electronics

MDPI

Article

HADNet: A Novel Lightweight Approach for Abnormal Sound Detection on Highway Based on 1D Convolutional Neural Network and Multi-Head Self-Attention Mechanism

Cong Liang [1,†], Qian Chen [2,†], Qiran Li [1], Qingnan Wang [3], Kang Zhao [4], Jihui Tu [1,*] and Ammar Jafaripournimchahi [5]

[1] School of Electronic Information and Electrical Engineering, Yangtze University, Jingzhou 434023, China; lcong.stu@yangtzeu.edu.cn (C.L.); lqr_licoding@163.com (Q.L.)
[2] Department of Traffic Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; qianchen@usst.edu.cn
[3] School of Physics, Electronics and Intelligent Manufacturing, Huaihua University, Huaihua 418099, China; wql@hhtc.edu.cn
[4] Yunnan Provincial Land and Resources Information Center, Kunming 650032, China; kzhao@whu.edu.cn
[5] Automotive Engineering Research Institute, Jiangsu University, Zhenjiang 212013, China; 1000005557@ujs.edu.cn
* Correspondence: tujh@yangtzeu.edu.cn
† These authors contributed equally to this work.

Abstract: Video surveillance is an effective tool for traffic management and safety, but it may face challenges in extreme weather, low visibility, areas outside the monitoring field of view, or during nighttime conditions. Therefore, abnormal sound detection is used in traffic management and safety as an auxiliary tool to complement video surveillance. In this paper, a novel lightweight method for abnormal sound detection based on 1D CNN and Multi-Head Self-Attention Mechanism on the embedded system is proposed, which is named HADNet. First, 1D CNN is employed for local feature extraction, which minimizes information loss from the audio signal during time-frequency conversion and reduces computational complexity. Second, the proposed block based on Multi-Head Self-Attention Mechanism not only effectively mitigates the issue of disappearing gradients, but also enhances detection accuracy. Finally, the joint loss function is employed to detect abnormal audio. This choice helps address issues related to unbalanced training data and class overlap, thereby improving model performance on imbalanced datasets. The proposed HADNet method was evaluated on the MIVIA Road Events and UrbanSound8K datasets. The results demonstrate that the proposed method for abnormal audio detection on embedded systems achieves high accuracy of 99.6% and an efficient detection time of 0.06 s. This approach proves to be robust and suitable for practical applications in traffic management and safety. By addressing the challenges posed by traditional video surveillance methods, HADNet offers a valuable and complementary solution for enhancing safety measures in diverse traffic conditions.

Keywords: abnormal audio detection; 1D CNN; self-attention mechanism; transformer; embedded system

check for updates

Citation: Liang, C.; Chen, Q.; Li, Q.; Wang, Q.; Zhao, K.; Tu, J.; Jafaripournimchahi, A. HADNet: A Novel Lightweight Approach for Abnormal Sound Detection on Highway Based on 1D Convolutional Neural Network and Multi-Head Self-Attention Mechanism. *Electronics* **2024**, *13*, 4229. https://doi.org/10.3390/electronics13214229

Academic Editor: Alexander Barkalov

Received: 13 September 2024
Revised: 18 October 2024
Accepted: 24 October 2024
Published: 28 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The improvement in people's standard of living has led to a growing demand for travel and transportation, which results in a continuous increase in the number of vehicles on our roadways. However, the rising frequency of traffic accidents has raised concerns, it prompts an increased demand for enhanced road safety measures. The ability to detect traffic anomalies can rapidly trigger the intervention of emergency response teams, which ensures the safety of individuals. Additionally, the timely detection of abnormal sounds can help mitigate the impact of traffic accidents by enabling quicker interventions. Therefore, road surveillance has become increasingly important. Currently, existing highway monitoring equipment predominantly relies on video surveillance as its primary method [1]. This

approach enables clear observation of abnormal occurrences, and exhibits a high detection rate for such events [2,3]. However, in certain situations, visual information alone may not be sufficient to accurately comprehend the vehicle's behavior or identify hazardous conditions, especially in cases of low visibility or when monitoring is obstructed [4,5]. One of the major advantages of audio detection over video detection is that it does not require light or visibility, it performs better at night or in special weather, and it can detect anomalies outside the camera's field of view. In such cases, incorporating road audio signal detection as a complementary tool to video analysis can enhance the detection capabilities of security systems [6,7]. Therefore, there is an urgent need for an audio detection method to tackle the growing challenge of highway anomaly detection, which has consequently become a worldwide research hotspot [8].

Nowadays, various road audio detection methods have been proposed, which fall into three main categories: Signals Analysis approaches [9], Machine Learning approaches [10], and Deep Learning approaches [11]. Signal Analysis methods involve converting sound signals into frequency bands and comparing them with the sound pressure of anomalous sounds stored in a database, which achieves classification. While this approach effectively compensates for the shortcomings of video detection, it often falls short in achieving the desired highway anomalous audio detection effect due to low-accuracy recognition. Machine Learning approaches have been proposed to enhance the accuracy of abnormal sound detection. In this approach, certain audio features, such as volume, energy, and over-zero rate, are manually extracted. These features are represented as vectors, and SVM classifiers are employed for training and classification [12]. However, Machine Learning methods are less robust and struggle to effectively distinguish abnormal sounds from background noise. Deep Learning methods have made significant progress in the field of abnormal audio detection [13], and achieve high accuracy in highway anomalous audio detection [14]. Despite their success, most works on abnormal sound detection based on Deep Learning have focused on achieving high accuracy and efficiency on high-performance servers or PCs [15]. This is mainly due to the large detection models used in these methods, which require a significant amount of computation and are difficult for deployment on embedded systems. This falls short of meeting the actual needs of practical applications. Consequently, developing an end-to-end abnormal sound detection method that fulfills the requirements of practical applications poses a significant challenge.

This paper focuses on three challenges for the practical application of anomalous sound detection on highways: (1) designing a lightweight network model, (2) overcoming the small sample problem, and (3) tackling the data imbalance issue. Firstly, while existing abnormal sound detection methods have indeed achieved satisfactory performance on high-performance servers, they come with a high computational cost and large size of the detection model. However, abnormal sound detection systems are commonly deployed on the environment of embedded systems, where considerations of system cost, the convenience of installation, and maintenance. Given that embedded devices often have limited hardware resources, existing abnormal sound detection approaches struggle to meet the practical needs of highway abnormal sound detection systems due to their complexity, numerous parameters, and large size [16]. Therefore, it is imperative to design a lightweight audio detection network model suitable for embedded environment, which can meet the real-time detection requirements of highway monitoring equipment. Secondly, crash and tire skid are prevalent vehicular hazards encountered on highway. Hence, the proposed highway anomaly sound detection system specifically addresses these two hazards. However, the limited number of available crash and skid events in the real environment poses a challenge for training with small samples. Finally, the limited nature of anomalous sounds causes the number of crash events, skid events and background noise samples we collect to be unbalanced. This imbalance may lead the model to overpredict a larger number of categories and ignore smaller ones, which results in biased evaluation results. Therefore, more attention should be paid to a few categories in the model. To address these challenges, we propose a lightweight network model for highway abnormal sound detection on the

embedded system environment, which is named the Highway Audio Detection Network (HADNet). We conducted a series of audio detection experiments using the MIVIA Road Events and UrbanSound8K datasets. The results showed that the accuracy and time of abnormal sound detection on the embedded system environment are 99.6% and 0.06 s, respectively. The experimental results presented in this paper demonstrate that the proposed method is more suitable for meeting the actual demands of abnormal audio detection.

In the light of the above considerations, the main contributions of our work can be summarized as follows:

- An end-to-end lightweight network based on a 1D Convolutional Neural Network (CNN) and Multi-Head Self-Attention Mechanism, called HADNet, is proposed for highway abnormal audio detection on the environment of embedding system. The 1D CNN is used to extract the local feature of audio, while the Multi-Head Self-Attention Mechanism has strong global modeling capabilities; hence, combining them allows for multi-level feature extraction, which enables the model to capture fine-grained details and understand global context simultaneously. Meanwhile, both the 1D CNN and Multi-Head Self-Attention Mechanism have strong parallel processing capabilities, which can accelerate computation, especially on large-scale data.
- To address the problem of small sample data, transfer learning is employed as a solution. The process involves pre-training the model on the large-scale UrbanSound8K dataset, followed by fine-tuning the model parameters using our dataset. The application of transfer learning has been shown to significantly enhance both the accuracy and robustness of the system.
- Additionally, the joint loss function is employed to address the problem of unbalanced sample data. This approach helps balance the dataset by adjusting the weights of the categories. Specifically, the classification performance is improved by increasing the weights assigned to anomalous sounds and decreasing the weights assigned to background noise.
- We evaluate the effectiveness of our proposed methodology for highway abnormal audio detection using the MIVIA Road Events and UrbanSound8K datasets in an embedded system environment. Our results demonstrate that the suggested approach exhibits superior quantitative and efficient performance compared to state-of-the-art techniques.

The paper is organized as follows: Section 2 describes the work related to this study. Section 3 details the implementation of the proposed method, including the feature extraction method for audio and the structure of HADNet, and highlights the reasons for our choice. Section 4 describes the dataset used in the experiments, details the experimental protocol and the results obtained, and compares them with existing methods. Section 5 concludes with conclusions and future work.

2. Related Works

In the field of intelligent transportation, various abnormal sound detection methods have been reported over the past few decades [17]. These efforts are divided into the following three stages.

Firstly, manual feature extraction methods were proposed for abnormal sound detection, such as volume, energy, and over-zero rate, coupled with classifiers like SVM, KNN, and Bayes [12,18]. While hand-crafted features demonstrated exemplary performance in previous works, they still need to improve due to poor robustness. Consequently, time-frequency transforms, such as MFCCs, log-mel spectrograms, and Gabor filter banks, were proposed to extract audio features. Since feature extraction in these methods is more biologically consistent, detection performance improved under adverse noise conditions. Nevertheless, it remains challenging to comprehensively present the characteristics of complex sounds, which results in relatively poor discriminability.

Secondly, methods combining CNNs with time-series approaches, such as LSTM and RNN, were proposed [19,20]. Since audio signals are time series data, time-series

approaches are used to extract the temporal feature. The CNN is used to extract local features from the spectrogram generated by the time-frequency transform. These methods achieve satisfactory results due to the extraction of spatial and temporal features, but the time-frequency transform suffers from information loss, and time-series methods have computational challenges, which results in limited real-time performance for the algorithm.

Thirdly, CNN and transformer encoder are employed for abnormal sound detection [13]. The CNN is still utilized to extract local features from the spectrogram generated by the time-frequency transform, while transformer encoder extracts temporal and global features. Due to the transformer encoder's solid, parallel processing capabilities, computation is accelerated. Consequently, the detection performance of abnormal sound has been improved compared with the previous stage, making this approach a hot topic.

To sum up, the following challenges remain for abnormal sound detection to be effectively applied in intelligent transportation: (1) Lightweight model. The abnormal sound detection model must be deployed on embedded systems for practical application. However, current algorithms cannot be deployed in embedded environments due to their high computational complexity. (2) End-to-end detection. Existing methods require an initial time-frequency conversion to convert the sound signal into a spectrogram. The conversion to a spectrogram blurs the characteristics of the anomalous sound, which not only distorts the audio features, but also incurs additional computational costs. Therefore, our paper proposes an end-to-end lightweight approach for abnormal sound detection on embedded systems.

3. Method

Accurate and rapid highway anomalous audio detection is of great importance to emergency responses and rescues during sudden traffic accidents. The key to effective audio detection is finding an appropriate lightweight model to balance detection accuracy and computational complexity on an embedded system. In this context, we proposed a novel lightweight network model named HADNet for detecting highway anomalous audio. The framework of our method is illustrated in Figure 1. It involves the following steps: (i) Local Feature Extraction: local features are extracted based on the 1D CNN from the raw audio [19,21]; (ii) Global Feature Extraction: the HADNet block, based on Multi-head attention mechanism, is employed to extract the global feature of audio; and (iii) Classification: audio classes are identified by employing the fully connected layer in conjunction with the SoftMax operation. This section provides a detailed description of the structure and implementation of HADNet.

3.1. One-Dimensional Convolutional Layer

At present, most methods for abnormal audio detection on highway rely on time-frequency conversion. This involves transforming the audio signal into various spectrograms, such as Fbank, MFCC, and audio spectrograms. Subsequently, multiple features are extracted from these spectrograms for audio classification. However, these methods have limitations: (1) Information loss. Time-frequency conversion leads to information loss, thereby reducing the accuracy of classification. (2) Computational time. The process of time-frequency conversion consumes computing time, which hinders real-time audio processing. In response to these challenges, we have adopted 1D convolution to directly extract audio features from the raw audio. This approach not only enhances detection accuracy, but also reduces computation time, which can make it more conducive to real-time audio processing.

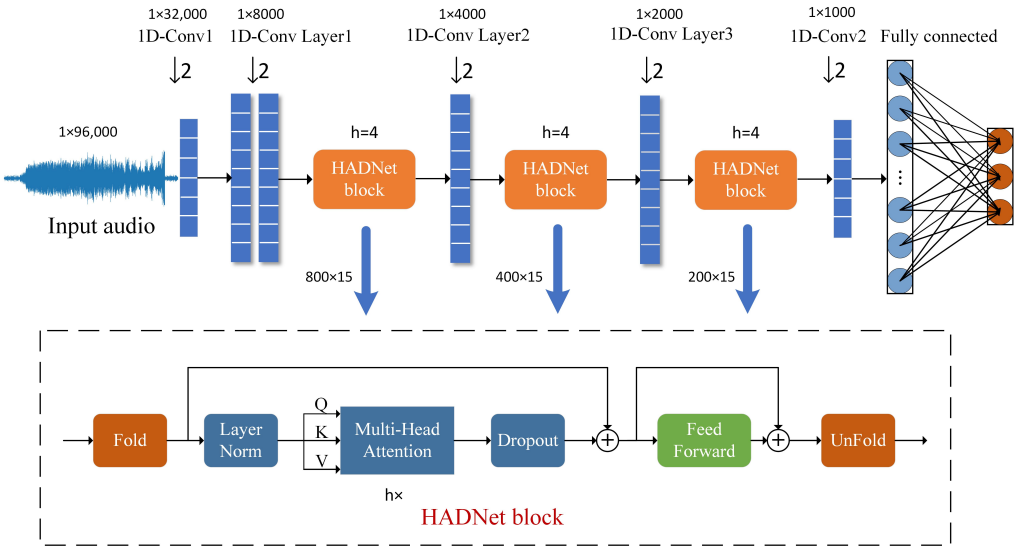


Figure 1. Flowchart of the methods.

The proposed 1D convolution model is depicted in Figure 1. It was designed with multiple hidden layers to learn high-level feature representations of audio signal [21]. HADNet consists of four 1D convolutional layers. The model takes 96,000 audio sampling points as input, the "2" in Figure 1 represents the downsampling factor for the audio MaxPooling operation. In addition, "1 × 32,000", "1 × 8000", etc. represent the output size of each audio layer. The first two 1D convolutional layers employ convolutional kernels of size 5, while the last two 1D convolutional layers use convolutional kernels of size 3. The downsampling factor for MaxPooling operations is set to 2. Table 1 provides details on the parameters and output sizes of each 1D convolutional layer in the HADNet model.

Table 1. Parameters of 1D-Conv in the HADNet.

Layer Type	Kernel Size	Number of Filters	Stride	Output Size
1D-Conv1	1 × 5	16	3	1 × 32,000
1D-Conv layer1	1 × 5	16	1	1 × 16,000
1D-Conv layer2	1 × 5	32	1	1 × 8000
1D-Conv layer3	1 × 3	64	1	1 × 4000
1D-Conv layer4	1 × 3	32	1	1 × 2000
1D-Conv2	1 × 3	16	2	1 × 1000

In the 1D convolutional layer, SiLU activation functions are applied after each convolution operation. Subsequently, MaxPooling is performed on the convolutional outputs to further reduce the dimensionality of the audio sequence. A spatial dropout layer is employed to introduce a controlled level of randomness and prevent overfitting, which randomly sets specific parameters in the audio sequence to zero. Finally, Batch Normalization is applied to normalize the output and accelerate the convergence of the network.

3.2. HADNet Block

After extracting features through 1D convolutional layers, the HADNet block is employed to learn global representations from the audio signals using a Multi-Head Self-Attention Mechanism. This approach minimizes the use of parameters and computational resources, thus making the proposed method more lightweight. The structure of the HADNet block is divided in the following three parts: (1) Fold operation: this operation transforms one-dimensional vectors into matrices after the convolutional part of HADNet; (2) Multi-Head Self-Attention Mechanism: this mechanism is used to extract the global features of the audio; and (3) Feed-Forward Neural Network layer: This layer is used for

audio classification. Each component of the HADNet block contributes to enhancing the detection accuracy and efficiency of the model, making it suitable for real-time applications in embedded systems.

3.2.1. Fold

To implement the input requirements of the Multi-Head Self-Attention Mechanism, the convolutional output of HADNet is transformed into a matrix by a sliding window, referred to as a Fold. The processing of the Fold is illustrated in Figure 2. Let F denote the output features of the convolution layer, where the size of F is $1 \times W$. The transformation matrix are obtained with the sliding window, the size of sliding window is N, the step length is $l = \frac{2 \times N}{3}$, and the overlap of adjacent windows is $\frac{N}{3}$. Since the sound signal is continuous, we incorporate a partial overlap between neighboring frames to prevent the loss of sound features. Therefore, the feature F is split into a $c \times k$ matrix through the Fold operation, where k = N, c $= \left| \frac{w}{l} \right| + 1$. The output of the Fold operation is fed into the Multi-Head Self-Attention, which is used to extract global features.

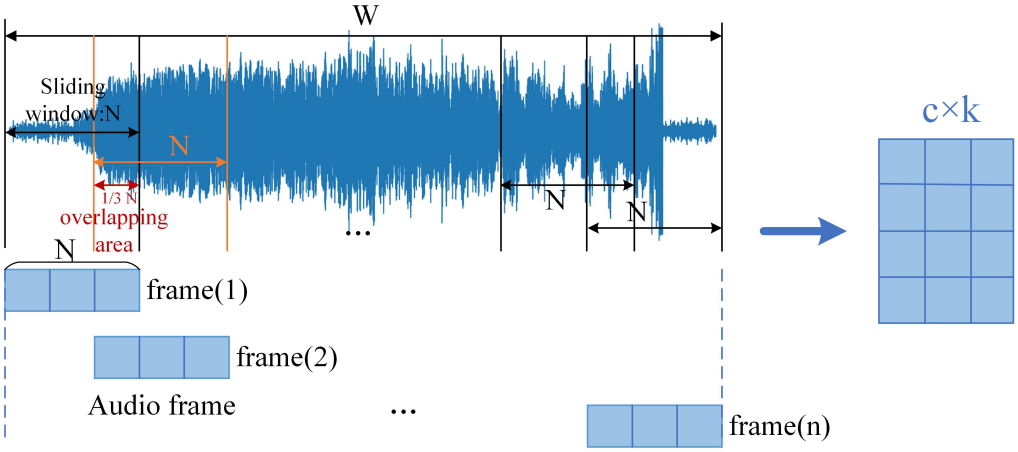


Figure 2. The processing figure of the Fold with a sliding window.

3.2.2. Multi-Head Self-Attention

Since Artificial Neural Networks (ANN) tend to lose significant amount of information during long-distance propagation, and are not very sensitive to feature importance, the attention mechanism proves effective in handling time series data. This mechanism allows neural networks to focus on relevant parts of the temporal sequence selectively; hence, the Multi-Head Self-Attention Mechanism is employed to weigh the importance of different time steps within a time series [22]. Based on the Self-Attention module, this mechanism extracts more expressive sequences and combines multiple attention representations for audio detection.

Figure 3 shows that multiple scaled dot-product attention modules are stacked together, which forms the Multi-Head Self-Attention Mechanism. Each attention module can express the weight of different parts of the input sequence, and these modules can operate in parallel. For the Multi-Head Self-Attention Mechanism, the input consists of three matrices: Queries (Q), Keys (K), and Values (V). Since Multi-Head Self-Attention uses Self-Attention (meaning the input sequence is both the Query and the Key), the matrices Q, K, and V have the same dimension. While one of the scaled dot product attention modules is running, the matrices Q, K, and V are linearly converted to create three sets of vectors for each head. These transformations are learned during training, and allow the model to adapt to the specific task. The following formula is used to calculate the Self-Attention mechanism:

$$Attention(q_i, K, V) = \sum_{j=1}^{m} softmax(\frac{q_i \cdot k_j^T}{\sqrt{d_k}})v_j \quad (1)$$

where the dot product between the input q_i and each vector k_j^T is used to calculate the similarity, resulting in the weight. The scale is computed by dividing the dot product by $\sqrt{d_k}$, which helps prevent the issue of gradient disappearance. $\sqrt{d_k}$ denotes the number of columns of the matrices Q, K, i.e., the dimensions of the vectors. The Softmax function is applied to normalize the weights, and the attention value of a block is obtained by computing the sum of the weights multiplied by their corresponding key values. The attention of the i-th space Head_i is calculated as

$$Head_i = Attention(W_i^Q Q, W_i^K K, W_i^V V) \tag{2}$$

where $W_i^Q \in R^{d \bmod el \cdot d_k}$, $W_i^K \in R^{d \bmod el \cdot d_k}$, and $W_i^V \in R^{d \bmod el \cdot d_k}$ are the parameter matrix required for the linear mapping of the Q, K, and V matrices in the i-th space, respectively. After the attention of each space is calculated, the head of each space is spliced and N heads are obtained. The multi-space attention MultiHead(Q,K,V) is calculated as follows:

$$MultiHead(Q, K, V) = Concat(Head_1, Head_2, ..., Head_h)W^0 \tag{3}$$

where $W^0 \in R^{d \bmod el \cdot d_k}$ is the weight of linear transformation, and Concat refers to the operation of concatenating vectors. The Multi-Head Attention Mechanism can learn additional semantic information from different spaces.

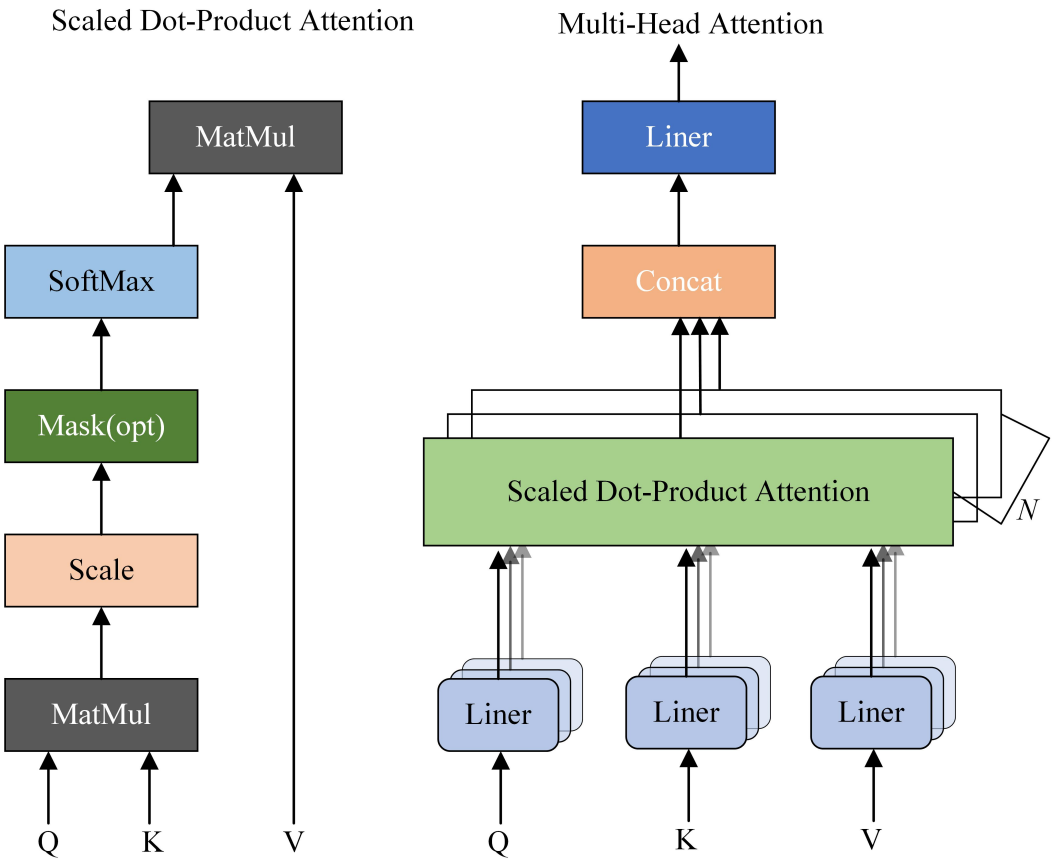


Figure 3. Structure diagram of Multi-Head Attention Mechanism.

3.2.3. Feed-Forward Neural Network

The output of the Multi-Head Self-Attention Mechanism is fed into a feed-forward neural network, which consists of an input layer, a hidden layer, and an output layer. The input layer receives the Multi-Head Self-Attention data, the hidden layers process

these data, and the output layer produces the final output. The FFN is calculated with the following formula:

$$FFN(x) = ((W_1x + b_1) \cdot Sigmoid(W_1x + b_1)) \cdot W_2 + b_2 \quad (4)$$

The Non-linear Activation Function of FFN is a Sigmoid-weighted Linear Unit (SiLU), where x is the input vector of FFN, W_1 and W_2 are their respective weights, b_1 and b_2 are their respective biases. These weights and biases are learned during the training process. FFN propagation flows in one direction, from the input layer to the output layer, without feedback loops. Each neuron in a given layer is connected to every neuron in the subsequent layer. The output of FFN is converted to a vector by Unfold operation.

3.3. Classification

Abnormal audio detection on the highway is actually a multi-classification problem. While traditional Softmax is widely used in audio classification, it has limitations. The Softmax loss function does not explicitly optimize the feature embedding to enforce higher similarity for intra-class samples and diversity for inter-class samples, which results in a performance degeneration for audio classification under large intra-class appearance variations. Therefore, instead of the traditional Softmax classifier, we employ a joint loss function for detecting abnormal audio. This approach is more effective in addressing the challenges posed by unbalanced training samples and overlapping classes.

The key idea of joint loss function is to introduce the center loss function into the traditional Softmax loss function; the center loss function not only learns the deep feature center of each class, but also penalizes the deviation of the expected feature from the corresponding class center. The goal of the joint loss function in this paper is used to maximize the interclass variance and minimize the intra-class variance, thereby enhancing the separability and discrimination of audio feature representation. The joint decision-making of the joint loss function further improves classification accuracy. The joint loss function is defined as follows:

$$\begin{aligned} L &= L_1 + \lambda L_2 \\ &= -\frac{1}{n}\left[\sum_{i=1}^{n} \log \frac{e^{W_{y(i)}^T x(i)}}{\sum_{i=1}^{m} e^{W_i^T x(i)}}\right] + \frac{1}{2}\sum_{i=1}^{n} \left\|x_i - c_{yi}\right\|_2^2 \end{aligned} \quad (5)$$

where λ is a hyper-parameter used to balance the weights between Softmax loss function L_1 and the center loss function L_2. When λ is zero, L is solely the Softmax loss. The variable m represents the total number of classes, $x(i) \in R^d$ denotes the deep feature of the i-th sample belonging to the y_i-th class, c_{yi} is the sample center of the class y_i. The embedding feature dimension d is set to 1000 in this paper. $W \in R^{d \times n}$, $W_{y(i)}^T$ denotes the j-th column of the weight matrix W, and N is the number of classes. The model parameter λ is adjusted by optimizing the joint loss function to obtain a better audio feature representation.

4. Experiments

To demonstrate the effectiveness of the HADNet model, a series of experiments were conducted to evaluate the accuracy, efficiency, and parameter settings of HADNet. The specific experimental design is as follows.

4.1. Dataset, Parameters Settings, and Performance Metric

4.1.1. Dataset

The performance of the HADNet model was evaluated using the MIVIA Road Events [23] and UrbanSound8K [24] datasets. MIVIA Road Events contains 200 tire skid events, 200 crash events, and background noise. Meanwhile, the UrbanSound8K dataset contains 8732 labeled sound excerpts ($\leq$4 s) of urban sounds from 10 classes: air_conditioner, car_horn, chil-

dren_playing, dog_bark, drilling, enginge_idling, gun_shot, jackhammer, siren, and street_music. Anomalous sounds on highways often overlap with background noise in real environments. To address this, we employed a method to superimpose various background noises onto anomalous sound events. This approach allows us to more accurately simulate the occurrence of anomalous events in the real world [25]. Consequently, background noises from the UrbanSound8K dataset are overlaid onto the MIVIA Road Events dataset [26]. These anomalous sounds include skids and crashes, while the background noise consists of elements like wind, rain, and sirens. To avoid any overlap between the training and test sets, we first split the 200 tire skid events and 200 crash events into training and test sets in a 7:3 ratio. After the split, background noise was added separately to the training and test sets. The function of superposition is outlined as follows:

$$out_j(n) = b_j(n) \oplus_{[s_i, e_i]} x_i(n) \tag{6}$$

where $\oplus_{[s_i, e_i]}$ is an operator that superimposes the anomalous sound signal $x_i(n)$ and the background sound signal $b_j(n)$ in the time interval $[s_i, e_i]$, with s_i and e_i are the start and end points of the abnormal sound, respectively.

Our dataset comprises 4000 audio events synthesized from MIVIA Road Events and UrbanSound8K. Each audio event, representing a continuous sound signal, has been segmented into 3-s audio clips. This segmentation results in a total audio duration of 200 min, utilizing a sampling rate of 32 kHz and expressed as 16-bit quantization per PCM sample. To enhance the dataset's variability, we have recorded each audio clip at various signal-to-noise ratios, specifically at -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, 25 dB, and 30 dB. This variation is intended to simulate different distances from the microphone [27]. The dataset comprises 3000 audio clips designated for training. Another 1000 audio clips are allocated for testing. Within the training set, we further divided the samples: samples are randomly divided into training samples (70%) and validation samples (30%). Detailed information on the dataset composition is provided in Table 2.

Table 2. Number of recordings and total duration of various audio events in the dataset.

Type	Train Set		Test Set	
	Events	Duration (s)	Events	Duration (s)
Skid	1000	3000	300	900
Crash	1000	3000	300	900
Background	1000	3000	400	1200

4.1.2. ENVIRONMENT and Parameters Settings

The training of the HADNet model occurred on a server, while the actual execution was performed on an embedded system. Therefore, the experimental environment in our paper is divided into three parts (see Table 3): the server, Raspberry Pi 4B, and Jeston Nano B01. The proposed method is implemented by using the PyTorch1.12.1 framework. To enhance the accuracy and adaptability of our model, we have incorporated transfer learning and fine-tuning techniques. Specifically, we utilized a pre-trained model that was fine-tuned on UrbanSound8K dataset. This approach leveraged the initial weights and learned representations, allowing us to achieve superior performance with reduced computational resources. During fine-tuning, a batch size of 16 and a fixed learning rate of 0.0002 were maintained, optimized through AdamW optimizer over 60 epochs. Random data augmentation techniques were also employed to mitigate overfitting, ensuring robust performance. Additionally, the balancing parameter λ in the joint loss function Equation (5) was carefully tuned to 0.002 to optimize model training for deployment on embedded systems.

Table 3. Experimental system environment (* indicates non-existence).

	Server	**Embedded System**	**Embedded System**
Model	Dell PowerEdge R730	Raspberry Pi 4B	Jeston Nano B01
CPU	Intel(R) Xeon(R) @2.20 GHz	ARMv7 1.5 GHz	ARM A57 @1.43 GHz
GPU	GeForce RTX2080Ti	*	128-core Maxwell
System	CentOS Linux 7	GNU/Linux 10	Ubuntu 18.04.6 LTS
RAM	11G	8G	4G
Language	Python 3.7.0	Python 3.7.0	Python 3.6.9
Library	Pytorch 1.10.1	Pytorch 1.12.1	Pytorch 1.7.1

4.1.3. Performance Metrics

The primary metric for assessing the accuracy of highway abnormal sound detection experiments is the recognition rate (RR) of abnormal sound events. In this evaluation, if audio containing abnormal sounds is incorrectly classified as background, it is considered a missed event. Conversely, if background noise is misclassified as an abnormal sound event, it results in a false alarm, simulating the occurrence in a real surveillance system. To measure detection accuracy, the following metrics are employed:

① Recognition Rate (RR): The ratio of correctly detected events to the total number of events. The formula calculation is as follows:

$$RR = \frac{TP + TN}{N} \tag{7}$$

② Miss Rate (MR): The ratio of abnormal events mistakenly detected as background to the total number of events. The formula calculation is as follows:

$$MR = \frac{FN}{N} \tag{8}$$

③ Error Rate (ER): The ratio of misclassified detected abnormal events to the total number of events. The formula calculation is as follows:

$$ER = \frac{FTP}{N} \tag{9}$$

④ False Alarm Rate (FAR): The ratio of background sounds incorrectly identified as abnormal events to the total number of events. The formula calculation is as follows:

$$FAR = \frac{FP}{N} \tag{10}$$

In the context of evaluating the detection accuracy, the following metrics are utilized:

- True Positive (TP): this value signifies the total count of accurately predicted abnormal events classified as positive cases.
- True Negative (TN): represents the total count of correctly predicted normal events classified as negative cases.
- False Positive (FP): denotes the number of normal events incorrectly identified as positive cases.
- False True Positive (FTP): represents the number of detected abnormal events misclassified.
- False Negative (FN): represents the number of abnormal events erroneously classified as negative cases.

These metrics comprehensively evaluate the detection accuracy, encompassing both actual positive and false positive scenarios. They are crucial for assessing the detection model's performance and identifying areas for improvement.

4.2. Accuracy Evaluation

In this experiment, four indices (Recognition Rate—RR, Miss Rate—MR, Error Rate—ER, and False Alarm Rate—FAR) are employed to comprehensively evaluate the performance of the proposed method. This evaluation includes a comparative analysis of existing abnormal sound detection methods on the highway based on both Machine Learning and Deep Learning. The chosen metrics offer a robust assessment, which can capture various aspects of the detection process.

First, we evaluated four indices for lightweight models based on MobileNetV1 [28], MobileNetV2 [29], MobileNetV3 [30], and MobileViT [31]. The experimental results are presented in Table 4. It can be seen from Table 4 that all of the lightweight models have achieved good results. Since the existing lightweight models for detecting abnormal sounds require time-frequency conversion, it will lead to the inevitable loss of audio information. In contrast, our proposed method uses 1D CNN to directly extract features from audio information, which avoids such information loss. Therefore, our method achieves more accurate results.

Table 4. Comparison of the proposed method with the lightweight approaches in terms of Recognition Rate (RR), Miss Rate (MR), Error Rate (ER), and False Alarm Rate (FAR).

Method	RR (%)	MR (%)	ER (%)	FAR (%)
HADNet (ours)	99.6	0.3	0.0	0.1
MobileViT-xxs	98.0	1.8	0.0	0.2
MobileViT-xs	99.1	0.4	0.0	0.5
MobileViT-s	99.0	0.2	0.0	0.8
MobileNetV1	98.8	0.5	0.2	0.5
MobileNetV2	99.5	0.4	0.0	0.1
MobileNetV3	99.4	0.4	0.0	0.2

Second, we also compared the proposed method to abnormal sound detection with other methods. The study tested the proposed method against the BoW model method in reference [32], the CNN and LSTM method in reference [20], and the CNN and GRU in reference [33]. All methods were tested using the same dataset and evaluated using the same assessment indicators. The experimental results are shown in Table 5. In the BoW model method [23], traditional feature extraction and an SVM classifier were employed to detect abnormal sounds on the highway. However, artificial feature extraction for constructing audio signal features made it challenging to achieve satisfactory detection results, and exhibited poor generalization ability. With the CNN and LSTM methods, and the CNN and GRU methods [32], ref. [20] employed time-frequency transform and Deep Learning techniques to extract high-level features of audio signals. While these methods achieved satisfactory results, time-frequency conversion led to information loss, and single CNN and LSTM models [33] suffered from incomplete feature extraction issues. Our proposed method, on the other hand, outperforms all competitive methods on the dataset. This is because our method utilizes both local and global features based on 1D CNN and the Multi-Head Self-Attention Mechanism, allowing for direct detection of abnormal sound from raw audio signals. Moreover, our approach achieves higher detection accuracy without signal loss, although it may have a slightly higher False Alarm Rate (FAR). In the next section, we will discuss the advancements of our approach in terms of detection efficiency.

Third, considering the limited amount of data in the dataset and the uneven distribution of categories, we used 5-fold cross-validation to ensure the consistency of the results across different data partitions, thus improving the adequacy and comprehensiveness of the evaluation. Each dataset is divided into five parts, four of which are used for training and the rest for testing. This approach not only evaluates the robustness of our method, but also estimates its performance more reliably, especially for small, unbalanced datasets. Figure 4 illustrates the cross-validation results, highlighting the performance metrics obtained in the

different divisions. These findings emphasize the effectiveness of the method in different data partitions and its usefulness.

Table 5. Comparison of the proposed method with the state-of-the-art approaches in terms of Recognition Rate (RR), Miss Rate (MR), Error Rate (ER), and False Alarm Rate (FAR).

Method	RR (%)	MR (%)	ER (%)	FAR (%)
HADNet (ours)	99.6	0.3	0.0	0.1
SVM	82.0	17.7	0.3	2.9
AReN	98.0	0.0	0.0	2.0
BLSTM	92.2	3.4	1.2	3.2
DENet	98.2	0.2	0.0	1.6

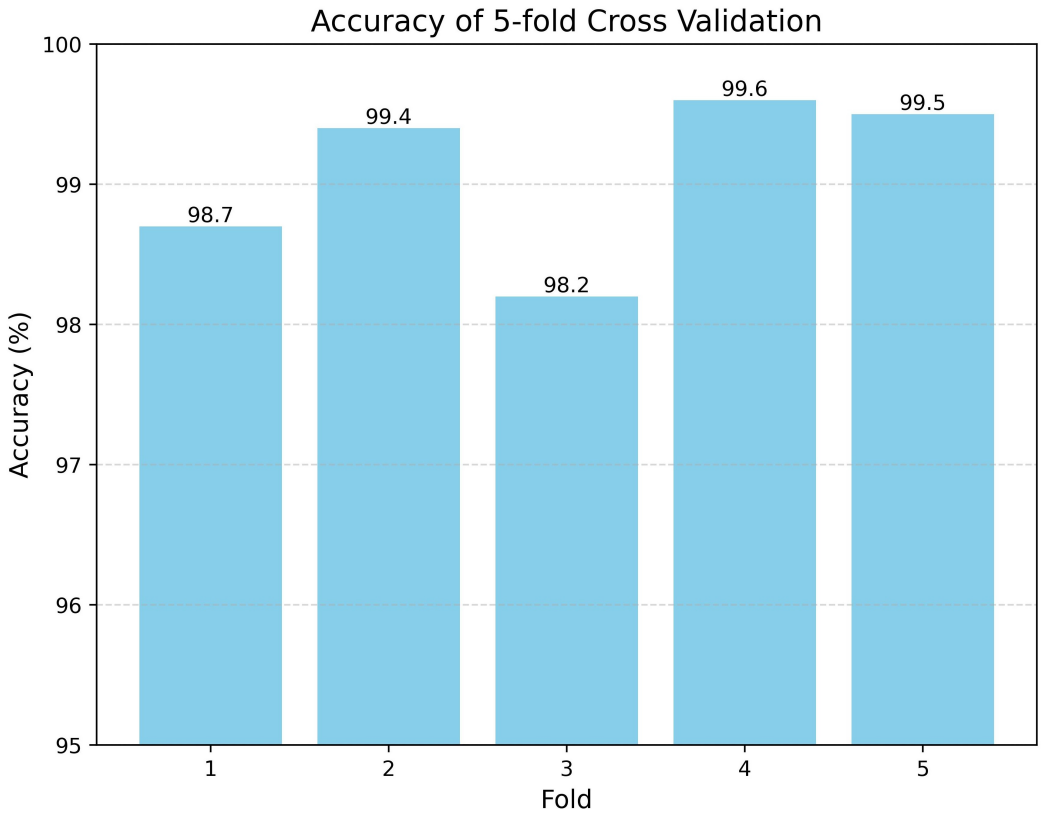


Figure 4. Results of the 5-fold cross-validation.

4.3. Efficiency Evaluation

In addition to precision assessment, another crucial aspect of detecting abnormal sounds on the highway is the detection speed. Real-time detection in the environment of an embedding system is advantageous for practical applications. Table 6 shows the run times of various methods used for abnormal sound detection over different hardware platforms. AReN and DENet have good detection efficiency on servers and PCs, but they cannot be deployed on embedded devices. The lightweight methods MobileVit, MobileNet, and our method achieve good results under parallel computation of GPU, but MobileVit and MobileNet consume some time to transform audio into Fbank feature spectrograms. The experimental results show that the proposed method have relatively higher running speeds than the other existing methods on the Raspberry Pi and Jeston Nano B01. Furthermore, the proposed method exhibits the smallest complexity and the highest accuracy.

Existing state-of-the-art methods AReN and DENet also achieve high detection efficiencies on the server, but compared to HADNet, DENet takes 71 milliseconds to detect 2.5 s of audio [33], while HADNet takes only 15 ms to detect 3 s of audio. When using PC

detection, DENet takes 182 ms to detect a 2.5-s tone [33], and AReN takes 200 ms to detect a 3-s tone [32]. It is worth noting that HADNet's detection time on the Jetson Nano B01 is only 60 ms, which effectively meets the real-time detection needs of embedded devices. In addition, HADNet obtains the smallest Params (0.27M) and FLOPs (0.23G) compared to all the inspected models. Therefore, HADNet is an excellent lightweight model suitable for deployment in embedded systems.

Table 6. Running time, Params and FLOPs of highway abnormal sound detection (* indicates non-existence).

Method	Server (ms)	Raspberry Pi 4B (ms)	Jetson Nano B01 (ms)	Params (M)	FLOPs (G)
HADNet	15	68	60	1.27	0.23
MobileViT-XXS	37	135	118	1.3	0.26
MobileViT-XS	39	194	138	2.3	0.7
MobileViT-S	40	230	140	5.6	1.44
MobileNetV1	38	155	125	4.2	0.57
MobileNetV2	36	122	118	3.5	0.32
MobileNetV3	34	130	116	5.48	0.23
AReN	78	200 (PC)	*	*	*
DENet	71	182 (PC)	*	*	*

4.4. Analysis of the Sensitivity to the Noise

Another worthwhile assessment to characterize the performance of HADNet involves analyzing its sensitivity to environmental noise. Experimental tests were conducted to evaluate the recognition rate of different models for anomalous sound events at varying SNR levels, which ranges from −5 dB to 30 dB. The manipulation of the SNR emulates real environmental scenarios, accounting for variations in both the distance between the detection device and the sound source and the magnitude of the signal energy [34].

The trend of the recognition rate with SNR is shown in Figure 5, where the recognition rate increases with the increase in SNR in all models. In fact, when the SNR exceeds 10 dB, the network model reaches high recognition performance and stability. However, at 5 dB, a significant decrease in the recognition rate is observed. Importantly, compared to other models, HADNet demonstrates superior noise immunity and enhanced robustness to environmental noise. This highlights its effectiveness in maintaining optimal performance, even in challenging noise conditions.

In Table 7, the detection ability of HADNet for abnormal sounds under different SNR levels is presented. With an SNR of 5dB, the network demonstrates a high miss rate (1.9%) and false alarm rate (3.0%), which is significantly affecting the detection of abnormal sounds in real environments. However, with an SNR greater than 10 dB, the network exhibits a notable reduction in both Miss Rate and False Alarm Rate for abnormal sounds. This observation underscores the ability of HADNet to maintain optimal performance even in noisy environments.

Table 7. Results of HADNet at different SNR values (from −5 to 30 dB).

SNR	RR (%)	MR (%)	ER (%)	FAR (%)
−5 dB	86.4	2.8	4.3	6.5
0 dB	89.6	2.3	3.6	4.5
5 dB	92.3	1.9	2.8	3.0
10 dB	96.2	0.7	1.2	1.9
15 dB	98.7	0.4	0.3	0.6
20 dB	99.4	0.3	0.0	0.3
25 dB	99.6	0.3	0.0	0.1
30 dB	99.6	0.3	0.0	0.1

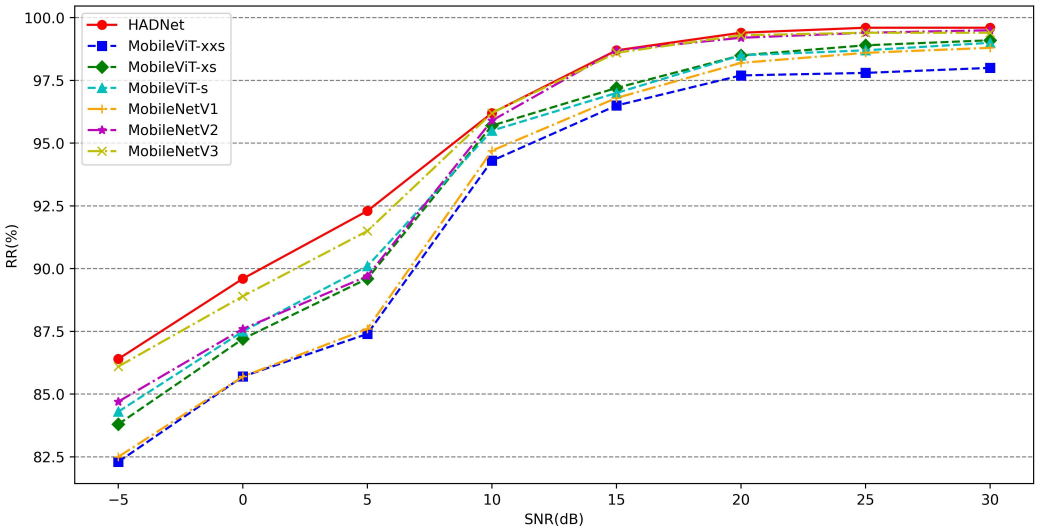


Figure 5. Recognition rate of HADNet and other lightweight models at different SNRs (from −5 to 30 dB).

4.5. Parameter Analysis

In the proposed method for abnormal sound detection, the balance parameter λ in the joint loss function Equation (5) plays a crucial role, and requires manual adjustment. The accuracy of detection is greatly influenced by this parameter. In our experiments, we examined various values of λ to evaluate their impact on detection accuracy, assessing the sensitivity of λ to the detection results. Figure 6 presents the experimental results for the dataset. The highest accuracy for the data were achieved with λ set to 0.002. However, a gradual decrease in detection accuracy was observed when λ exceeded 0.004. Therefore, a threshold value of 0.002 is selected for this paper.

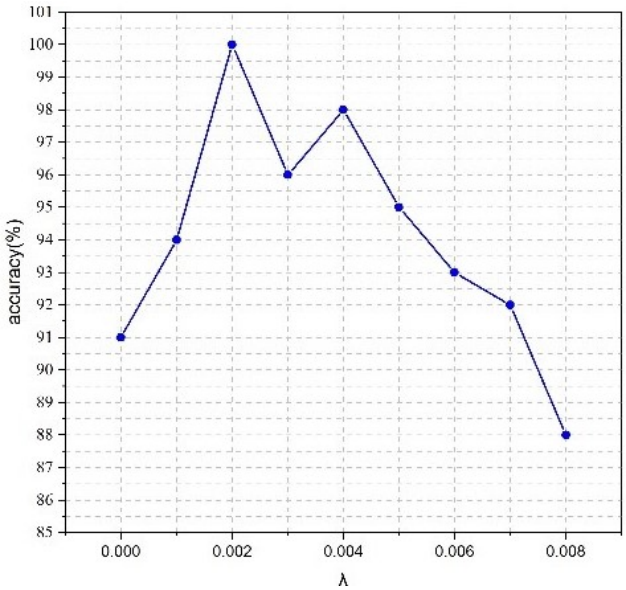


Figure 6. Accuracy curve with different λ value.

5. Conclusions

In this study, a novel lightweight method based on 1D CNN and Multi-Head Self-Attention Mechanism is used to detect abnormal sound on highway, which is named HADNet. First, a 1D CNN is utilized to extract local features from the raw audio signal, which avoids the loss of information and computation cost caused by time-frequency conversion. Secondly, we employ a block model based on the Multi-Head Self-Attention

Mechanism to extract global features from the audio signal. This approach not only effectively mitigates the issue of vanishing gradients, but also enhances both detection accuracy and effectiveness. Finally, the joint loss function is employed to detect the abnormal audio. This choice is crucial for addressing issues related to unbalanced training data and overlaps between classes, which ultimately enhances the overall performance of the HADNet model. The proposed method was thoroughly evaluated on the MIVIA Road Events and UrbanSound8K datasets. The experimental results show that HADNet achieves a good balance between recognition rate and detection efficiency for abnormal sounds. Specifically, the method achieves a high recognition rate, low rate of miss and false alarm, and exhibits high detection efficiency on the embedded platform. Consequently, it effectively fulfills the practical requirements for detecting abnormal sounds on highway. Overall, HADNet offers a promising solution for enhancing road safety measures and facilitating timely responses to traffic incidents through effective abnormal sound detection.

In future work, we aim to enhance our dataset by incorporating additional samples of abnormal sounds. This augmentation will provide a more comprehensive and diverse dataset, contributing to our detection model's improved robustness and effectiveness.

Author Contributions: Conceptualization, J.T. and Q.C.; methodology, C.L. and J.T.; data curation, C.L. and Q.C.; investigation, Q.C. and Q.W.; software, C.L. and Q.L.; formal analysis, J.T., Q.C. and Q.W.; validation, Q.L.; supervision, J.T.; project administration, Q.W.; funding acquisition, Q.C. and K.Z.; writing—original draft preparation, C.L.; writing—review and editing, J.T., K.Z. and A.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 62303325, Shanghai Science and Technology Innovation Action Plan Morning Star Project (Sail Special) (No. 23YF1429700), Yunnan Province Technical Innovation Talent Development Projects (No. 202405AD350058), and Major Science and Technology Projects of Yunnan Province (No. 202202AD080010).

Data Availability Statement: Dataset available on request from the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, Y.; Liu, W.; Huang, Q. Traffic anomaly detection based on image descriptor in videos. *Multimed. Tools Appl.* **2016**, *75*, 2487–2505. [CrossRef]
2. Khan, S.W.; Hafeez, Q.; Khalid, M.I.; Alroobaea, R.; Hussain, S.; Iqbal, J.; Almotiri, J.; Ullah, S.S. Anomaly detection in traffic surveillance videos using deep learning. *Sensors* **2022**, *22*, 6563. [CrossRef] [PubMed]
3. Pamula, T. Road traffic conditions classification based on multilevel filtering of image content using convolutional neural networks. *IEEE Intell. Transp. Syst. Mag.* **2018**, *10*, 11–21. [CrossRef]
4. Maha Vishnu, V.; Rajalakshmi, M.; Nedunchezhian, R. Intelligent traffic video surveillance and accident detection system with dynamic traffic signal control. *Clust. Comput.* **2018**, *21*, 135–147. [CrossRef]
5. Chandrakala, S.; Jayalakshmi, S. Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies. *ACM Comput. Surv. (CSUR)* **2019**, *52*, 1–34. [CrossRef]
6. Almaadeed, N.; Asim, M.; Al-Maadeed, S.; Bouridane, A.; Beghdadi, A. Automatic detection and classification of audio events for road surveillance applications. *Sensors* **2018**, *18*, 1858. [CrossRef]
7. Rehman, A.U.; Ullah, H.S.; Farooq, H.; Khan, M.S.; Mahmood, T.; Khan, H.O.A. Multi-modal anomaly detection by using audio and visual cues. *IEEE Access* **2021**, *9*, 30587–30603. [CrossRef]
8. Latif, S.; Cuayáhuitl, H.; Pervez, F.; Shamshad, F.; Ali, H.S.; Cambria, E. A survey on deep reinforcement learning for audio-based applications. *Artif. Intell. Rev.* **2023**, *56*, 2193–2240. [CrossRef]
9. Stowell, D.; Giannoulis, D.; Benetos, E.; Lagrange, M.; Plumbley, M.D. Detection and classification of acoustic scenes and events. *IEEE Trans. Multimed.* **2015**, *17*, 1733–1746. [CrossRef]
10. Mnasri, Z.; Rovetta, S.; Masulli, F. Anomalous sound event detection: A survey of machine learning based methods and applications. *Multimed. Tools Appl.* **2022**, *81*, 5537–5586. [CrossRef]
11. Purwins, H.; Li, B.; Virtanen, T.; Schlüter, J.; Chang, S.Y.; Sainath, T. Deep learning for audio signal processing. *IEEE J. Sel. Top. Signal Process.* **2019**, *13*, 206–219. [CrossRef]
12. Wei, P.; He, F.; Li, L.; Li, J. Research on sound classification based on SVM. *Neural Comput. Appl.* **2020**, *32*, 1593–1607. [CrossRef]
13. Momynkulov, Z.; Dosbayev, Z.; Suliman, A.; Abduraimova, B.; Smailov, N.; Zhekambayeva, M.; Zhamangarin, D. Fast Detection and Classification of Dangerous Urban Sounds Using Deep Learning. *CMC-Comput. Mater. Contin.* **2023**, *75*, 2191–2208. [CrossRef]

14. Kotus, J.; Lopatka, K.; Czyzewski, A. Detection and localization of selected acoustic events in acoustic field for smart surveillance applications. *Multimed. Tools Appl.* **2014**, *68*, 5–21. [CrossRef]
15. Crocco, M.; Cristani, M.; Trucco, A.; Murino, V. Audio surveillance: A systematic review. *ACM Comput. Surv. (CSUR)* **2016**, *48*, 1–46. [CrossRef]
16. Mateen, A.; Hanif, M.Z.; Khatri, N.; Lee, S.; Nam, S.Y. Smart roads for autonomous accident detection and warnings. *Sensors* **2022**, *22*, 2077. [CrossRef] [PubMed]
17. Mittal, U.; Chawla, P. Acoustic based emergency vehicle detection using ensemble of deep learning models. *Procedia Comput. Sci.* **2023**, *218*, 227–234. [CrossRef]
18. Rebai, K.; Achour, N.; Azouaoui, O. Road intersection detection and classification using hierarchical SVM classifier. *Adv. Robot.* **2014**, *28*, 929–941. [CrossRef]
19. Vrysis, L.; Tsipas, N.; Thoidis, I.; Dimoulas, C. 1D/2D deep CNNs vs. temporal feature integration for general audio classification. *J. Audio Eng. Soc.* **2020**, *68*, 66–77. [CrossRef]
20. Li, Y.; Li, X.; Zhang, Y.; Liu, M.; Wang, W. Anomalous sound detection using deep audio representation and a BLSTM network for audio surveillance of roads. *IEEE Access* **2018**, *6*, 58043–58055. [CrossRef]
21. Li, F.; Liu, M.; Zhao, Y.; Kong, L.; Dong, L.; Liu, X.; Hui, M. Feature extraction and classification of heart sound using 1D convolutional neural networks. *EURASIP J. Adv. Signal Process.* **2019**, *2019*, 59. [CrossRef]
22. Wang, M.; Yao, Y.; Qiu, H.; Song, X. Adaptive Memory-Controlled Self-Attention for Polyphonic Sound Event Detection. *Symmetry* **2022**, *14*, 366. [CrossRef]
23. Foggia, P.; Petkov, N.; Saggese, A.; Strisciuglio, N.; Vento, M. Audio surveillance of roads: A system for detecting anomalous sounds. *IEEE Trans. Intell. Transp. Syst.* **2015**, *17*, 279–288. [CrossRef]
24. Salamon, J.; Jacoby, C.; Bello, J.P. A dataset and taxonomy for urban sound research. In Proceedings of the 22nd ACM international conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 1041–1044.
25. Lopatka, K.; Kotus, J.; Czyzewski, A. Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations. *Multimed. Tools Appl.* **2016**, *75*, 10407–10439. [CrossRef]
26. Chandrakala, S.; Jayalakshmi, S. Generative model driven representation learning in a hybrid framework for environmental audio scene and sound event recognition. *IEEE Trans. Multimed.* **2019**, *22*, 3–14. [CrossRef]
27. Foggia, P.; Petkov, N.; Saggese, A.; Strisciuglio, N.; Vento, M. Reliable detection of audio events in highly noisy environments. *Pattern Recognit. Lett.* **2015**, *65*, 22–28. [CrossRef]
28. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
29. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
30. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
31. Mehta, S.; Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arXiv:2110.02178.
32. Greco, A.; Petkov, N.; Saggese, A.; Vento, M. AReN: A deep learning approach for sound event recognition using a brain inspired representation. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 3610–3624. [CrossRef]
33. Greco, A.; Roberto, A.; Saggese, A.; Vento, M. DENet: A deep architecture for audio surveillance applications. *Neural Comput. Appl.* **2021**, *33*, 11273–11284. [CrossRef]
34. Cao, J.; Cao, M.; Wang, J.; Yin, C.; Wang, D.; Vidal, P.P. Urban noise recognition with convolutional neural network. *Multimed. Tools Appl.* **2019**, *78*, 29021–29041. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.