Open Peer Review on Qeios

Acoustic Vehicle Classification using Deep Learning Trained on a Spectrogram and Scalogram Fusion

Khairul Khaizi Mohd Shariff[1], Rajeswari Raju[1], Ihsan Yassin[1], Farzad Eskandari[2], Megat Syahirul Amin Megat Ali[1]

1 Universiti Teknologi Mara
2 Allameh Tabataba'i University

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

This paper explores an audio-based on-road vehicle classification method that utilizes visual representations of sound through spectrograms, scalograms, and their fusion as features, classified using a modified VGG16 Convolutional Neural Network (CNN) architecture. The proposed method offers a non-intrusive, potentially less costly, and environmentally adaptable alternative to traditional sensor-based and computer vision techniques. Our results indicate that the fusion of scalogram and spectrogram features provides enhanced accuracy and reliability in distinguishing between vehicle types. Performance metrics such as training and loss, alongside precision and recall of classes, support the efficacy of a richer feature set in improving classification outcomes. The fusion features demonstrate a marked improvement in distinguishing closely related vehicle classes like 'Cars' and 'Trucks'. These findings underline the potential of our approach in refining and expanding vehicle classification systems for intelligent traffic monitoring and management.

Khairul Khaizi Mohd Shariff[1], **Rajeswari Raju**[2], **Ihsan Mohd Yassin**[1,*], **Farzad Eskandari**[3], and **Megat Syahirul Amin Megat Ali**[1]

[1] *Microwave Research Institute, Universiti Teknologi Mara (UiTM), Shah Alam*

[2] *College of Computing, Informatics and Mathematics, Universiti Teknologi MARA (UiTM) Terengganu, Malaysia*

[3] *Allameh Tabataba'ie University, Islamic Republic of Iran*

[*]Correspondence: ihsan_yassin@uitm.edu.my

Introduction

Vehicle classification is a fundamental aspect of road use management, with wide-ranging implications for infrastructure design, traffic management safety, and environmental policy [1][2][3]. The task is defined as systematic categorization of vehicles based on their physical characteristics. Typically, vehicles are broadly classified into categories such as motorcycles, passenger vehicles, and heavy goods vehicles, with further subdivisions within each category based on specific attributes [4]. The classification criteria and categories can vary significantly across different jurisdictions but generally serve the same underlying purposes of managing and regulating road use.

In infrastructure planning, roads and bridges are tailored for anticipated traffic, incorporating robust designs for heavy vehicles, and using vehicle classification data for cost-effective, durable, and safe roads, ensuring future traffic needs are met. Vehicle classification data streamlines traffic management by informing the design of control devices, speed limits, and lane configurations, facilitating efficient and safer roads [5][6][7]. Urban areas leverage this data for traffic regulations, while ITS use it for dynamic traffic strategies like variable speed limits or rerouting, reducing congestion. From a safety perspective, vehicle classification informs regulations addressing varying risks of different vehicle types, leading to targeted enforcement and safer roads. Heavier vehicles, for example, require stricter regulations due to their increased risk factors. Vehicle classification also brings benefits from an environmental standpoint, vehicle classification aids in shaping policies and initiatives tailored to reduce emissions and noise pollution. Policymakers use this data to create targeted measures like low emission zones and forecast transport emission trends, supporting broader environmental strategies.

Effective traffic management and the development of intelligent transportation systems rely heavily on accurate and robust traffic monitoring technologies [8]. Current vehicle classification methods come with inherent drawbacks that necessitate a balance between accuracy, cost, and adaptability. On-road vehicle classification predominantly falls into two categories: sensor-based systems and computer vision techniques. Sensor-based systems, such as inductive loop detectors [9], magnetic sensors [10], and vibration sensors[11], are physically installed on or near roadways. These systems classify vehicles based on physical characteristics like axle count, weight, speed, and length. Although sensor-based methods are widely used for their reliability and accuracy in data collection, they come with significant drawbacks. The installation and maintenance of these systems are often costly and disruptive to traffic. Additionally, they can be susceptible to wear and tear, leading to frequent needs for recalibration and repair.

Computer vision techniques represent a more modern approach, utilizing cameras and image processing algorithms to classify vehicles (see examples in [12][13][14]). These systems analyze video or images to identify vehicle characteristics, often employing machine learning techniques to improve accuracy [13]. The advantage of computer vision methods is their flexibility and the richness of data they can provide, not just in terms of classification but also in monitoring traffic patterns and driver behavior. However, these methods face challenges in varying light conditions, occlusions, and complex urban environments where multiple vehicles and pedestrians are present. Furthermore, the accuracy of such systems heavily relies on the quality of the algorithms used and the training data, which must be extensive, and representative of the diverse vehicle types and conditions encountered on roads.

In considering both sensor-based and computer vision techniques, a common challenge emerges: the balance between accuracy, cost, and adaptability. While sensor-based methods are generally more reliable and accurate for specific measurements like weight or axle count, their high installation and maintenance costs can be prohibitive. On the other hand, computer vision offers a less intrusive and potentially more cost-effective solution but at the expense of accuracy and robustness under variable conditions. Both methods also need to continuously adapt to changes in vehicle technology and types, requiring ongoing updates and maintenance. These limitations underscore the necessity for innovative approaches that balance these factors. The exploration of vehicle sound analysis presents an intriguing frontier. Sound, an inherent and distinctive characteristic of moving vehicles, offers a potentially rich dataset for classification purposes (see [15], [16] for several related works). By harnessing the acoustic signatures of vehicles, there is an opportunity to circumvent some of the physical and environmental constraints that hamper existing methods.

Our proposed solution involves the strategic placement of microphones along roadways to capture the unique acoustic footprints of passing vehicles. These acoustic signals will then be transformed into visual representations, specifically scalograms and spectrograms, which capture the frequency and time information of sound signals in a two-dimensional format. Scalograms, generated through wavelet transforms, provide detailed representations of signal frequency variations over time, while spectrograms, derived from Fourier transforms, illustrate the signal's frequency content over time. These visual representations encapsulate the complex audio patterns associated with different types of vehicles, turning raw sound into structured, analyzable images.

To leverage these structured sound representations for vehicle classification, a modified VGG16 Convolutional Neural Network (CNN) was employed. CNNs have demonstrated remarkable success in image recognition tasks due to their ability to learn hierarchical patterns and features from visual data [17]. By training the modified VGG16 model on a diverse dataset of scalogram and spectrogram images, categorized by vehicle type, it has learned to distinguish the nuanced differences in the acoustic signatures of motorcycles, passenger cars, trucks, and other vehicle categories. This approach offers a non-intrusive, potentially cost-effective solution that is adaptable to various environmental conditions without the need for extensive physical infrastructure. Moreover, the continuous improvement in audio processing and machine learning techniques promises enhancements in accuracy and applicability, making sound-based vehicle classification a compelling avenue for innovation in traffic monitoring and management systems.

Despite the promising potential of using visual representations of sound for vehicle classification, relying on a single type of transformation, such as scalograms or spectrograms alone, may not fully capture the intricate acoustic patterns of different vehicles. Each representation has its strengths—scalograms provide excellent time-frequency resolution for non-stationary signals, while spectrograms offer a clear depiction of energy distribution over frequency bands. However, individually they might miss subtle yet crucial features necessary for a robust classification. The proposed solution is to employ a fusion of both scalogram and spectrogram images, thereby combining the comprehensive time-frequency information from scalograms with the distinct energy distribution patterns of spectrograms. This fusion aims to create a richer, more descriptive feature set for the Convolutional Neural Network (CNN) to learn from, enhancing the model's ability to differentiate between vehicle types with higher accuracy. By leveraging the complementary strengths of both representations, the system is hypothesized to achieve superior performance in vehicle classification tasks, particularly in complex acoustic environments. This innovative approach promises to address the limitations of single-method representations and sets a new benchmark for accuracy and reliability in sound-based vehicle classification.

The remainder of this paper is organized as follows: Section 2 presents the methodology of the research, followed by the results and discussion in Section 3. Finally, concluding remarks are presented in Section 4.

Related Works

Recent research trends in the field of vehicle detection and classification have primarily focused on leveraging advanced machine learning techniques, particularly deep learning models, for improved accuracy and efficiency. In [18], an ensemble of deep learning models, including fully connected neural networks (FCNet), convolutional neural networks (CNN), and recurrent neural networks (RNN), was developed to detect emergency vehicles based on siren sounds. This system aimed to enhance traffic management by reducing waiting times at intersections. Similarly, [19] explored the use of CNNs for acoustic detection of emergency vehicles, highlighting its potential benefits for hearing-impaired drivers and integration into driver assistance systems. Chiang et al. [20] investigated vehicle detection and classification using Distributed Acoustic Sensor (DAS) systems, employing a CNN network for analyzing vehicle models and sizes. In a different approach, [21] focused on classifying vehicle sub-types for acoustic traffic monitoring using a Convolutional Neural Network (CNN), demonstrating significant improvements in accuracy. [22] and [8] both introduced innovative methods using DAS technology for traffic monitoring, with [Zhipeng] developing a system for monitoring traffic flow and vehicle speed, and [8] proposing a long-range traffic monitoring system using fiber-optic Distributed Acoustic Sensing (DAS). Lastly, [23] and [24] both proposed novel methods for vehicle traffic monitoring using acoustic signals, with[23] utilizing Mel-Frequency Cepstral Coefficients (MFCC) and Long Short-Term Memory (LSTM) networks, and [24] introducing a Long-Term Correlation Feature Network (LTCFN) for field vehicle acoustic and seismic signal classification. These studies collectively represent the cutting-edge advancements in vehicle detection and classification, highlighting the growing reliance on deep learning and sensor fusion technologies for traffic monitoring and management.

In [18], an efficient system for detecting emergency vehicles based on siren sounds leveraging an ensemble of deep learning models was developed. The system aims to detect emergency vehicles approaching intersections to reduce waiting times and improve traffic management. The approach involves creating and evaluating an ensemble of fully connected neural networks (FCNet), convolutional neural networks (CNN), and recurrent neural networks (RNN). The ensemble was trained on Mel Frequency Cepstral Coefficients (MFCC) features extracted from audio data, with the dataset sourced from Google's AudioSet Ontology. The ensemble method combines predictions from the three base models using majority voting, namely FCNet (based purely on dense layers without convolutional layers), CNN_Net (includes up to 6 2D convolutional layers), and RNN_Net (comprises various long short-term memory (LSTM) layers). The dataset consists of siren sounds from four different types of emergency vehicles. The models were evaluated based on their training and testing accuracy, with the ensemble model achieving the highest accuracy.

Similarly, [19] explored the development of an acoustic method for detecting emergency vehicles using convolutional neural networks (CNNs), intending to address the challenge of drivers not hearing emergency vehicle sirens due to various factors. The proposed system can assist in traffic safety, especially for hearing-impaired drivers, and can be integrated into driver assistance or autonomous vehicle systems. The authors employed a CNN model trained on a dataset comprising siren sounds and city traffic noises. The dataset used was "Emergency Vehicle Siren Sounds," sourced from public internet platforms like Google and YouTube, and stored in.wav audio format. The audio data were transformed using spectrograms, a visual representation of the spectrum of frequencies in a sound, for analysis through CNNs. The proposed work was developed in Python, leveraging TensorFlow and Keras for machine learning, and deployment on the Google Colab platform. The model achieved an average accuracy of 93.3% and a recognition speed of approximately 0.0004±5% seconds.

Chiang et al. [20] investigated the use of Distributed Acoustic Sensor (DAS) systems for vehicle detection and classification, focusing on recognizing vehicle models and sizes. They employed a 13-layer CNN network for feature extraction from DAS signals, followed by a softmax function for classification. The study utilized controlled experiments with DAS data collected along a 4.8 km road. Five different vehicle models were studied at varying speeds (30, 40, 50, 60, 70 km/h). The model's performance was evaluated using accuracy and a confusion matrix, achieving mean accuracy of 94% for vehicle type and 95% for size classification.

Reference [21] focused on improving the classification of vehicle sub-types for acoustic traffic monitoring (ATM) using a well-optimized Convolutional Neural Network (CNN). The authors used simple CNN architectures with various feature extraction methods including Mel Frequency Cepstral Coefficients (MFCC), Gammatone Frequency Cepstral Coefficients (GFCC), and Mel spectrograms for classifying acoustic signals into four classes: car, truck, bike, and no vehicle. Temporal stretching was used as the primary data augmentation strategy to address class imbalance issues in the dataset. The research utilized the IDMT Traffic dataset, which includes stereo audio recordings of different vehicle types under various conditions. The proposed methodology achieved an accuracy of 98.95%, significantly improving upon the state-of-the-art baseline for the IDMT Traffic dataset. The study compared the performance of its CNN model with more complex networks like VGGnet, demonstrating that simpler models can be equally or more effective.

The integration of DAS technology with machine learning in[22] for traffic monitoring represents an innovative approach in the field. The authors developed a system for monitoring traffic flow and vehicle speed using Distributed Acoustic Sensing (DAS) and object detection methods. The approach involves collecting DAS data using roadside DAS arrays combined with deep learning-based object detection techniques. The research has significant implications for intelligent traffic management systems, offering a novel approach to monitoring traffic flow and vehicle speed, with potential contribution to the development of smart city infrastructures and efficient traffic management solutions.

Similar work by [8] proposed a novel long-range traffic monitoring system using fiber-optic Distributed Acoustic Sensing (DAS) with optimized pulse compression, a first in traffic-monitoring DAS systems. The work utilized optical pulse compression in a coherent optical time-domain reflectometry (COTDR) setup for higher sensitivity and range. A new vehicle detection and tracking algorithm based on a novel transformed domain, an evolution of the Hough Transform, was introduced. Real-world testing was conducted using a dark fiber in a telecommunication cable running along a 40 km road, demonstrating the system's practical applicability. The fiber-optic cable used was installed along a road, detecting vibrations induced by passing vehicles. The sensor configuration appears to produce high-sensitivity and high-resolution vibration measurement signals, essential for the vehicle detection and tracking algorithm. The system classifies vehicles based on size (cars and trucks) and travel direction (westbound and eastbound). The classification used Support Vector

Machines (SVM) and takes advantage of the linear nature of the OPC-COTDR measurements, leveraging features such as peak-to-peak amplitude, center amplitude, lateral area, and velocity for vehicle classification. These features are instrumental in differentiating between vehicle types and detecting non-vehicle classes. The classification system achieved a respectable accuracy of 97.7% for vehicle passing events. It was particularly effective in differentiating cars from trucks and in reducing misclassification of non-vehicle events.

In [23], a large-scale monitoring of vehicle traffic using acoustic signals was proposed using Mel-Frequency Cepstral Coefficients (MFCC) and Long Short-Term Memory (LSTM) networks, addressing the high cost and privacy concerns associated with traditional vision-based traffic sensors. The study explores a novel approach in acoustic vehicle classification, combining feature extraction (MFCC) with LSTM networks, a deep learning technique known for its efficiency in modeling sequential data. The MFCC features were obtained by transforming frequency scales into the Mel Scale, which mimics human perception of sound. This process involves several steps, including signal segmentation, power spectrum generation, and applying a discrete cosine transform. The study used a dataset from the Fraunhofer Institute for Digital Media Technology, comprising traffic noise data recorded from different road types under various conditions. The dataset included samples for four vehicle categories: motorcycle, car, truck, and no traffic. Audio data was pre-processed by converting stereo to mono, downsampling, and limiting the duration to two seconds. The dataset was divided into training, validation, and testing groups, and MFCC features were extracted using MATLAB functions. The LSTM model achieved an accuracy of 82-86.2% across the four vehicle categories, indicating the system's reliability in classifying vehicles based on their acoustic signatures.

Sun et al. [24] proposed a field-vehicle-type recognition system using acoustic and seismic sensors, a key aspect in border protection tasks. The research aims to improve signal classification by addressing limitations in existing methods that primarily focus on frequency-domain characteristics of signals and neglect their time-domain aspects. The paper introduces a Long-Term Correlation Feature Network (LTCFN) for classifying field vehicle acoustic and seismic signals. This network integrates an AlexNet-type feature extractor with an LSTM-based classifier, leveraging both intraframe network and fusion methods for feature vector extraction, and an interframe classifier for analyzing time correlation and overall classification. The proposed method replaces traditional handcrafted feature extraction with a neural network approach, utilizing the AlexNet architecture for processing each signal frame. This shift to deep learning enables the extraction of features more representative of the signal's characteristics. The LTCFN network generated feature maps by fusing acoustic and seismic features within the same frame into feature vectors, which are then combined into a comprehensive feature map. This approach overcomes limitations of single-frame signal classification that may be prone to environmental noise or distance-related inaccuracies. Consequently, the LSTM network was employed to learn the potential characteristics of the feature map, going beyond simple addition or ratio operations used in traditional methods, enabling more sophisticated analysis and interpretation of a diverse range of vehicle categories. The LTCFN demonstrated an impressive classification accuracy of up to 96%, outperforming other fusion methods and showing superior anti-noise performance as it was able to effectively learn time-correlated characteristics of feature maps, enhancing its performance compared to traditional majority-voting methods.

Methodology

Hardware & Software Specification

The hardware and software specifications for the research is shown in Table 1. The NVIDIA GeForce RTX 3090 is equipped with 10,496 CUDA cores, critical in accelerating DLNN training, offering parallel processing capabilities essential for handling the extensive computations involved. These cores expedite matrix and vector operations, allowing simultaneous processing of data points and significantly reducing computation time. The high parallelism optimizes resource allocation and improves hardware performance, leading to faster model training and iteration.

Table 1. Hardware & Software Specifications

Item	Specifications
Central Processing Unit (CPU)	Advanced Micro Devices (AMD) Threadripper 3990X
Graphics Processing Unit (GPU)	NVidia RTX 3090, 24 GB VRAM
Random Access Memory (RAM)	64 GB DDR4
MATLAB	r2023a
Operating System	Microsoft Windows 11 Professional

Data Collection

The dataset used in this study was like[23]. It was obtained from the Fraunhofer Institute for Digital Media Technology (IDMT), Germany. This dataset's comprehensive coverage of different road types and conditions, along with high-quality audio recording, provides a solid foundation for the study's vehicle classification analysis.

Data was gathered from roads located in Ilmenau, Germany under four different conditions, including one country road and three urban roads. Additionally, data collection occurred under both dry and wet states, offering a range of acoustic environments that affect traffic noise. This diversity in recording environments adds to the dataset's robustness.

The system used for collecting audio consisted of two sets of microphones, namely a set of condenser microphones from the SE electronics brand, model sE8, and a set of MEMS (Micro-Electro-Mechanical Systems) microphones. The microphones were positioned 50 centimeters away from the road, ensuring the capture of clear traffic sound. All audio events were captured simultaneously at a rate of 48 kHz, with each sample of data lasting for two seconds. This uniform length for audio samples is crucial for consistent analysis. For this study, two-second audio samples were used as this was the average length in the dataset.

In the study, the collected traffic noise data was converted into spectrogram and scalogram features using MATLAB, a process critical for effective pattern recognition by the custom VGG16 deep learning model. Spectrograms, generated through Short-Time Fourier Transform (STFT), visually represent the spectrum of frequencies of a signal as it varies over time, offering a detailed view of the signal's frequency content and temporal evolution. This representation is particularly beneficial for VGG16, as its convolutional layers are adept at recognizing and learning from the intricate patterns in the frequency-time domain, crucial for distinguishing different vehicle sounds. Scalograms, on the other hand, are created using Continuous Wavelet Transform (CWT), providing a multi-resolution analysis of the signal. This approach is advantageous for capturing both high-frequency events occurring over short durations and low-frequency components over extended periods. The Scalogram's ability to represent non-stationary signals at various scales complements VGG16's pattern recognition capabilities, enabling it to discern subtle and complex features inherent in traffic noise, thus enhancing the overall accuracy of vehicle classification.

MATLAB's image fusion method was employed to combine the spectrogram and scalogram images into a single image, producing a richer and more informative feature set for classification. This method integrates complementary information from multiple images to enhance the quality and increase the utility of the synthesized image. We hypothesize that given similar network architectures, the fusion trained network would produce better classification accuracies due to its enriched feature set relative to individual spectrogram or scalogram features.

Pre-Processing

In preparation for the vehicle classification task, the first step involves standardizing the input images to a uniform size. This was achieved by setting a target size (224 × 224 pixels) and resizing all images in the dataset to these dimensions. This standardization is crucial as it ensures that all input images are compatible with the neural network architecture, specifically the VGG16 DLNN model used.

The dataset was then split into training and validation subsets: 70% for training and 30% for validation. The images were randomly distributed. This is important for the learning task, where the model was trained on the training data and evaluated on an unseen separate set (validation data) to gauge its performance and generalization capabilities. This approach helps in understanding the model's predictive accuracy and avoiding overfitting to the training data.

CNN Architecture

The VGG16 model is a CNN architecture known for its depth and simplicity, consisting of 16 layers that include 13 convolutional layers, 3 fully connected layers, and pooling layers interspersed between some of the convolutional layers [25]. Developed by the Visual Graphics Group from the University of Oxford (hence the name VGG), it was a breakthrough due to its uniform use of 3 x 3 convolutional filters throughout the network and its depth, significantly deeper than its predecessors at the time. The consistent use of small filters in successive layers allows it to capture a wide range of features in the input images, leading to highly effective recognition capabilities. Despite its relatively simple architecture, VGG16's depth and effectiveness in feature extraction make it a powerful tool for image classification tasks.

Initially designed for the 1,000-case ImageNet classification task, the VGG16 network was customized for our purpose.

The last three layers of the standard VGG16 model were replaced with a four-class output based on the number of vehicle categories (motorcycle, car, truck and no vehicle). A fully connected layer with the same number of neurons as the number of classes was added, followed by a softmax layer to transform the output into probability distributions, and finally, a classification layer that produces the actual class prediction. Training was done from scratch without the use of transfer learning.

Training Parameters and Models

The choice of parameters to train the CNN is crucial (Table 2). The specified training options employ the adaptive moment estimation (ADAM) optimizer [26], renowned for its efficiency in computation and low memory requirement, making it particularly suitable for tasks with large datasets and parameter spaces like image classification. ADAM is an adaptive learning rate method that stands out for its ability to update network weights iteratively based on training data, thus contributing to faster convergence and improved accuracy. The minibatch size was set to 128 to allow the model to update weights more frequently, leading to more robust learning, while the maximum epochs were set to 100 to ensure sufficient iterations over the dataset for the network to learn and generalize well from the acoustic features represented in the images.

Table 2. Training Parameters & Specifications

Item	Specification
Training Algorithm	ADAM
Minibatch Size	128
Maximum Epochs	100
Initial Learning Rate	0.0001
Validation Check	Performed every 10 iterations.
Dataset Randomization (Shuffling)	Performed every epoch.

The inclusion of validation data promotes model reliability and generalization. By evaluating the model against a validation set every ten iterations, it's possible to monitor and prevent overfitting, ensuring the model's performance truly reflects its ability to generalize to new, unseen data. This iterative validation approach allows for continuous assessment of the model's performance, providing opportunities to tweak or halt training as needed based on validation results. Additionally, the training dataset was randomized and shuffled after every epoch to prevent the model from learning spurious patterns and encouraging more robust feature learning.

The initial learning rate of 0.0001 was a conservative choice to mitigate the risks of overshooting minima in the loss landscape—a common challenge in training deep neural networks. Starting with a smaller learning rate allows the model to make gradual, more precise updates to weights, leading to steady convergence. The training plots were examined after 10 epochs to provide valuable insights into the model's learning trajectory and convergence behavior, as well as to monitor issues like overfitting or underfitting, guiding further refinements in the training process.

Validation Methods

In combination, accuracy and loss plots alongside the confusion matrix provide a comprehensive evaluation framework for training DLNNs. While accuracy and loss plots offer a macro view of model performance and learning stability, the confusion matrix delivers detailed insights into class-specific performance. This approach enables a thorough understanding of the model's strengths and weaknesses, guiding optimization and refinement for better overall performance and reliability. Accuracy plots depict the proportion of correctly classified instances, reflecting the model's effectiveness, while loss plots show the model's error rate, indicating the disparity between predicted outputs and actual labels. These plots provide insights into the model's learning trend, revealing patterns of improvement, overfitting, or underfitting, and are instrumental in assessing the overall model health and progression over training epochs.

Post-training, the confusion matrix was used to visualize the performance of the model, presenting the counts of actual versus predicted labels in a matrix format. This matrix presents a snapshot of not only the overall accuracy but also the specific types of classification errors and is particularly useful in identifying consistent misclassifications, aiding in targeted improvements to the model or data.

Results & Discussions

Data Samples

Figure 1 to Figure 3 are samples of visual representation of acoustic signatures for the different classes, processed through various time-frequency analysis techniques. There are noticeable differences between the visual representation of the features.

For the spectrogram representation, the energy distribution of the Car class is relatively even with some concentration in the mid-frequency range, suggesting a steady-state engine noise. In scalogram representation, there are distinct horizontal bands indicating consistent frequencies over time, typical for the harmonic structure of engine sounds. These characteristics are combined in the fusion image showing a blend of steady-state energy and harmonic bands.

The motorcycle sound is represented as a more varied energy distribution, possibly due to a higher frequency operation or less noise dampening compared to a car, while the scalogram shows more intense and varied frequency components over time, which could be due to the acceleration and deceleration cycles of motorcycles being more pronounced. The fusion enhances the variability in the motorcycle's acoustic signature, which could be useful in detecting transient behaviors like acceleration.

The spectrogram of truck audio appears similar to the car but with energy spread over a broader frequency range, possibly due to the larger engine size and associated harmonics. The scalogram features exhibit intense low-frequency components and clearer banding than the car or motorcycle, possibly due to the deeper sounds emitted by large truck engines. Meanwhile, the fusion image shows a robust representation of the truck's acoustic features, with strong low-frequency components and harmonic bands that could be key identifiers for this type of vehicle.

Finally, there appears to be almost uniform spectrogram distribution with less intensity in the No Traffic class indicating ambient noise or silence. Similarly, the scalogram indicates minimal features suggestive of a lack of dominant frequency components, which aligns with the absence of vehicular noise. The fusion does not add significant features, indicating the robustness of this method in cases where there is no significant acoustic event to detect.

Based on the appearance of the images, the spectrogram provides a time-frequency representation where the intensity of color reflects the energy or power at a given frequency and time. The scalogram, derived from the CWT, gives a more nuanced view of how frequency components vary over time. The fusion of these two methods aims to leverage the advantages of both representations for more accurate vehicle detection and classification in a noisy environment.

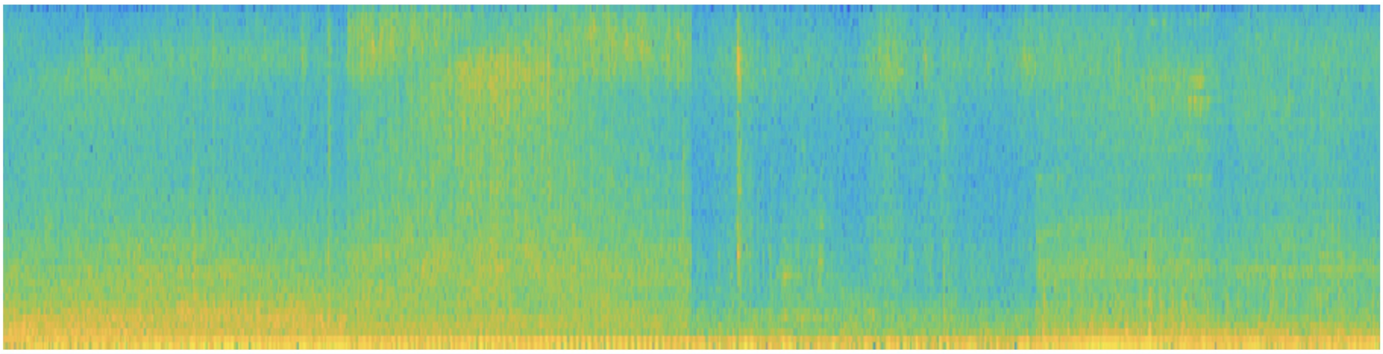


Figure 1. Spectrogram Features for (a) Car, (b) Motorcycle, (c) No Traffic, (d) Truck

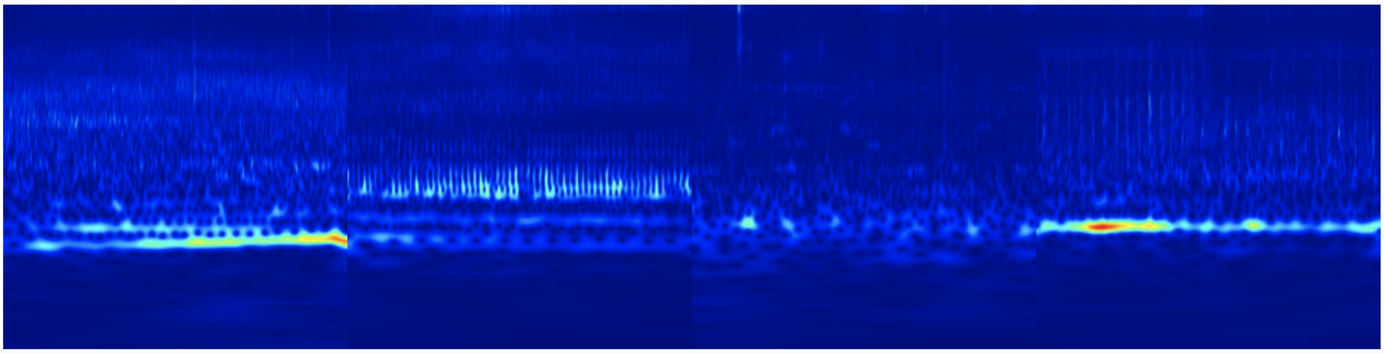


Figure 2. Scalogram Features for (a) Car, (b) Motorcycle, (c) No Traffic, (d) Truck

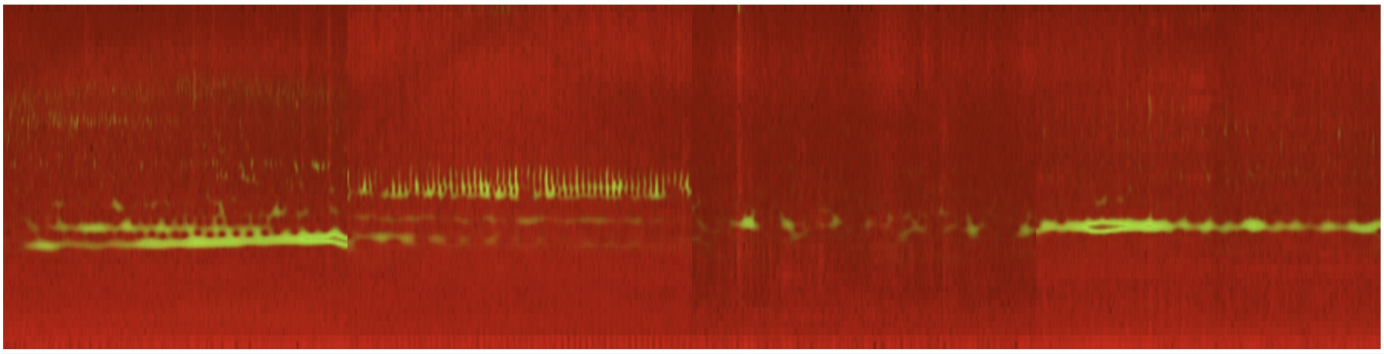


Figure 3. Fusion Features for (a) Car, (b) Motorcycle, (c) No Traffic, (d) Truck

Comparison between Spectrogram- Scalogram-, and Fusion-Trained Models

The training and loss plots for the modified VGG16 architecture trained on spectrogram, scalogram, and fusion features are shown in Figure 4 and Figure 5, while Table 3 shows the models' final performances at epoch 100. All models demonstrated a robust and consistent increase in training and validation accuracy, while the loss metrics for the training set of all models consistently decreased, reflecting their ability to effectively minimize classification errors over time regardless of the features used. This observation demonstrates the suitability of the VGG16 architecture for spectrogram, scalogram and fusion of features. Additionally, the scalogram model edged out slightly in terms of validation accuracy and stability, making it potentially more reliable for generalizing to new data.

Table 3. Accuracy and Loss Values at Final Epoch

Model	Final (at epoch 100)			
	Training Accuracy (%)	Validation Accuracy (%)	Training Loss	Validation Loss
VGG16-Spectrogram	100.00	85.28	1.32×10^{-6}	1.4584
VGG16-Scalogram	100.00	86.36	4.14×10^{-7}	1.7637
VGG16-Fusion	100.00	87.99	1.87×10^{-6}	1.2849

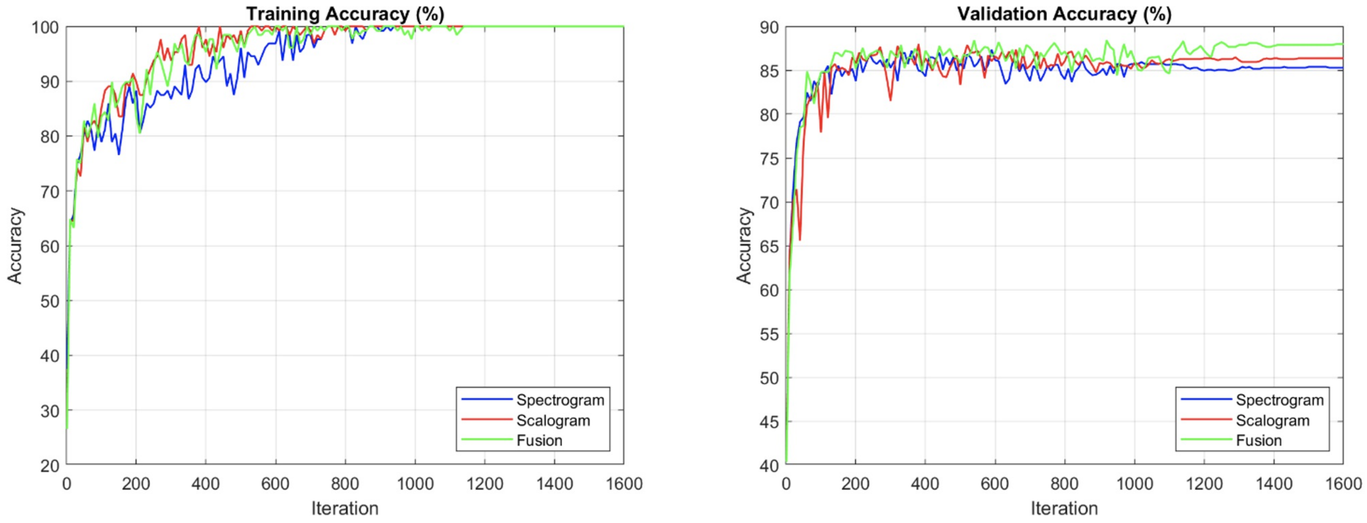


Figure 4. Training Plot for Spectrogram, Scalogram and Fusion Features

Figure 4. Training Plot for Spectrogram, Scalogram and Fusion Features

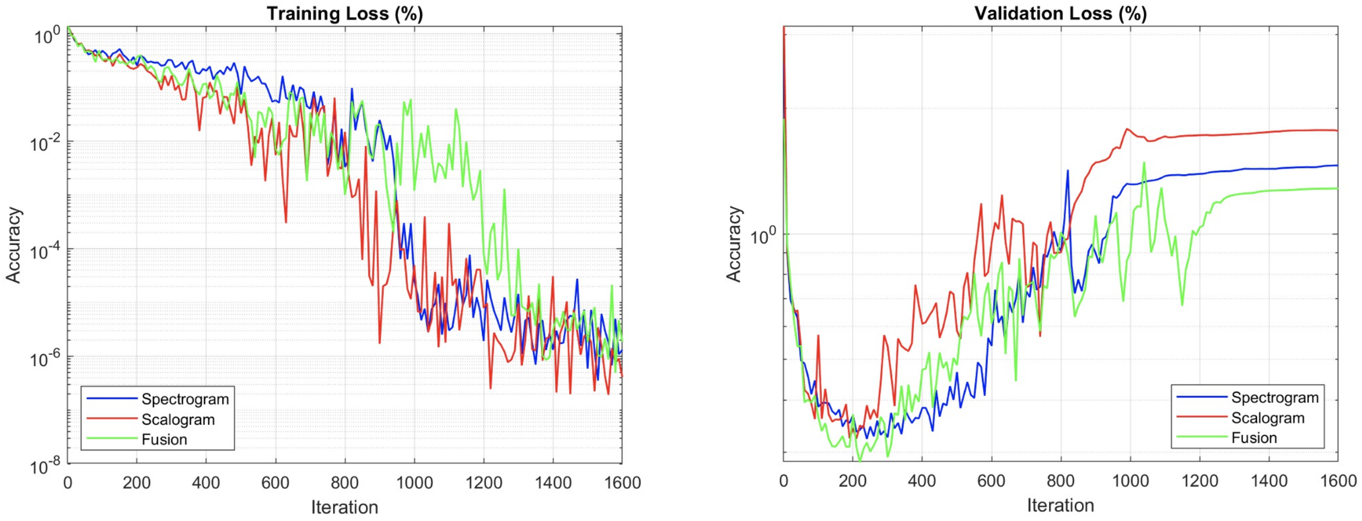
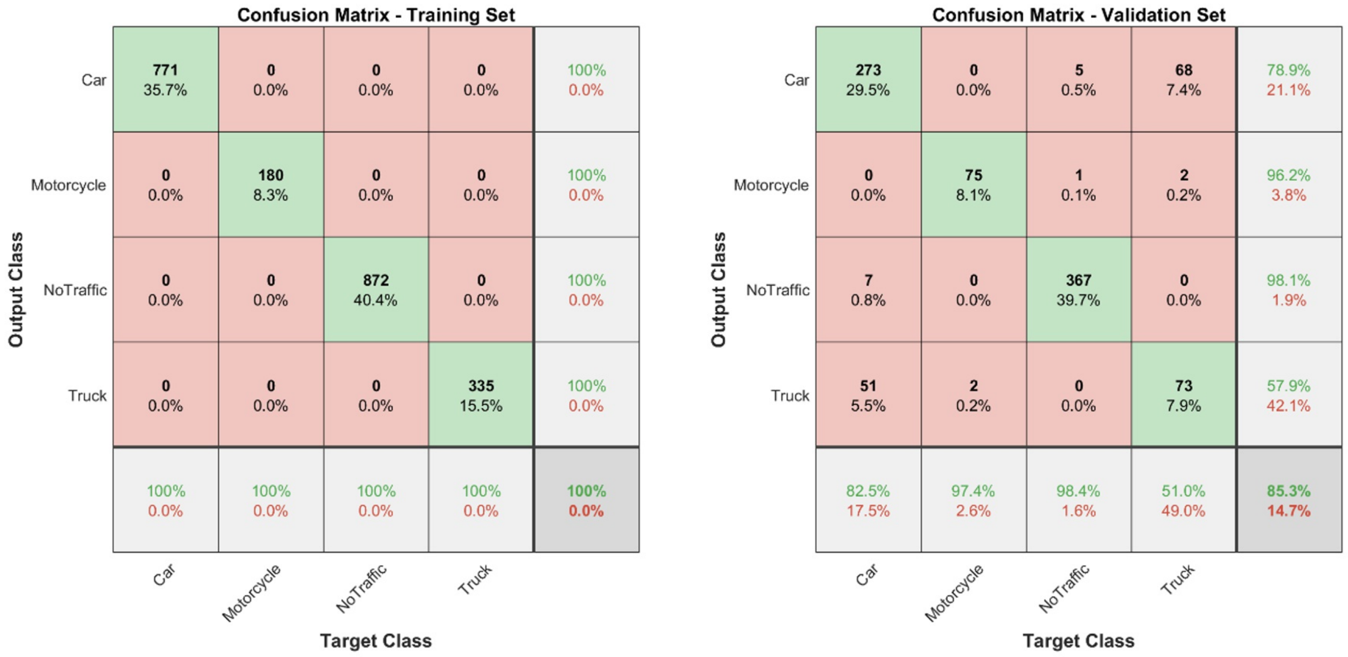


Figure 5. Loss Plot for Spectrogram, Scalogram and Fusion Features

The confusion matrices in Figure 6, Figure 7, and Figure 8 offer a comparative analysis of the classification performance of the custom VGG16 models trained on the three features. For the spectrogram features, the matrices demonstrated high precision for the 'No Traffic' and 'Motorcycle' classes, as indicated by the high percentage values on the diagonal of the matrix. However, there was notable confusion between the 'Car' and 'Truck' classes, with a significant number of 'Truck' instances misclassified as 'Car'. The scalogram- and fusion-trained models all suffer from these issues, albeit less pronounced compared to the spectrogram model. Notably, the 'Truck' class has improved precision and recall in the scalogram and fusion feature models, suggesting that the addition of scalogram features enabled improved class separation for this category. Nevertheless, all models appear to struggle with distinguishing features between these two classes as the feature signatures for the ‘Car’ and ‘Truck’ closely resemble each other, possibly due to their almost similar engine sounds.

Overall, the fusion-trained model appears to outperform the model trained on spectrogram and scalogram features, especially in distinguishing between the 'Car' and 'Truck' classes. This could be attributed to a richer feature set due to the combination of advantages from spectrogram and scalogram features. The results suggest that fusion features are better suited for tasks requiring fine-grained differentiation between classes with overlapping feature spaces.



Confusion Matrix - Training Set

Output Class / Target Class

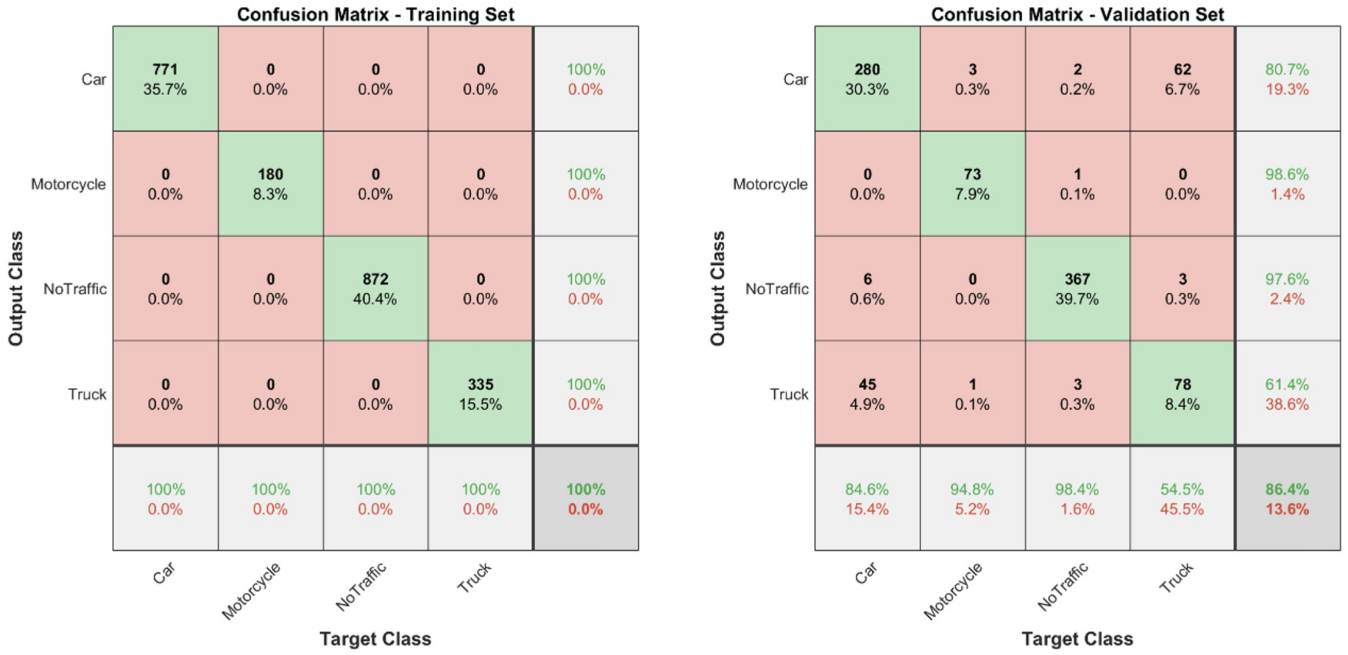
	Car	Motorcycle	NoTraffic	Truck	
Car	771 35.7%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
Motorcycle	0 0.0%	180 8.3%	0 0.0%	0 0.0%	100% 0.0%
NoTraffic	0 0.0%	0 0.0%	872 40.4%	0 0.0%	100% 0.0%
Truck	0 0.0%	0 0.0%	0 0.0%	335 15.5%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%

Confusion Matrix - Validation Set

Output Class / Target Class

	Car	Motorcycle	NoTraffic	Truck	
Car	273 29.5%	0 0.0%	5 0.5%	68 7.4%	78.9% 21.1%
Motorcycle	0 0.0%	75 8.1%	1 0.1%	2 0.2%	96.2% 3.8%
NoTraffic	7 0.8%	0 0.0%	367 39.7%	0 0.0%	98.1% 1.9%
Truck	51 5.5%	2 0.2%	0 0.0%	73 7.9%	57.9% 42.1%
	82.5% 17.5%	97.4% 2.6%	98.4% 1.6%	51.0% 49.0%	85.3% 14.7%

Figure 6. Training and Testing Confusion Plots for Spectrogram Features



Confusion Matrix - Training Set

Output Class / Target Class

	Car	Motorcycle	NoTraffic	Truck	
Car	771 35.7%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
Motorcycle	0 0.0%	180 8.3%	0 0.0%	0 0.0%	100% 0.0%
NoTraffic	0 0.0%	0 0.0%	872 40.4%	0 0.0%	100% 0.0%
Truck	0 0.0%	0 0.0%	0 0.0%	335 15.5%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%

Confusion Matrix - Validation Set

Output Class / Target Class

	Car	Motorcycle	NoTraffic	Truck	
Car	280 30.3%	3 0.3%	2 0.2%	62 6.7%	80.7% 19.3%
Motorcycle	0 0.0%	73 7.9%	1 0.1%	0 0.0%	98.6% 1.4%
NoTraffic	6 0.6%	0 0.0%	367 39.7%	3 0.3%	97.6% 2.4%
Truck	45 4.9%	1 0.1%	3 0.3%	78 8.4%	61.4% 38.6%
	84.6% 15.4%	94.8% 5.2%	98.4% 1.6%	54.5% 45.5%	86.4% 13.6%

Figure 7. Training and Testing Confusion Plots for Scalogram Features

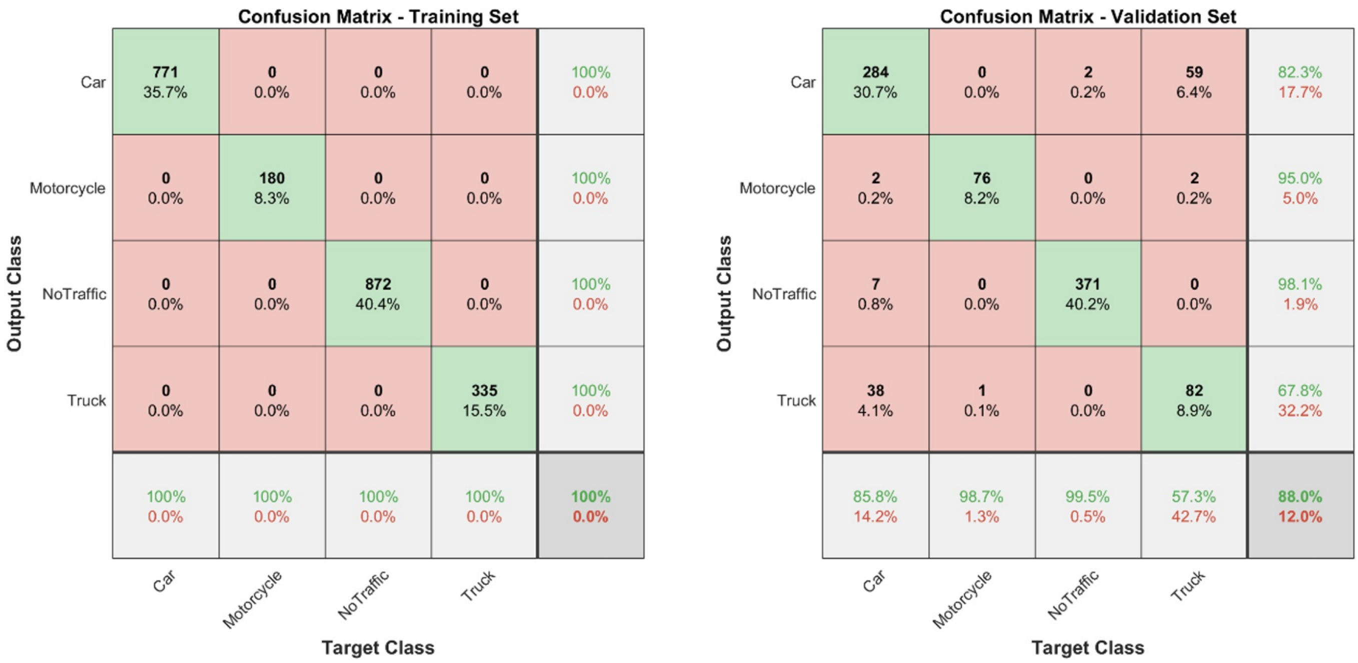


Figure 8. Training and Testing Confusion Plots for Fusion Features

Conclusions

In conclusion, our study explored innovative methodologies in on-road vehicle classification, addressing the critical need for cost-effective, accurate, and adaptable systems. We introduced an acoustic-based classification approach utilizing visual representations of sound—specifically spectrograms, scalograms, and their fusion—classified with a modified VGG16 CNN. Our findings demonstrate that this novel method holds considerable promise, providing a non-intrusive, potentially less costly, and environmentally adaptable alternative to traditional sensor-based and computer vision techniques. The fusion of scalogram and spectrogram features emerged as a superior approach, offering enhanced accuracy and reliability in distinguishing between vehicle types. The performance of the models, reflected in the training and loss metrics as well as the precision and recall of the classes, supports the hypothesis that a richer feature set leads to better classification outcomes. Despite the challenges of differentiating closely related vehicle classes such as 'Cars' and 'Trucks', the fusion-trained model showed a marked improvement, indicating the potential of our approach to refine and expand the capabilities of vehicle classification systems.

As we move forward, the continuous evolution of audio processing and machine learning technologies promises further advancements, paving the way for sound-based vehicle classification to become a cornerstone in the development of intelligent traffic monitoring and management systems. Further research and real-world testing will be essential to optimize this technology and explore its integration with existing infrastructure, marking a significant step toward more intelligent, efficient, and safe roadways.

Acknowledgements

The authors would like to thank Universiti Teknologi Mara, Malaysia for financial support for the research.

References

1. ^C.-J. Lin, S.-Y. Jeng, and H.-W. Lioa, "A Real-Time Vehicle Counting, Speed Estimation, and Classification System Based on Virtual Detection Zone and YOLO," Math Probl Eng, vol. 2021, pp. 1–10, Nov. 2021, doi: 10.1155/2021/1577614.
2. ^Y. Song et al., "Road-Users Classification Utilizing Roadside Light Detection and Ranging Data," Dec. 2020. doi: 10.4271/2020-01-5150.
3. ^S. Maity, A. Bhattacharyya, P. K. Singh, M. Kumar, and R. Sarkar, "Last Decade in Vehicle Detection and Classification: A Comprehensive Survey," Archives of Computational Methods in Engineering, vol. 29, no. 7, pp. 5259–5296, Nov. 2022, doi: 10.1007/s11831-022-09764-1.
4. ^Z. Chen, T. Ellis, and S. A. Velastin, "Vehicle type categorization: A comparison of classification schemes," in 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), IEEE, Oct. 2011, pp. 74–79. doi: 10.1109/ITSC.2011.6083075.
5. ^D. ul Khairi, F. Ayaz, N. Saeed, K. Ahsan, and S. Z. Ali, "Analysis of Deep Convolutional Neural Network Models for the Fine-Grained Classification of Vehicles," Future Transportation, vol. 3, no. 1, pp. 133–149, Jan. 2023, doi: 10.3390/futuretransp3010009.
6. ^N. G. Ripoll, L. E. G. Aguilera, F. M. Belenguer, A. M. Salcedo, and F. J. Ballester Merelo, "Design, Implementation, and Configuration of Laser Systems for Vehicle Detection and Classification in Real Time," Sensors, vol. 21, no. 6, p. 2082, Mar. 2021, doi: 10.3390/s21062082.
7. ^H. Shokravi, H. Shokravi, N. Bakhary, M. Heidarrezaei, S. S. Rahimian Koloor, and M. Petrů, "A Review on Vehicle Classification and Potential Use of Smart Vehicle-Assisted Techniques," Sensors, vol. 20, no. 11, p. 3274, Jun. 2020, doi: 10.3390/s20113274.
8. a, b, c, dI. Corera, E. Piñeiro, J. Navallas, M. Sagues, and A. Loayssa, "Long-Range Traffic Monitoring Based on Pulse-Compression Distributed Acoustic Sensing and Advanced Vehicle Tracking and Classification Algorithm," Sensors, vol. 23, no. 6, Mar. 2023, doi: 10.3390/s23063127.
9. ^R. Ma, Z. Zhang, Y. Dong, and Y. Pan, "Deep Learning Based Vehicle Detection and Classification Methodology Using Strain Sensors under Bridge Deck," Sensors, vol. 20, no. 18, p. 5051, Sep. 2020, doi: 10.3390/s20185051.
10. ^P. Sarcevic, S. Pletl, and A. Odry, "Real-Time Vehicle Classification System Using a Single Magnetometer," Sensors, vol. 22, no. 23, p. 9299, Nov. 2022, doi: 10.3390/s22239299.
11. ^H. Zhao, D. Wu, M. Zeng, and S. Zhong, "A Vibration-Based Vehicle Classification System using Distributed Optical Sensing Technology," Transportation Research Record: Journal of the Transportation Research Board, vol. 2672, no. 43, pp. 12–23, Dec. 2018, doi: 10.1177/0361198118775840.
12. ^M. Abdel-Aty, Z. Wang, O. Zheng, and A. Abdelraouf, "Advances and applications of computer vision techniques in vehicle trajectory generation and surrogate traffic safety indicators," Accid Anal Prev, vol. 191, p. 107191, Oct. 2023, doi: 10.1016/j.aap.2023.107191.

13. a, b M. A. Berwo et al., *"Deep Learning Techniques for Vehicle Detection and Classification from Images/Videos: A Survey," Sensors, vol. 23, no. 10, p. 4832, May 2023, doi: 10.3390/s23104832.*
14. ^ *P. Premaratne, I. Jawad Kadhim, R. Blacklidge, and M. Lee, "Comprehensive review on vehicle Detection, classification and counting on highways," Neurocomputing, vol. 556, p. 126627, Nov. 2023, doi: 10.1016/j.neucom.2023.126627.*
15. ^ *O. E. A. Agudelo, C. E. M. Marín, and R. G. Crespo, "Correction to: Sound measurement and automatic vehicle classification and counting applied to road traffic noise characterization," Soft comput, vol. 25, no. 18, pp. 12089–12089, Sep. 2021, doi: 10.1007/s00500-021-05852-9.*
16. ^ *K.-H. N. Bui, H. Oh, and H. Yi, "Traffic Density Classification Using Sound Datasets: An Empirical Study on Traffic Flow at Asymmetric Roads," IEEE Access, vol. 8, pp. 125671–125679, 2020, doi: 10.1109/ACCESS.2020.3007917.*
17. ^ *A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," Artif Intell Rev, vol. 53, no. 8, pp. 5455–5516, Dec. 2020, doi: 10.1007/s10462-020-09825-6.*
18. a, b *U. Mittal and P. Chawla, "Acoustic Based Emergency Vehicle Detection Using Ensemble of deep Learning Models," Procedia Comput Sci, vol. 218, pp. 227–234, 2023, doi: 10.1016/j.procs.2023.01.005.*
19. a, b *A. A. Lisov, A. Z. Kulganatov, and S. A. Panishev, "Using convolutional neural networks for acoustic-based emergency vehicle detection," Modern Transportation Systems and Technologies, vol. 9, no. 1, pp. 95–107, Mar. 2023, doi: 10.17816/transsyst20239195-107.*
20. a, b *C.-Y. Chiang, M. Jaber, K. K. Chai, and J. Loo, "Distributed Acoustic Sensor Systems for Vehicle Detection and Classification," IEEE Access, vol. 11, pp. 31293–31303, 2023, doi: 10.1109/ACCESS.2023.3260780.*
21. a, b *M. Ashhad, U. Goenka, A. Jagetia, P. Akhtari, S. K. Ambat, and M. Samuel, "Improved Vehicle Sub-type Classification for Acoustic Traffic Monitoring," in 2023 National Conference on Communications, NCC 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/NCC56989.2023.10067994.*
22. a, b *Z. Ye et al., "Traffic flow and vehicle speed monitoring with the object detection method from the roadside distributed acoustic sensing array," Front Earth Sci (Lausanne), vol. 10, Jan. 2023, doi: 10.3389/feart.2022.992571.*
23. a, b, c, d *A. I. Yassin, K. K. M. Shariff, M. A. Kechik, A. M. Ali, and M. S. M. Amin, "Acoustic Vehicle Classification Using Mel-Frequency Features with Long Short-Term Memory Neural Networks," TEM Journal, vol. 12, no. 3, pp. 1490–1496, Aug. 2023, doi: 10.18421/TEM123-29.*
24. a, b, c *L. Sun, Z. Zhang, H. Tang, H. Liu, and B. Li, "Vehicle Acoustic and Seismic Synchronization Signal Classification Using Long-Term Features," IEEE Sens J, vol. 23, no. 10, pp. 10871–10878, May 2023, doi: 10.1109/JSEN.2023.3263572.*
25. ^ *K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in 3rd International Conference on Learning Representations, 2015, pp. 1–14.*
26. ^ *Diederik P. Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization," in Proc. 3rd International Conference for Learning Representations, 2015, pp. 1–15.*