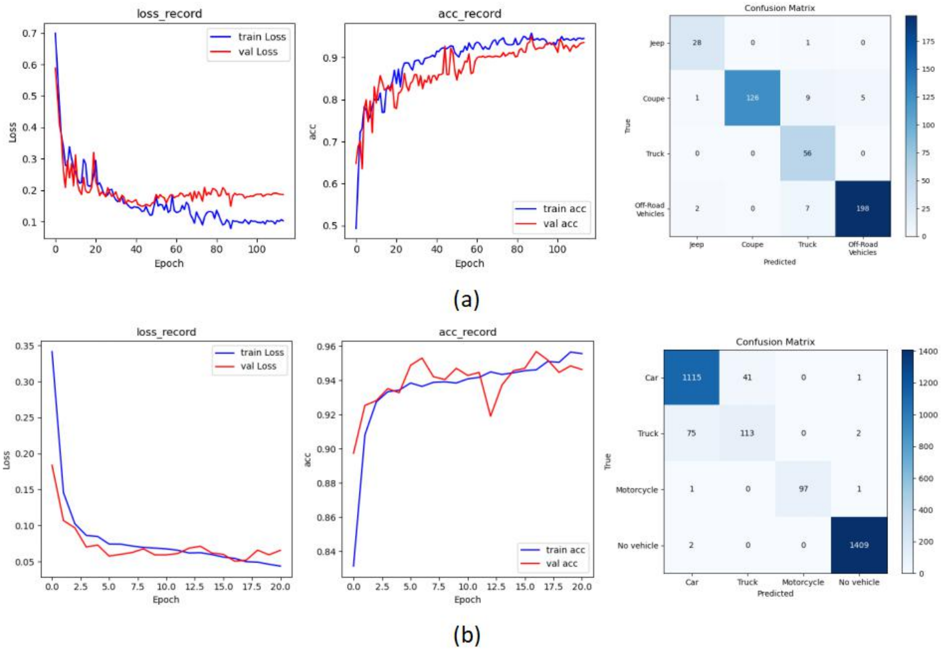

FIGURE 7. Visualization results (**a**) Loss curves, accuracy curves, and confusion matrices for the FVSD dataset(**b**)Loss curves, accuracy curves, and confusion matrices for the IDMT-Traffic dataset

4. **Conclusion.** TFANet, a novel approach for sound-based vehicle type detection, has been developed in this study. This approach utilizes MFCC as the main feature and employs ResNet18 as the classifier. Moreover, it presents a time-frequency hybrid attention mechanism, specifically developed to enhance the effectiveness of classification. This study selects the MFCC as the primary feature to accurately capture the fundamental characteristics of sound, considering the intricate nature of the natural field environment and the notable impact of environmental factors like wind noise on vehicle audio signals. During the analysis, various type classifiers were compared, and it was found that ResNet18 performed better than other networks in the task of classifying vehicle categories. The time-frequency hybrid attention mechanism, known as the core innovation of TFANet, plays a pivotal role. The system effectively recognizes and allocates suitable importance to the characteristics that are most crucial for making classification decisions. This ensures that the network focuses on pertinent information while disregarding irrelevant or distracting data during the learning phase. Comprehensive experiments were carried out using the FVSD and IDMT-Traffic to empirically confirm the effectiveness of TFANet. This strategy notably enhances the accuracy and robustness of the classification on both datasets, particularly in its capacity to differentiate between analogous categories and mitigate the impact of extraneous signals. These findings offer both theoretical insights into vehicle sound classification and robust technical support for practical applications in areas such as intelligent traffic monitoring and environmental protection.

Despite significant advancements in sound-based vehicle type detection achieved by TFANet, its accuracy diminishes notably when identifying vehicle categories with limited sample sizes. For instance, the F1 score for the Truck category is merely 0.66, significantly lower than that observed for other vehicle categories. Future research endeavors will prioritize refining the TFANet model's precision in discerning vehicle categories characterized by limited sample sizes. We will explore data augmentation techniques and unsupervised learning methods to leverage unlabeled data for increasing sample sizes. Additionally, enhancements to the time-frequency hybrid attention mechanism are envisaged to adapt to the nuances of imbalanced datasets. Through the implementation of these strategies, we anticipate an improvement in TFANet's performance, thereby providing a robust technical foundation for applications in intelligent traffic surveillance and environmental protection.

Acknowledgment. This work was funded in part by the Foundation of Science and Technology on Near-surface Detection Laboratory Fund Project, with project number 6142414210302.

REFERENCES

[1] T.-Y. Wu, Z. Lee, L. Yang, and C.-M. Chen, "A provably secure authentication and key exchange protocol in vehicular ad hoc networks," *Security and Communication Networks*, vol. 2021, 9944460, p. 17, 2021.

[2] S. Sathruhan, O.-K. Herath, T. Sivakumar, and A. Thibbotuwawa, "Emergency vehicle detection using vehicle sound classification: A deep learning approach," *2022 6th SLAAI International Conference on Artificial Intelligence*, pp. 1–6, 2022.

[3] T.-M. Nithya, P. Dhivya, S.-N. Sangeethaa, and P.-R. Kanna, "Tb-mfcc multifuse feature for emergency vehicle sound classification using multistacked cnn–attention bilstm," *Biomedical Signal Processing and Control*, vol. 88, p. 105688, 2024.

[4] X. Wang, "Vehicle image detection method using deep learning in uav video," *Computational Intelligence and Neuroscience*, vol. 2022, 8202535, 2022.

[5] B. Shobha and R. Deepu, "A review on video-based vehicle detection and tracking using image processing," in *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions.* IEEE, 2018, pp. 183–186.

[6] C.-M. Chen, Q. Miao, S. Kumar, and T.-Y. Wu, “Privacy-preserving authentication scheme for digital twin-enabled autonomous vehicle environments,” *Transactions on Emerging Telecommunications Technologies*, vol. 34(11), 2023.

[7] G. Allegro, A. Fascista, and A. Coluccia, “Acoustic dual-function communication and echo-location in inaudible band,” *Sensors*, vol. 22(3), p. 1284, 2022.

[8] K.-W. Cheng, H.-M. Chow, S.-Y. Li, and T.-W. Tsang, “Spectrogram-based classification on vehicles with modified loud exhausts via convolutional neural networks,” *Applied Acoustics Volume*, vol. 205, 2023, 109254.

[9] G. Chi, Z. Xu, E. Sowah, and W. Li, “Robust vehicle classification using sound signals and attention module,” in *International Conference on Smart Transportation and City Engineering 2021.* SPIE, 2021, 1205004, pp. 16–23.

[10] J.-Y. Wang, “Research on traffic vehicle acoustic signal recognitionmethod based on multi-feature fusion,” in *East China Jiaotong University*, 2022.

[11] N. Sharma, A.-K. Jairath, B. Singh, and A. Gupta, “Detection of various vehicles using wireless seismic sensor network,” in *2012 International Conference on Advances in Mobile Network, Communication and Its Applications.* IEEE, 2012, pp. 149–155.

[12] A. Aljaafreh and L. Dong, “An evaluation of feature extraction methods for vehicle classification based on acoustic signals,” in *2010 International Conference on Networking, Sensing, and Control (ICNSC).* IEEE, 2010, pp. 570–575.

[13] S.-S. Yang, Y.-G. Kim, and H. Choi, “Vehicle identification using discrete spectrums in wireless sensor networks,” *Journal of Networks*, vol. 3, no. 4, pp. 51–63, 2008.

[14] O. Agudelo, C. Marín, and R. Crespo, “Sound measurement, and automatic vehicle classification and counting applied to road traffic noise characterization,” *Soft Computing*, vol. 25, pp. 12 075–12 087, 2021.

[15] J. Zeng, Y. Liu, M. Wang, and X. Zhang, “Environmental sound classification based on attention feature fusion and improved residual network,” *Automatic Control and Computer Sciences*, vol. 57, pp. 371–379, 2023.

[16] H. Li, A. Chen, J. Yi, W. Chen, D. Yang, G. Zhou, and W. Peng, “Environmental sound classification based on car-transformer neural network model,” *Circuits, Systems, and Signal Processing*, vol. 42, pp. 5289–5312, 2023.

[17] Manisha, W. Clifford, E. McLaughlin, and P. Stynes, “A deep learning emotion classification framework for low resource languages.” Springer, 2023, pp. 113–121.

[18] U. Haider, M. Hanif, H. Kobayashi, L. Parajuli, D. Shimotoku, A. Rashid, and S. Safeer, “Signal classification using hybrid feature space with machine learning,” in *2023 15th International Conference on Computer and Automation Engineering.* IEEE, 2023, pp. 376–380.

[19] L. Mutanu, J. Gohil, K. Gupta, P. Wagio, and G. Kotonya, “A review of automated bioacoustics and general acoustics classification research,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22(21), p. 8361, 2022.

[20] S. Mohine, B. Bansod, R. Bhalla, and A. Basra, “Acoustic modality-based hybrid deep 1d cnn-bilstm algorithm for moving vehicle classification,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 16 206–16 216, 2022.

[21] M. Ashhad, U. Goenka, A. Jagetia, P. Akhtari, S. Ambat, and M. Samuel, “Improved vehicle subtype classification for acoustic traffic monitoring,” in *2023 National Conference on Communications.* IEEE, 2023, pp. 1–6.

[22] H. Chen and Z. Zhang, “Hybrid neural network based on novel audio feature for vehicle type identification,” *Scientific Reports*, vol. 11, no. 7648, 2021.

[23] L. Sun, Z. Zhang, H. Tang, H. Liu, and B. Li, “Vehicle acoustic and seismic synchronization signal classification using long-term features.” *IEEE Sensors Journal*, vol. 23, no. 10, pp. 10 871–10 878, 2023.

[24] S. Mohine, P. Gupta, B. Bansod, R. Bhalla, and A. Basra, “Evaluation of acoustic modality features for moving vehicle identification.” *Multidimensional Systems and Signal Processing*, vol. 33, pp. 1349–1365, 2022.

[25] Y. Luo, L. Chen, Q. Wu, and X. Zhang, “Sound-convolutional recurrent neural networks for vehicle classification based on vehicle acoustic signal.” *2021 International Conference on Smart City and Green Energy*, pp. 98–102, 2021.

[26] C. Wang, Y. Song, H. Liu, H. Liu, J. Liu, B. Li, and X. Yuan, “Real-time vehicle sound detection system based on depthwise separable convolution neural network and spectrogram augmentation.” *Remote Sensing*, vol. 14(19), p. 4848, 2022.

[27] M. Ashhad, O. Ahmed, S. Ambat, Z. Haq, and M. Alam, “Mvd: A novel methodology and dataset for acoustic vehicle type classification.” *arXiv preprint*, vol. 2309.03544, 2023.

[28] A. Tripathi and A. Mishra, “Environment sound classification using an attention-based residual neural network,” *Neurocomputing*, vol. 460, pp. 409–423, 2021.

[29] J. Abeßer, S. Gourishetti, A. Kátai, T. Clauß, P. Sharma, and J. Liebetrau, “Idmt-traffic: An open benchmark dataset for acoustic traffic monitoring research,” *2021 29th European Signal Processing Conference*, pp. 551–555, 2021.

[30] G. Sharma, K. Umapathy, and S. Krishnan, “Trends in audio signal feature extraction methods.” *Applied Acoustics*, vol. 158, 2020, 107020.

[31] B. Selbes and M. Sert, “Multimodal vehicle type classification using convolutional neural network and statistical representations of mfcc,” *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, 2017.

[32] M. Sidhu, N. Latib, and K. Sidhu, “Mfcc in audio signal processing for voice disorder: a review,” *Multimedia Tools and Applications*, 2024.

[33] A.Vaswani, N. Shazeer, and N. Parmar, “Attention is all you need,” *Advances in Neural Information Processing Systems*, p. 30, 2017.

[34] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*, pp. 770–778, 2016.

[35] S.-J. Bu, H.-J. Moon, and S.-B. Cho, “Adversarial signal augmentation for cnn-lstm to classify impact noise in automobiles,” *2021 IEEE International Conference on Big Data and Smart Computing*, pp. 60–64, 2021.

[36] K. Shariff, S. Zainuddin, and M. Ali, “Detection of wet road surfaces from acoustic signals using scalogram and optimized alexnet,” *12th Symposium on Computer Applications & Industrial Electronics*, pp. 159–163, 2022.