machines

MDPI

Article

Sound-Based Intelligent Detection of FOD in the Final Assembly of Rocket Tanks

Tantao Lin [1], Yongsheng Zhu [1,*], Zhijun Ren [1], Kai Huang [1], Xinzhuo Zhang [1], Ke Yan [1] and Shunzhou Huang [2]

[1] Key Laboratory of Education Ministry for Modern Design & Rotor-Bearing System, Xi'an Jiaotong University, Xi'an 710049, China
[2] Shanghai Aerospace Equipments Manufacturer Co., Ltd., Shanghai 200245, China
* Correspondence: yszhu@mail.xjtu.edu.cn

Abstract: The traditional method of relying on human hearing to detect foreign object debris (FOD) events during rocket tank assembly processes has the limitation of strong reliance on humans and difficulty in establishing objective detection records. This can lead to undetected FOD entering the engine with the fuel and causing major launch accidents. In this study, we developed an automatic, intelligent FOD detection system for rocket tanks based on sound signals to overcome the drawbacks of manual detection, enabling us to take action to prevent accidents in advance. First, we used log-Mel transformation to reduce the high sampling rate of the sound signal. Furthermore, we proposed a multiscale convolution and temporal convolutional network (MS-CTCN) to overcome the challenges of multi-scale temporal feature extraction to detect suspicious FOD events. Finally, we used the proposed post-processing strategies of label smoothing and threshold discrimination to refine the results of FOD event detection and ultimately determine the presence of FOD. The proposed method was validated through FOD experiments. The results showed that the method had an accuracy rate of 99.16% in detecting FOD and had a better potential to prevent accidents compared to the baseline method.

Keywords: rocket tank; foreign object debris (FOD); sound detection; temporal convolution

Citation: Lin, T.; Zhu, Y.; Ren, Z.; Huang, K.; Zhang, X.; Yan, K.; Huang, S. Sound-Based Intelligent Detection of FOD in the Final Assembly of Rocket Tanks. *Machines* **2023**, *11*, 187. https://doi.org/10.3390/machines11020187

Academic Editor: Davide Astolfi

Received: 13 December 2022
Revised: 13 January 2023
Accepted: 17 January 2023
Published: 31 January 2023

1. Introduction

The tank is a crucial component of a rocket, serving as a storage unit for cryogenic propellants and providing structural support and protection for the rocket [1]. It is typically composed of multiple tanks that are joined together through a docking process, which includes drilling, riveting, and bolting. However, this process may inadvertently result in the presence of Foreign Object Debris (FOD) such as small bolts, nuts, screws, titanium wire, and other metal remnants inside the tank. FOD is a major contributor to launch accidents of space products, as they can enter the engine along with the propellant during launch if not detected and removed in a timely manner, leading to serious accidents. Thus, the development of FOD detection technology for rocket tanks is vital for ensuring the reliability of aerospace engineering.

The commonly used Particle Impact Noise Detection (PIND) method [2–8] in the field of FOD in the aerospace industry requires the manufacture of a large excitation platform to be effective in detecting FOD in large rocket tanks, which is not practical. The methods based on machine vision [9,10], on the other hand, require cameras to enter the complex internal structures of the tanks and are susceptible to being affected by factors such as occlusion and lighting, which can also cause secondary damage to the tanks. Therefore, the most common method used for detecting FOD in rocket tanks is the rotation listening method, which involves rotating the tank slowly while it is being inspected and listening for the sound of metal FOD sliding against or colliding with the tank's

walls. The presence of FOD is determined by human experience, but this method has some drawbacks, including low detection efficiency, high workload, and high dependence on human judgment. Additionally, it is difficult to form data records of detection and quantitative evaluation results. Because of these drawbacks, there is a need for a more efficient, accurate, and automatic method for detecting FOD in rocket tanks.

Nowadays, acoustic-based condition monitoring methods using microphones have gained popularity in special applications due to their non-destructive and non-contact advantages [11], such as in trackside acoustic diagnostic systems [12], arc magnet internal defect detection [13], and milling process monitoring [14]. These systems offer a new approach for detecting FOD in rocket tanks by using microphone sensors to acquire acoustic signals and applying suitable signal processing and FOD identification algorithms for intelligent, quantitative assessment of FOD status. However, in the field of acoustic-based diagnosis, it is common to intercept a small segment of the signal for processing, which is suitable for high-speed rotating machinery such as gears [15,16], bearings [17,18] and motors [19,20]. This approach may not be effective for FOD detection in tanks, as FOD can cause random scratching or collision events with the slowly rotating (1-2 r/min) inner walls of the tank, and the response time for FOD events can range from a few tens of milliseconds to a few seconds. Therefore, sound-based FOD detection must be able to detect suspicious FOD sound events of varying lengths within the entire input signal and determine the presence of an FOD event through post-processing criteria.

For the task of long-signal sound event detection, there are two main types of methods: traditional methods and intelligent methods. Traditional methods, such as those based on Gaussian Mixture Models–Hidden Markov Models (GMM-HMM) [21] or Non-negative Matrix Factorization (NMF) [22], rely on manual feature design and are dependent on user expertise. In contrast, data-driven deep learning models, such as those based on Convolutional Recurrent Neural Networks (CRNN) [23–25], are increasingly popular due to their ability to perform automatic feature extraction. These models utilize recurrent neural network structures, such as Long Short-Term Memory (LSTM) [26] and Gate Recurrent Units (GRUs) [27], to model the time sequence characteristics of sound events. However, some studies have demonstrated that recurrent neural network structures have limitations in modeling long-term dependencies [28]. In contrast, Temporal Convolutional Networks (TCNs) are able to extract features from both short-term and long-term time sequences by expanding the receptive field through multiple layers of dilated convolutions with different dilation rates and have been shown to be effective in tasks such as sound event detection [29] and localization [30].

Unfortunately, TCN-based methods still have the following limitations for FOD detection tasks. First, locally, the frictional sound generated by the FOD in the tank consists of many short-weak shocks with random intervals. The duration of these shocks and the time interval between shocks can be long or short and need to be analyzed from different scales. However, traditional TCNs use only one size of convolution kernel in dilated convolution, and a single size of convolution kernel does not capture the local multi-scale features [15]. Secondly, the TCN-based event detection method only makes a simple threshold judgment on the prediction result of each frame, and when the prediction probability is greater than a set value, a specific event is considered to have occurred in that frame. This will lead to unstable detection results for the FOD events consisting of intermittent short-weak shocks.

To solve the above problems, the MS-CTCN method is proposed in this paper for sound-based FOD detection. The main contributions of this paper are as follows:

(1) A new method for detecting FOD in rocket tanks using sound signals and deep neural networks is proposed, addressing the deficiencies of traditional methods.

(2) The Multi-Kernel Size (MKS) convolution is introduced in TCN-based temporal feature extraction for local feature fusion and solving FOD multi-scale short weak shock extraction.

(3) A post-processing strategy with label smoothing and decision-making based on FOD response frames is proposed for stable detection.

(4) A simulation tank for FOD detection experiments is constructed to test the proposed method's effectiveness and performance using various FOD sound signals.

The remainder of this paper is organized as follows: in Section 2, we review relevant previous research. The fundamental theories utilized in this study are outlined in Section 3. The proposed method is described in detail in Section 4. The experimental setup is presented in Section 5. The results and discussion are provided in Section 6. Finally, we provide our conclusions in Section 7.

2. Literature Review

2.1. FOD Detection in Rocket Tanks

PIND [2–8] is considered the most well-established method in the field of FOD. This method utilizes a specialized excitation platform to generate mechanical impacts that cause FOD to collide with the inner walls of the object being inspected. The collision signals are then captured using acceleration sensors or acoustic emission sensors and analyzed using signal processing techniques to identify the presence of FOD [7,8]. Additionally, researchers have developed FOD detection methods using machine vision technology [9,10], which use specialized cameras to scan the interior of the object and compare images to determine the presence of FOD. However, these methods are not suitable for detecting FOD in rocket tanks due to their complex structures. The PIND method is limited in its application to larger objects, and the excitation process can potentially cause damage to the tanks. Machine-vision-based methods, on the other hand, have the risk of introducing additional FOD and are affected by factors such as lighting and obstruction. As a result, the current method for detecting FOD in rocket tanks during final assembly involves slowly rotating the tank under the motor's drive and listening for the acoustic signals produced by the FOD colliding or slipping against the walls. However, this method is labor-intensive, relies heavily on human experience, and produces limited data records and quantitative results.

As hardware and software technologies have advanced, microphone-based acoustic state detection technology for machinery has been widely studied [12–14,31]. It has the advantage of non-destructive and non-contact [11]. Therefore, conducting research on FOD detection in rocket tanks based on sound signals has promising potentials. However, these methods for rotating machinery deal with periodic signals with fast rotational speeds and are not applicable to rocket storage tank sound signals with slow rotational speeds and random non-stationary FOD events.

2.2. Sound Event Detection

Sound event detection refers to the process of detecting and identifying specific sound events within a given recording or live audio stream. Currently, deep learning-based [32,33] sound event detection techniques have the advantage of automatically extracting features and stronger fitting capabilities compared to traditional machine learning methods based on GMM-HMM [21] or NMF [22], which makes them widely used. Convolutional neural networks (CNNs) are particularly effective in extracting local features due to their local connections and weight sharing, which is important for many sound event detection tasks [34,35]. However, CNNs are not suitable for capturing long-term temporal dependencies in signals, which leads to poor performance when detecting events with long time spans. To address this limitation, some methods combine CNN with other types of models, such as RNN [26,27] or Transformers [36], to capture both local and long-term temporal features. CRNNs [23–25,37] use CNNs as the feature extractors and feed the extracted features into RNNs, which are able to model temporal dependencies in the input signal. The CNN–Transformer uses a self-attention mechanism to model the temporal relationship among features[38]. The performance of the CNN–Transformer is comparable with CRNN, and it has the advantage of being more computationally efficient due to its parallel computation nature [39]. However, Transformer models have been shown to have a weaker inductive bias when compared to some other models such as CNNs [40]. This means that they may perform poorly when there is are a limited number of data. Another variant of

CNN, known as TCN [41], is another option for capturing short- and long-term temporal dependencies in the input signals while maintaining a relatively lower computational complexity. TCN uses multiple layers of convolutional layers, each with a different dilation factor, to expand the receptive field of the network. By using dilated convolutions, TCN can capture long-term dependencies in the input signal without the need for recurrent connections or self-attention mechanisms, which makes it computationally more efficient than RNNs or Transformers while still able to effectively extract temporal features in the input signal. It is considered to be a simple yet powerful alternative to complex architectures, and its performance on sound event detection [29,42] and localization [30] has been proven.

However, the utilization of TCN for FOD detection is associated with certain limitations. The FOD response is composed of a series of brief and weak shocks with unpredictable intervals and duration events, which can vary in length. As a result, utilizing a single kernel-sized TCN may not effectively capture the multi-scale properties [15] in the temporal sequence. Furthermore, the current event detection methods rely solely on simple thresholding to make final decisions, which may not be suitable for detecting weaker FOD events.

3. Theoretical Background

3.1. Logarithmic Mel Transform

The sound signal of the rotating tank is characterized by (a) time-varying non-stationary and (b) high sampling rate. To address these characteristics, this study incorporates the logarithmic Mel transform[30,43], which is a method that is based on the Short-Time Fourier Transform (STFT) [44] and can capture the time-frequency information of time-varying signals. The Mel transform also reduces the dimensionality of the STFT features of high sampling rate signals while preserving useful information [45]. Additionally, the Mel transform simulates the varying sensitivity of the human ear [45] to different frequencies of sound, thereby introducing further a priori knowledge. The process of logarithmic Mel time-frequency transformation is as follows.

STFT operation. A window function of static length is used to intercept a very short part (frame) of the total time-varying signal, and the discrete Fourier transform is used to obtain the local spectrum of each frame, as expressed in Equation (1).

$$\boldsymbol{S}(t,k) = \sum_{n=1}^{N_\mathrm{f}} \boldsymbol{x}_t[n] \cdot \boldsymbol{w}[n] \cdot e^{-j\frac{2\pi kn}{N_\mathrm{f}}} \tag{1}$$

where $\boldsymbol{S}(t,k) \in \mathbb{R}^{T\times N_\mathrm{f}}$ represents the k-th discrete frequency component of the short-time spectrum at time t; $t = 1, 2, \cdots, T$ and $k = 1, 2, \cdots, N_\mathrm{f}$ denote the frame number and discrete frequency number, respectively; and $\boldsymbol{x}_t[n]$ and $\boldsymbol{w}[n]$ refer to the t-th framed signal and window function, respectively.

Mel frequency filter banks construction. The main parameters for constructing Mel filter banks include (a) number of Mel filters, F, (b) minimum frequency in Hz, $f_{\mathrm{Hz}_{min}}$, and (c) maximum frequency in Hz, $f_{\mathrm{Hz}_{max}}$. For the most general case, $f_{\mathrm{Hz}_{min}}$ is equal to zero and $f_{\mathrm{Hz}_{max}}$ is equal to half the sampling frequency. Then, Mel-scale minimum frequency and maximum frequency $f_{\mathrm{Mel}_{max}}$ and $f_{\mathrm{Mel}_{min}}$ can be computed using Equation (2). After that, Mel frequency filters can be constructed as expressed in Equations (3)–(5).

$$f_\mathrm{Mel} = 2595 \cdot \log_{10}(1 + f_\mathrm{Hz}/700) \tag{2}$$

$$\boldsymbol{H}(k,m) = \begin{cases} 0 & , f_\mathrm{Hz}(k) < f_{\mathrm{Hz_c}}(m-1) \\ \frac{f_\mathrm{Hz}(k) - f_{\mathrm{Hz_c}}(m-1)}{f_{\mathrm{Hz_c}}(m) - f_{\mathrm{Hz_c}}(m-1)} & , f_{\mathrm{Hz_c}}(m-1) \le f_\mathrm{Hz}(k) \le f_{\mathrm{Hz_c}}(m) \\ \frac{f_\mathrm{Hz}(k) - f_{\mathrm{Hz_c}}(m+1)}{f_{\mathrm{Hz_c}}(m) - f_{\mathrm{Hz_c}}(m+1)} & , f_{\mathrm{Hz_c}}(m) \le f_\mathrm{Hz}(k) \le f_{\mathrm{Hz_c}}(m+1) \\ 0 & , f_\mathrm{Hz}(k) \ge f_{\mathrm{Hz_c}}(m+1) \end{cases} \tag{3}$$

$$f_{\mathrm{Hz_c}}(m) = 700(10^{f_{\mathrm{Mel_c}}(m)/2595} - 1) \tag{4}$$

$$f_{\mathrm{Mel_c}}(m) = m \cdot (f_{\mathrm{Mel}_{max}} - f_{\mathrm{Mel}_{min}})/F \tag{5}$$

where $\boldsymbol{H}(k,m) \in \mathbb{R}^{N_f \times F}$ denotes the gain at the k-th Hz-scale frequency $f_{\mathrm{Hz}}(k)$ in the m-th Mel filter; f_{Mel_c} and f_{Hz_c} denotes the center frequency of the filter in the Hz-scale and Mel-scale, respectively; and $m = 1, 2, \cdots, F$ denotes the Mel filter number.

Mel filter bank features extraction. The constructed Mel filter banks $\boldsymbol{H}$ are used to extract features from the original short-time spectrum $\boldsymbol{S}$. Eventually, the two-dimensional MFB features $\boldsymbol{X}(t,m)$ are obtained according to Equation (6), which includes the subsequent absolute and logarithmic operations.

$$\boldsymbol{X}(t,m) = 20 \times \log_{10}(|\boldsymbol{S}(t,k)| \cdot \boldsymbol{H}(k,m)) \in \mathbb{R}^{T \times F} \tag{6}$$

3.2. Convolutional Neural Network (CNN)

CNNs are particularly adept at extracting features from two-dimensional (2D) data, such as images. Compared to traditional Artificial Neural Networks (ANNs), CNNs have a smaller number of trainable parameters and have been widely and successfully applied in various fields [46,47], including the field of sound-based intelligent equipment maintenance [48,49], where it has demonstrated impressive results in various studies. The fundamental operations in a convolutional layer include convolution, activation, batch normalization, and max pooling.

The convolution operation with the most used Rectified Linear Unit (ReLU) activation function is described in Equations (7) and (8).

$$\boldsymbol{y}_{\mathrm{a}} = \mathrm{Re\,LU}(\boldsymbol{W}_{\mathrm{cn}} \otimes \boldsymbol{x}_{\mathrm{cn}} + \boldsymbol{b}_{\mathrm{cn}}) \tag{7}$$

$$\mathrm{ReLU}(\boldsymbol{x}) = \max\{\boldsymbol{x}, 0\} \tag{8}$$

where $\boldsymbol{W}_{\mathrm{cn}}$, $\boldsymbol{b}_{\mathrm{cn}}$, and $\otimes$ denote weight, bias, and convolution operation, respectively; $\boldsymbol{x}_{\mathrm{cn}} \in \mathbb{R}^{W \times H}$ refers to input feature map; and $\boldsymbol{y}_{\mathrm{a}}$ denotes feature map after convolution and activation.

Batch Normalizing (BN) transform is described in Equations (9) and (10). The purpose of using BN is to reduce internal covariate shift to better train the network [50].

$$\widehat{\boldsymbol{y}_{\mathrm{a}}} = \frac{\boldsymbol{y}_{\mathrm{a}} - \mu}{\sqrt{\sigma^2 + \varepsilon}} \tag{9}$$

$$\boldsymbol{y}_{\mathrm{bn}} = \gamma \widehat{\boldsymbol{y}_{\mathrm{a}}} + \beta \tag{10}$$

where μ and σ are the expectation and the variance, $\boldsymbol{y}_{\mathrm{bn}}$ represents the output features over a mini-batch, γ and β are two parameters to be learned, and ϵ is a constant close to zero.

Maximum pooling is a sample-based discretization process. The objective is to down-sample an input representation (image, hidden-layer output matrix, etc.), reducing its dimensionality and enabling features that are more robust after convolution [51]. The mathematical description is given as follows:

$$\boldsymbol{y}_{\mathrm{mp}}(i) = \max\{\boldsymbol{y}_{\mathrm{bn}}(i : i + r - 1)\} \tag{11}$$

where $\boldsymbol{y}_{\mathrm{mp}}(i)$ denotes the maximum value in the i-th corresponding pooling region and r is the pool size.

3.3. Temporal Convolutional Network (TCN)

Unlike 2D-CNN, the processing object of TCN is 1D time sequences. The basic TCN consists of 1D dilated convolutions with different dilation rates. As the dilation rate increases, the receptive field of the convolution kernel increases accordingly [25]. By concatenating multiple layers of dilated convolutions with gradually increasing dilated rates, the network can learn the time sequence correlation of sound signals that expand from local to global. As illustrated in Figure 1, a three-layer structure with a kernel size of 3 and a dilated convolution with dilation rates of 1, 2, and 4 is given, and the receptive field changes from an initial 3 to 9. The output in the i-th layer at moment t with an odd kernel size is described in Equation (12).

$$x_{i+1}(t) = \sum_{u=-\xi}^{\xi} w_i(u) \cdot x_i(t + u \cdot d_i) + b_i(u) \tag{12}$$

where $\xi = \text{int}(k_i/2)$, $w_i(u)$ and $b_i(u)$ are the u-th element's weight and bias of convolution kernel in the i-th layer.

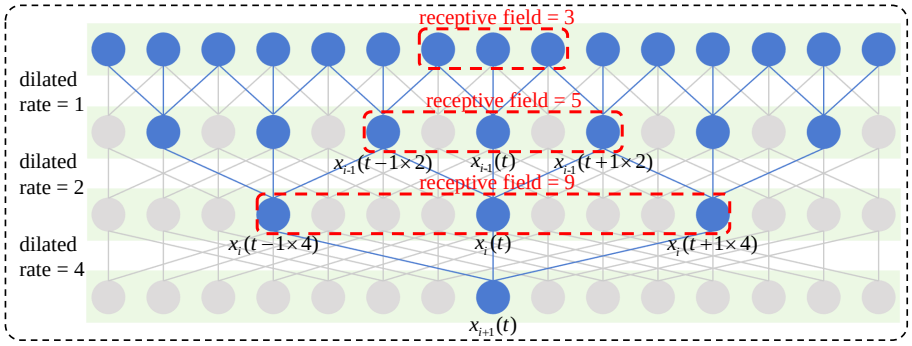


Figure 1. 1D dilated convolution with convolution kernel size 3.

For the i-th layer of dilated convolution, the output length T_{i+1} is shown in Equation (13).

$$T_{i+1} = \frac{T_i + 2p_i - d_i(k_i - 1) - 1}{s_i} + 1 \tag{13}$$

where p_i, d_i, k_i, and s_i denote padding size, dilated rate, kernel size, and stride of dilated convolution. To make the input and output sequence equal in length, specifically, $T_{i+1} = T_i = T$, the boundary padding $p_i = 1/2[s_i(T-1) - T + d_i(k_i - 1) + 1]$.

4. Proposed Method

During FOD detection process of rocket tanks, the tank is rotated by a motor, causing any FOD inside the tank to be lifted along the inner wall. When it reaches a certain height, the FOD falls off the wall due to gravity and generates an abnormal sound from impacts and friction. These mechanical sound waves are transmitted to the outside of the tank and captured by microphones placed externally. The subsequent detection system processes the sound signals acquired by the microphones and automatically determines the presence of FOD. By gradually moving the position of the microphone along the axial direction of the rocket tank, the FOD detection operation can be completed for the entire tank. The proposed method, the MS-CTCN-based FOD detection system, is the core of this process. As illustrated in Figure 2, it includes five sequential steps: (1) preprocessing, (2) CNN block, (3) MS-TCN block, (4) classification block, and (5) post-processing.

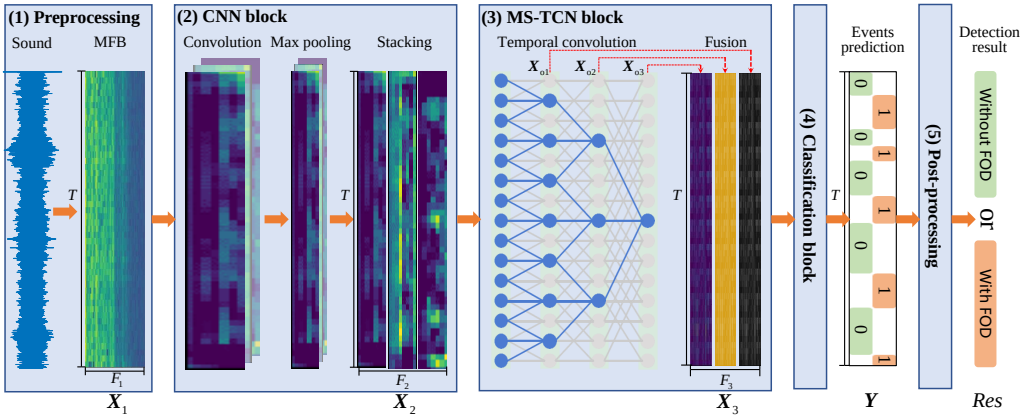


Figure 2. The MS-CTCN-based FOD detection system.

4.1. Preprocessing

The utilization of high sampling rate sound signals as input directly necessitates deep and extensive network models, which incurs significant computational costs. Therefore, a signal-processing method with a preliminary feature extraction function and dimensionality reduction capability is required. Additionally, the time-varying characteristics of the FOD signal must be considered. As outlined in Section 3.1, the problem of time-varying feature conversion of FOD signals is addressed by first applying the STFT to the raw time-domain high sampling rate signals, followed by extracting the MFB features, which effectively reduces the dimensionality while preserving relevant information.

For convenience, the continuous sound signal is divided into time frames with a duration of 2048 sampling points at a 48k sampling rate (42.7 ms), and the frame overlap rate is 50%. The log-Mel transform with the feature dimension $F = 128$ is applied to each frame to extract the MFB features $\boldsymbol{x}_t^{F_0}$, which leads to the input sequence $\boldsymbol{X}_1 = [\boldsymbol{x}_1^{F_1}, \boldsymbol{x}_2^{F_1}, \cdots, \boldsymbol{x}_T^{F_1}] \in \mathbb{R}^{T \times F_1}$ for the subsequent network, where T and F_1 present the length of the time sequence and the dimension of the MFB features, respectively.

4.2. CNN Block

The MFB feature is actually a 2D image with the location of each pixel point in the length and height directions in time and frequency, respectively. The lack of a feature extractor to further generalize the MFB into a feature form that the network can appreciate will make it difficult to develop prediction results. In image recognition, 2D CNNs, with less parameters and better feature extraction capability than fully connected networks, are frequently used to automatically comprehend 2D image information and further abstract the local translation-invariant [52] features in images. The 2D convolution operation described in Section 3.2 is also introduced before the TCN in the detection system. The purpose of the CNN block is to extract frequency invariant features from the time sequence $\boldsymbol{X}_1$ obtained in the previous step.

However, in pooling operation, unlike the traditional 2D max-pooling, which has the same pooling size in both dimensional directions, the max-pooling is only applied in the frequency direction [53], specifically, the height direction of the feature map. Its purpose is to ensure that the time resolution remains unchanged when using TCN to extract temporal features.

After a series of convolution blocks, the obtained multi-channel feature maps are stacked in the frequency direction. One can thereby obtain the single-channel feature map $\boldsymbol{X}_2 = [\boldsymbol{x}_1^{F_2}, \boldsymbol{x}_2^{F_2}, \ldots, \boldsymbol{x}_T^{F_2}] \in \mathbb{R}^{T \times F_2}$ with unchanged sequence length T, where F_2 represents the feature dimension after stacking. If the number of channels of the last layer of convolution operation is C_{cn}, the height of the output feature map is H_{cn}, and $F_2 = C_{\text{cn}} \cdot H_{\text{cn}}$.

4.3. MS-TCN Block

The signal of FOD slipping from the tank is characterized by short and weak shocks, and the interval between these short and weak shocks can be long or short. The duration of the FOD event consisting of a series of short and weak shocks can also be long or short. To prevent missed detection, the local response of the FOD signal should be extracted and the long temporal features of the FOD need to be concerned. Therefore, a TCN with dilated convolution is developed to increase the receptive field by increasing the cavity rate layer-by-layer for obtaining a wider range of long-time features. Meanwhile, MKS is joined to the dilated convolution operation to accommodate the short and weak shock features with different time intervals. Finally, the features extracted from different layers are fused to achieve the expected multi-scale feature fusion and enhance the feature expression level.

In the proposed multi-scale temporal convolution block as shown in Figure 3, three dilated convolution layers with different dilation rates are used to obtain global receptive fields at different scales. In each layer, an MKS dilation convolution operation is used for considering local features at different scales, and multi-scale feature fusion is performed on this basis.

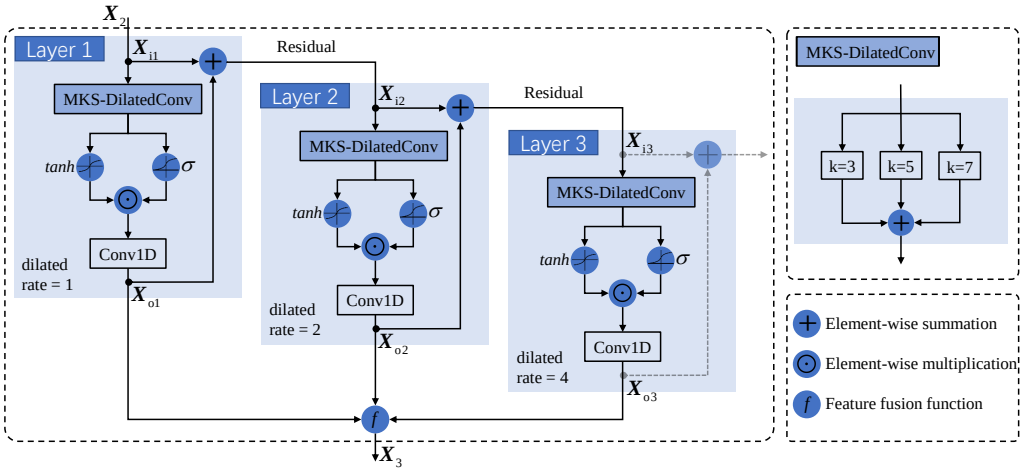


Figure 3. Multi-scale temporal convolution block.

The MKS dilated convolution operation in layer performs convolution and feature fusion as described in Equations (14) and (15).

$$x_{\text{out}}(t) = \sum_{j=1}^{3} \sum_{u=-\xi_j}^{\xi_j} w_{ij}(u) \cdot x_{\text{in}}(t + u \cdot d_i) + b_{ij}(u) \tag{14}$$

$$\xi_j = \text{int}(k_{ij}/2) \tag{15}$$

where k_{ij}, $w_{ij}(u)$, and $b_{ij}(u)$ denote the kernel size the u-th weight, and the bias of j-th kernel in MKS dilated convolution, respectively, and $x_{\text{in}(t)}$ and $x_{\text{out}(t)}$ denote input and output feature at the time, respectively. In the proposed method, we set $k_{i1} = 3$, $k_{i2} = 5$ and $k_{i3} = 7$ in the MKS-dilated convolution, as presented in Figure 3.

The activation function $f_a(x)$ after dilated convolution and feature fusion is combined with the tanh function and the sigmoid function, as expressed in Equations (16)–(18). This activation mode is better than the ReLU function in the time sequence processing of the sound signal [54].

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{16}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{17}$$

$$f_a(\boldsymbol{x}) = \tanh(\boldsymbol{x}) \odot \sigma(\boldsymbol{x}) \tag{18}$$

where $\odot$ denotes element-wise multiplication.

The data flow for the multi-scale temporal convolution block can be expressed as follows. The input sequence $\boldsymbol{X}_{\mathrm{i}1} = \boldsymbol{X}_2$ enters Layer 1, and MKS dilated convolution, activation, and a 1D convolution are performed. The output of Layer 1, $\boldsymbol{X}_{\mathrm{o}1}$, flows in two directions. One is be added to $\boldsymbol{X}_{\mathrm{i}1}$ to form the residual input of Layer 2 to enhance the model's trainable ability [55], and another one is fused with the output features $\boldsymbol{X}_{\mathrm{o}2}$ and $\boldsymbol{X}_{\mathrm{o}3}$ of Layer 2 and Layer 3. This multi-scale fusion is completed by feature stacking. Similarly, the same operation is performed on Layer 2 and Layer 3; the difference, however, is that the input to layers 2 and 3 are different from layer 1, the input to Layer 2 is $\boldsymbol{X}_{\mathrm{i}2} = \boldsymbol{X}_{\mathrm{i}1} + \boldsymbol{X}_{\mathrm{o}1}$, and the input to Layer 3 is $\boldsymbol{X}_{\mathrm{i}3} = \boldsymbol{X}_{\mathrm{i}2} + \boldsymbol{X}_{\mathrm{o}2}$. Then, the optimal output of the multi-scale temporal convolution block is obtained as described in Equation (19).

$$\boldsymbol{X}_3 = \begin{bmatrix} \boldsymbol{X}_{\mathrm{o}1} \\ \boldsymbol{X}_{\mathrm{o}2} \\ \boldsymbol{X}_{\mathrm{o}3} \end{bmatrix} = [\boldsymbol{x}_1^{F_3}, \boldsymbol{x}_2^{F_3}, \ldots, \boldsymbol{x}_T^{F_3}] \in \mathbb{R}^{T \times (3 \times \bar{F})} \tag{19}$$

where $\bar{F}$ denotes the size in the frequency direction of the output feature maps of each layer, and $F_3 = 3 \times \bar{F}$.

4.4. Classification Block

We expect fine detection of each frame of the signal to improve the interpretability and persuasiveness of the detection. Therefore, the state of each frame needs to be output. This can be regarded as an FOD state classification for each frame.

The classification block includes several fully connected layers and the sigmoid activation function of the last layer, which outputs the probability of each frame belonging to each sound event. If the number of sound event categories is C, then $\boldsymbol{P}_y \in \mathbb{R}^{T \times C}$, and $0 \leq \boldsymbol{P}_y(t, c) \leq 1$. In this article, the sound event category only includes two types, with FOD (abnormal) and without FOD (normal), i.e., $C = 2$. In other words, we do not distinguish between the types of FOD. Thus, the network's final output is shown in Equation (20).

$$\boldsymbol{Y} = [y_1, \cdots, y_T] = \arg\max(\boldsymbol{P}_y) \in \mathbb{R}^{T \times 1} \tag{20}$$

where $y_i \in \{0, 1\} (i = 1, \cdots, T)$ represents the detection result of each frame and 0 and 1 represent normal and abnormal, respectively.

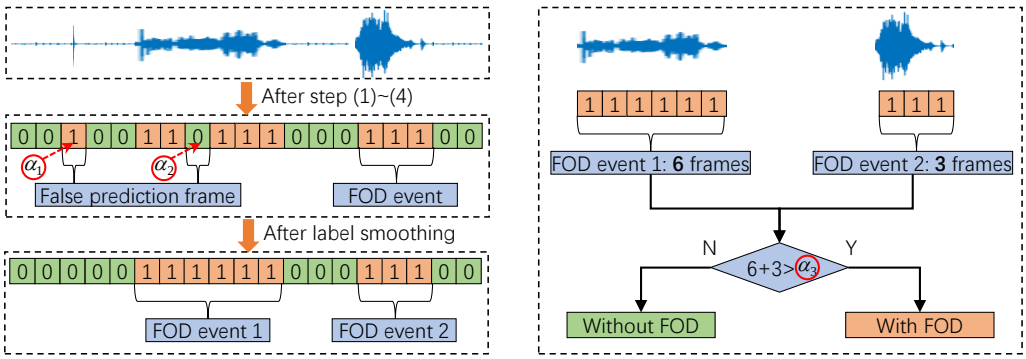
4.5. Post-Processing

The output of the above network represents only the state of each frame of the signal, while the final output required by the detection system is whether this segment of the signal contains FOD. In addition, the predicted state of each frame is not guaranteed to be 100% correct. Therefore, a post-processing strategy is needed to improve the robustness of the final decision. It includes (1) predicted label smoothing and (2) final discriminant criteria as described in Figure 4.

(1) Predicted label smoothing: From experimental results, it was found that the FOD sound events always last for some time (more than the length of each frame) in each long signal segment. Since the duration of each frame is chosen to be small when the sound signal is split, this makes the FOD sound events, once they occur, often last for several frames. In addition, the continuous response of each FOD event is locally composed of short and weak shock and transition intervals. Most of these transition segments are the same or close to the normal waveform, which may cause the system to output incorrect results. Therefore, it is necessary to adopt a smoothing strategy to correct these unstable prediction results. As shown in Figure 4a, the method proposed in this paper ignores frames predicted to be labeled "FOD" if the number of consecutive frames for which FOD is detected is less than the consecutive threshold α_1. Only when the number of consecutive abnormal frames

is greater than or equal to α_1, are the segments of these frames considered to be an FOD event. In addition, two FOD events with the frame interval less than the interval threshold α_2 are considered as the same event. By the above strategies, the elimination of instability of event detection results is achieved.

(2) Final discriminatory criteria: the continuous sound signal of the rocket tank is processed by pre-processing, network model prediction, and a frame smoothing strategy to obtain all suspected FOD events within the detection period of the rocket tank. As shown in Figure 4b, if the sum of the frames of all suspected FOD events exceeds the allowed threshold α_3, the FOD is believed to be presented in this part of the tank.



(**a**) (**b**)

Figure 4. Post-processing: (**a**) predicted label smoothing, (**b**) final discriminant criteria.

The specific algorithm pseudo-code for post-processing is shown in Algorithm 1.

Algorithm 1: Post-processing

Input : prediction result $\mathbf{Y} = [y_1, \cdots, y_T] \in \mathbb{R}^{T\times 1}$, continuous threshold value α_1, interval threshold α_2, allowable frame length threshold α_3

Output: final detection result $Res \in \{0, 1\}$, where 0 and 1 denote normal and abnormal respectively.

Initialize event counter $n = 0$
for $i \leftarrow 1$ **to** T **do**
  if *there are m consecutive y values of 1 starting from i* **and** $m > \alpha_1$ **then**
    $n = n + 1$
    n-th indicator of event start time $s_n = i$
    n-th indicator of evnet frame count $c_n = m$
  end
end
for $i \leftarrow 2$ **to** n **do**
  if $s_i - (s_{i-1} + c_{i-1}) < \alpha_2$ **and** $s_{i-1} \neq 0$ **then**
    $\boldsymbol{c}_i = (\boldsymbol{s}_i - \boldsymbol{s}_{i-1}) + \boldsymbol{c}_i$
    $\boldsymbol{s}_{i-1} = 0$
    $\boldsymbol{c}_{i-1} = 0$
  end
  if $\sum \boldsymbol{c} > \alpha_3$ **then**
    $Res = 1$
  else
    $Res = 0$
  end
end

5. Experimental Setup

5.1. Signal and Data Description

In the experimental part of this research, the rocket tank with a complex internal structure was simulated, and the rocket tank model and related detection devices as shown in Figure 5. were manufactured. The rocket tank model has a three-stage structure, including fuel tank Section 1, a connecting section, and fuel tank Section 2. Each section has a diameter of 800 mm and a length of 600 mm. A motor drives the two rollers on the left side to rotate the tank with a speed of 2 r/min. At the same time, 16 microphones arranged on the carbon fiber bracket collect sound signals in different positions. The abnormal state with FOD is simulated by adding metal FOD to the inner wall of the tank. The types of FOD configured in the experiment are shown in the right of Figure 5, including titanium alloy wires, nuts, bolts, and rivets.

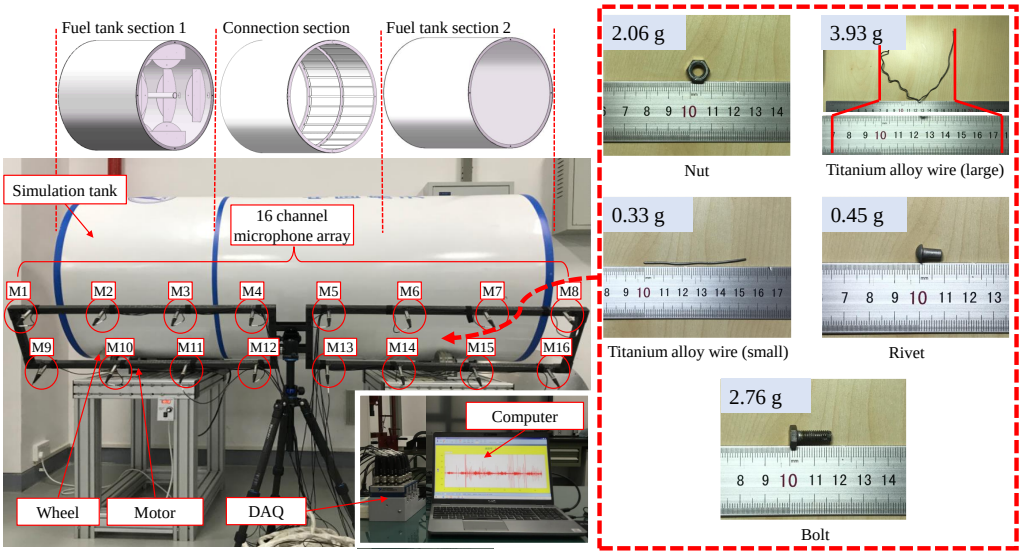


Figure 5. The test bench of sound-based detection for FOD.

The signals of different FOD types are presented in Figure 6. The amplitude of the original time domain signals is close and dominated by the background noise of the drive motor. Thus, it is difficult to distinguish them from each other. From the view of the short time spectrum, the high-frequency portion is dark, so the energy of the signal is mainly concentrated in the low frequency. This indicates that if STFT is used as the input, it will bring many meaningless high-frequency features and increase the computational effort in vain. In contrast, in the Mel spectrum, the central frequency distribution of the Mel filter is densely decayed from low to high frequencies. Therefore, the Mel spectrum of the FOD signal increases the resolution of the low-frequency band compared to the short-time spectrum, and the useful information is further amplified.

In the process of the production of the data set as shown in Figure 7, each 60 s signal is divided into many small segments of 5.46 s (256 frames), and the overlap between segments is 2.73 s (128 frames) for data augmentation. Then they are converted to MFB features with the length of 256 and feature dimension of 128, and labels for each frame are added manually. The final data set is shown in Table 1. Abnormal samples contain both normal frames and abnormal frames, while normal samples have only normal frames. Thus, only abnormal samples are used for model training, while both normal and abnormal samples are used for testing to evaluate the performance of the method.

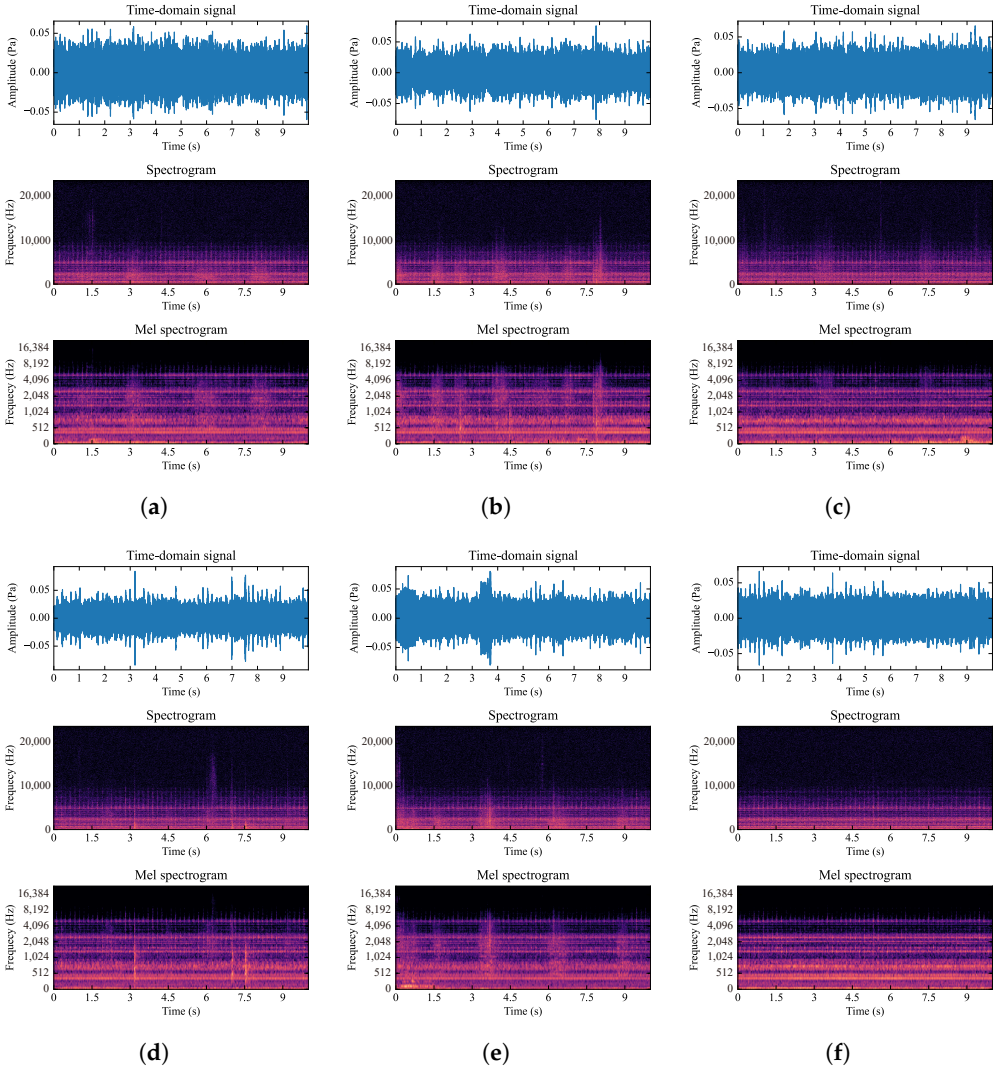


Figure 6. Signals of different FOD: (**a**) Nut, (**b**) Ti-wire (3.93 g), (**c**) Ti-wire (0.33 g), (**d**) Rivet, (**e**) Bolt, (**f**) Without FOD. In the spectrum and Mel spectrum, a brighter color means higher energy.

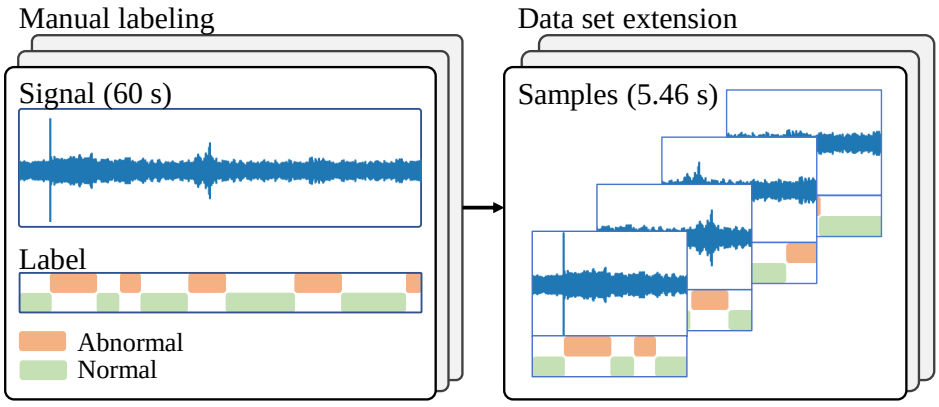


Figure 7. Manual labeling and data set expansion.

Table 1. Data set information.

Acoustic Condition of the Tank		Number of the Training Samples	Number of the Testing Samples
With FOD (abnormal)	Nut	960	240
	Ti-wire (3.39 g)	960	240
	Ti-wire (0.33 g)	960	240
	Rivet	960	240
	Bolt	960	240
Without FOD (normal)	-	-	1200
Total	-	4800	2400

5.2. Metrics

In this paper, the metrics [56] of precision P, recall R, and F-score F are used to evaluate the effectiveness of the proposed sound-based FOD detection method, and the definition of each indicator is shown in Equations (21)–(23). For the FOD detection task, precision can reflect the level of missed detection and recall can reflect the level of false alarm, while F-score is a comprehensive consideration of the above two factors.

$$P = \frac{TP}{TP + FP} \tag{21}$$

$$R = \frac{TP}{TP + FN} \tag{22}$$

$$F = \frac{2PR}{P + R} \tag{23}$$

where TP, FP, and FN denote true positive, false positive, and false negative, respectively. They are intermediate metrics from the confusion matrix.

In addition, the performance of the method is evaluated by calculating the overall accuracy of the samples with and without FOD, as described in Equation (24).

$$P_{\text{total}} = \frac{TP + TN}{N} \tag{24}$$

where TN denotes true negative and N denotes the total number of test samples.

5.3. Comparison Methods and Parameter Setting

The network model used in this paper consists of a three-layer 2D-CNN, a three-layer TCN, and a two-layer fully connected layer. The detailed network structure parameters are shown in Table 2. Since the process of FOD detection can be considered essentially as a frame-by-frame multi-label detection task, the network uses a binary cross-entropy (BCE) loss function. At the same time, a sigmoid function is used to output the probability of the accuracy of the signal detection results for each frame.

To verify the effectiveness of the MS-CTCN-based method for FOD detection, the method in this paper is compared with a CRNN that considers the sound signal's time sequence characteristics and a CNN that only contains convolution layers and fully connected layers. In both CNN and CRNN, the same three-layer 2D-CNN structure and the resulting post-processing method as the proposed method in this paper are adopted, while the recurrent structure in CRNN adopts the bidirectional gated recurrent unit (Bi-GRU).

The factors that affect the model performance during the training of all models are parameter initialization strategy, batch size, learning rate, and epoch number. For the sake of fairness, the same settings are adopted in all model training processes. All models are built based on Pytorch and use its default parameter initialization method. The batch size,

learning rate, and number of epochs are 64, 0.001, and 80, respectively. In addition, the most classical Adam optimizer is used to train the network parameters.

Table 2. The main parameters of the proposed method.

Operation	Output Size	Configuration
Preprocessing	256 × 128	-
Network input	1 × 256 × 128	-
CNN1(32, 5, 5)	32 × 256 × 128	Max-pooling (1, 4)
CNN2(32, 5, 5)	32 × 256 × 32	Max-pooling (1, 4)
CNN3(32, 3, 3)	32 × 256 × 4	Max-pooling (1, 2)
Reshape	256 × 128	-
TCN1	256 × 128	Dilation rate = 1
TCN2	256 × 128	Dilation rate = 2
TCN3	256 × 128	Dilation rate = 4
Connection	256 × 384	-
FC1	256 × 128	-
FC2	256 × 2	-
Sigmoid	256 × 2	-
Post-processing	1	$\alpha_1 = 2, \alpha_2 = 1, \alpha_3 = 10$

When post-processing the prediction results, three thresholds, α_1, α_2, and α_3, should be determined. α_1 and α_2 are determined to have the most appropriate values of 2 and 1 through a grid search; that is, the isolated abnormal frames are discarded and two abnormal events with an interval of one frame are combined into the same event. As shown in Figure 8, the total number of abnormal frames for most samples in the dataset is between 25 and 150, and there are almost no abnormal samples with fewer than 10 frames. The same grid search method is used and α_3 is set to 10.

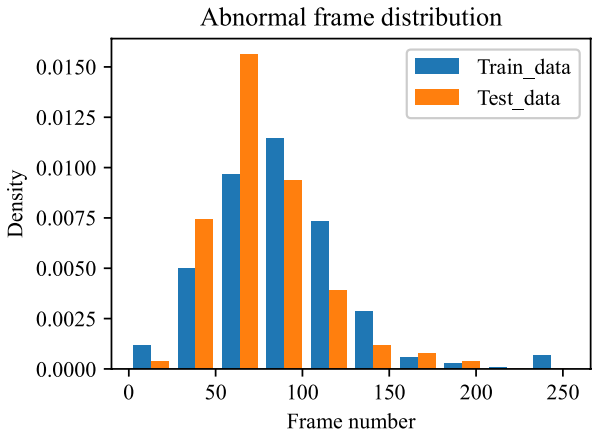


Figure 8. The statistical distribution of the number of frames of the samples with FOD.

6. Results and Discussion

6.1. Overall Performance Analysis of MS-CTCN

The final detection performance of the proposed method is presented in Table 3. As can be observed from the results, the CNN method, which does not take into account temporal characteristics, incorrectly classifies all samples as normal states, rendering it unsuitable for this system. In comparison, the MS-CTCN method, which incorporates temporal characteristics, demonstrates an improvement in precision, recall, F-score, and overall accuracy of 0.29%, 5.07%, 3.50%, and 2.70%, respectively, when compared to the CRNN method. These results indicate that the performance of the MS-CTCN method is superior.

Table 3. Performance of different models.

Model	*P*(%)	*R*(%)	*F*(%)	P_{total}(%)
CNN	100	0.08	0.17	50.04
CRNN	99.12	93.75	96.36	96.46
MS-CTCN	99.41	98.82	99.17	99.16

Due to the extremely poor performance of the CNN frame-by-frame method, only CRNN and MS-CTCN are further compared. The overall two-category confusion matrix is shown in Figure 9. For the case of misreporting the normal condition as abnormal, there are 10 cases of CRNN and 7 cases of MS-CTCN. The performance of the two methods is very close in terms of false alarm rate. However, in terms of the miss detection rate, MS-CTCN has only 13 cases, while CRNN has 75 cases. Thus, MS-CTCN performs better for FOD detection.

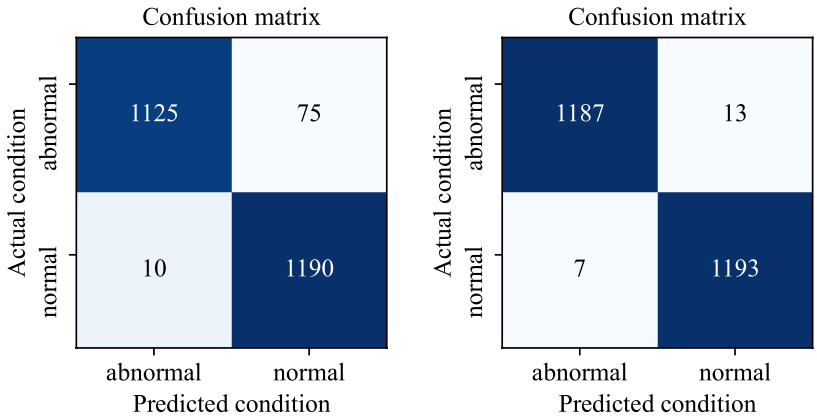


Figure 9. Confusion matrix: (**a**) Confusion matrix of CRNN, (**b**) Confusion matrix of MS-CTCN.

In addition, the detection performance of each kind of FOD is analyzed, as shown in Table 4. For nuts, large titanium alloy wires, and bolts with large masses, the detection accuracy can reach 100%. However, for rivets and small titanium alloy wires with a mass of less than 0.5 g, the detection accuracy of the simple CRNN method is not as good as that of the MS-CTCN method. Especially for small titanium alloy wires with a mass of only 0.33 g, the detection accuracy of the MS-CTCN method is 20.42% higher than that of the CRNN. Therefore, the method proposed in this paper has superior detection performance for smaller FOD.

Table 4. The accuracy of CRNN and MS-CTCN for different FOD.

Type of FOD	CRNN(%)	MS-CTCN(%)
Nut	100	100
Ti-wire (3.93)	100	100
Ti-wire (0.33)	77.50	97.92
Rivet	93.75	97.50
Bolt	100	100

6.2. Frame-Wise Event Detection Performance Analysis

To evaluate the performance of the proposed method for frame-wise event detection, a segment of bolt signal was further analyzed by different methods, as shown in Figure 10. The four red signal segments in Figure 10a are the four continuous bolt events marked manually. Due to the subjective nature of the manual marking, it is not possible to mark the exact start and end times of the events correctly, but an approximate range can be determined. Figure 10b shows the results using CNN. The CNN has poor feature extraction

ability for signals with faint FOD because the characteristics of the time sequences are not considered. Only a few frames predict the FOD response, but they do not agree with the true labels. In addition, FOD events occur less frequently than normal events. In most cases, most of the data frames input to the system are normal frames without FOD, so the training of CNN belongs to a typical CNN training problem under imbalanced sample conditions, where the CNN tends to predict all frames as FOD-free events due to the influence of majority class samples. The performance of CRNN is shown in Figure 10c, where it isolates only two FOD events, one that has its start time grossly misestimated. In contrast, the MS-CTCN method in Figure 10d successfully detects four FOD events, and the duration of each event includes the range of manually labeled events. In addition, the post-processed smoothed labels successfully eliminated the instability of the FOD event start and end positions. Therefore, the method in this paper is more suitable for frame-wise detection of FOD events.

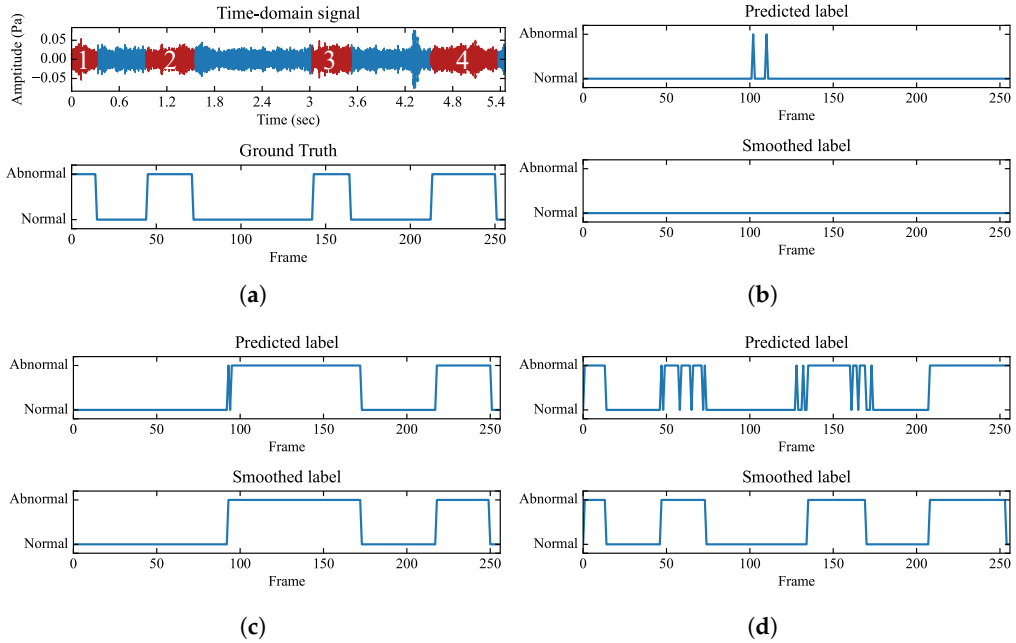


Figure 10. Frame-by-frame prediction and smoothing results for different methods: (**a**) FOD signal and true label, (**b**) CNN's prediction results and smoothed results, (**c**) CRNN's prediction results and smoothed results, (**d**) MS-CTCN's prediction results and smoothed results.

6.3. Performance of MS-TCN with Different Kernel Sizes

Experiments were conducted in MS-TCN blocks using convolution kernel sizes of 3, 5, and 7 to verify the excellent performance of convolution using MKS in this paper. The results are shown in Table 5. It is inferred that the possible reason is that the global time sequence features extracted at different scales are best suited for FOD sound event detection after the convolution kernel of size 3 expands the receptive field with dilation rates of 1, 2, and 4. In MKS, the addition of the temporal convolution with kernel sizes 5 and 7 corresponds to the enrichment of local multiscale time sequence features that are beneficial for FOD detection; i.e., local multiscale time sequence feature extraction is achieved.

Table 5. Model performance when using dilated convolution with different convolution kernel sizes.

Model	$P(\%)$	$R(\%)$	$F(\%)$
KS-3	99.82	99.67	98.22
KS-5	99.74	96.00	97.83
KS-7	99.82	96.08	97.91
MKS	99.40	98.92	99.16

Compared to using a single-size convolutional kernel, the F-score with MKS increased by at least 0.94% and the recall increased by at least 2.25%, but the precision decreased slightly, though not more than 0.42%. Corresponding to Equations (21) and (22), the decrease in precision reflects an increase in the system's false reporting of normal samples in the test set as abnormal, and the increase in recall indicates a decrease in the number of missed detections by the system. Compared to misreporting, missed detections are more important to avoid in the actual testing task, and the reduction in missed detections in MKS is 5.36 times greater than the increase in misreporting. Therefore, MKS is more beneficial to improve the overall performance of the system. On the other hand, it also shows that the additional use of MKS with convolutional kernels of the size 5 and 7 is more beneficial for the FOD detection task.

To explore the event detection performance of the models with different convolutional kernel sizes, we performed frame-wise abnormal detection of the signal in Figure 10a with the corresponding model, and the results are shown in Figure 11. Before smoothing, MKS predicts the start and end positions of FOD events more stably compared to KS-3, KS-5, and KS-7. Figure 11d shows the start and end positions of the four FOD events. The oscillations of MKS are significantly smaller than Figure 11a–c. In addition, for the signals in the figure, KS-5 performs the worst, a result that is also consistent with its worst F-score in Table 5. In conclusion, the size of the convolution kernel has an important effect on the FOD detection in the dilated convolution where TCN has been used. For convolution kernel sizes of 3, 5, and 7, size 3 works best, followed by 7, and then 5, while the method using MKS is better than that using the three aforementioned convolution kernels alone.

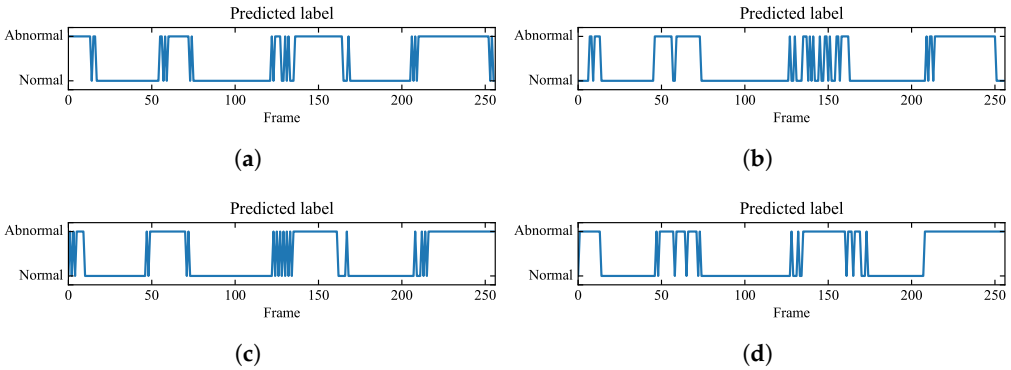


Figure 11. The frame-by-frame prediction results of the model when using different convolution kernel sizes for dilated convolution for the signal in Figure 10: (**a**) KS-3's prediction label, (**b**) KS-5's prediction label, (**c**) KS-7's prediction label, (**d**) MKS's prediction label.

7. Conclusions

In this paper, an intelligent FOD-detection model is proposed based on MS-CTCN. The results show that the method has superior FOD detection accuracy to the traditional CNN and CRNN models. The impacts of different sizes of convolution kernels on the FOD detection performance in MS-CTCN are investigated, and the optimal network parameters applicable to FOD detection are given.

The aim of this study is to identify the presence of FOD. In future work, the research on sound-based FOD detection will be devoted to achieving the identification of FOD types and the localization of FOD through multi-channel sound signals. As a starting point, researchers can focus on developing methods for identifying different types of FOD based on their unique sound signatures. This could involve using advanced signal-processing techniques, such as frequency-domain analysis or machine learning algorithms [57], to extract relevant features from the sound signals. Another area of focus can be on developing methods for localizing FOD based on multi-channel sound signals. This could involve using techniques such as beamforming [58], which uses information from multiple microphone channels to pinpoint the location of a sound source, or using techniques from source separation to separate the sounds from the FOD from the background noise. Incorporating

these new techniques into the detection model could enable the model to identify not just the presence of FOD but also the type of FOD and its location, providing more efficient and reliable support for quality assurance in the rocket-production process.

Author Contributions: Conceptualization, T.L., Y.Z., and Z.R.; methodology, T.L. and S.H.; software, T.L.; validation, T.L., Y.Z. and Z.R.; formal analysis, T.L.; investigation, K.H. and X.Z.; resources, K.Y.; data curation, K.H. and X.Z.; writing—original draft preparation, T.L. and Z.R.; writing—review and editing, Y.Z. and K.Y.; visualization, K.H.; supervision, Y.Z.; project administration, S.H.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Defense Industrial Technology Development Program grant number JCKY2017203A007.

Data Availability Statement: Not applicable.

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

References

1. Li, J.; Liang, G. Simulation of mass and heat transfer in liquid hydrogen tanks during pressurizing. *Chin. J. Aeronaut.* **2019**, *32*, 2068–2084. [CrossRef]
2. Porto, J.F.D.; Loescher, D.H.; Olson, H.C.; Plunkett, P.V. SEM/EDAX Analysis of PIND Test Failures. In Proceedings of the 19th International Reliability Physics Symposium, Las Vegas, NV, USA, 7–9 April 1981; pp. 163–166. ISSN: 0735-0791. [CrossRef]
3. Scaglione, L. Neural network application to particle impact noise detection. In Proceedings of the 1994 IEEE International Conference on Neural Networks (ICNN'94), Orlando, FL, USA, 28 June–2 July 1994; Volume 5, pp. 3415–3419. [CrossRef]
4. Wang, S.; Chen, R.; Zhang, L.; Wang, S. Detection and material identification of loose particles inside the aerospace power supply via stochastic resonance and LVQ network. *Trans. Inst. Meas. Control.* **2012**, *34*, 947–955. [CrossRef]
5. Zhai, G.; Chen, J.; Li, C.; Wang, G. Pattern recognition approach to identify loose particle material based on modified MFCC and HMMs. *Neurocomputing* **2015**, *155*, 135–145. [CrossRef]
6. Wang, G.T.; Liang, X.W.; Xue, Y.Y.; Li, C.; Ding, Q. Algorithm Used to Detect Weak Signals Covered by Noise in PIND. *Int. J. Aerosp. Eng.* **2019**, *2019*, e1637953. [CrossRef]
7. Gao, H.l.; Zhang, H.; Wang, S.j. Research on Auto-detection for Remainder Particles of Aerospace Relay Based on Wavelet Analysis. *Chin. J. Aeronaut.* **2007**, *20*, 75–80. [CrossRef]
8. Wang, S.j.; Gao, H.l.; Zhai, G.f. Research on Feature Extraction of Remnant Particles of Aerospace Relays. *Chin. J. Aeronaut.* **2007**, *20*, 253–259. [CrossRef]
9. Liu, Y.; Li, S.; Wang, J.; Zeng, H.; Lu, J. A computer vision-based assistant system for the assembly of narrow cabin products. *Int. J. Adv. Manuf. Technol.* **2015**, *76*, 281–293. [CrossRef]
10. Xu, H.; Han, Z.; Feng, S.; Zhou, H.; Fang, Y. Foreign object debris material recognition based on convolutional neural networks. *EURASIP J. Image Video Process.* **2018**, *2018*, 21. [CrossRef]
11. Wang, Y.S.; Liu, N.N.; Guo, H.; Wang, X.L. An engine-fault-diagnosis system based on sound intensity analysis and wavelet packet pre-processing neural network. *Eng. Appl. Artif. Intell.* **2020**, *94*, 103765. [CrossRef]
12. Zhang, S.; He, Q.; Ouyang, K.; Xiong, W. Multi-bearing weak defect detection for wayside acoustic diagnosis based on a time-varying spatial filtering rearrangement. *Mech. Syst. Signal Process.* **2018**, *100*, 224–241. [CrossRef]
13. Huang, Q.; Xie, L.; Yin, G.; Ran, M.; Liu, X.; Zheng, J. Acoustic signal analysis for detecting defects inside an arc magnet using a combination of variational mode decomposition and beetle antennae search. *ISA Trans.* **2020**, *102*, 347–364. [CrossRef] [PubMed]
14. Tran, M.Q.; Liu, M.K.; Elsisi, M. Effective multi-sensor data fusion for chatter detection in milling process. *ISA Trans.* **2022**, *125*, 514–527. [CrossRef]
15. Yao, Y.; Zhang, S.; Yang, S.; Gui, G. Learning Attention Representation with a Multi-Scale CNN for Gear Fault Diagnosis under Different Working Conditions. *Sensors* **2020**, *20*, 1233. [CrossRef]
16. Parey, A.; Singh, A. Gearbox fault diagnosis using acoustic signals, continuous wavelet transform and adaptive neuro-fuzzy inference system. *Appl. Acoust.* **2019**, *147*, 133–140. [CrossRef]
17. Zhang, G.; Wang, J.; Han, B.; Jia, S.; Wang, X.; He, J. A Novel Deep Sparse Filtering Method for Intelligent Fault Diagnosis by Acoustic Signal Processing. *Shock Vib.* **2020**, *2020*, e8837047. [CrossRef]
18. Li, X.; Wan, S.; Liu, S.; Zhang, Y.; Hong, J.; Wang, D. Bearing fault diagnosis method based on attention mechanism and multilayer fusion network. *ISA Trans.* **2022**, *128*, 550–564. [CrossRef]
19. Glowacz, A.; Glowacz, W.; Glowacz, Z.; Kozik, J. Early fault diagnosis of bearing and stator faults of the single-phase induction motor using acoustic signals. *Measurement* **2018**, *113*, 1–9. [CrossRef]
20. Nakamura, H.; Asano, K.; Usuda, S.; Mizuno, Y. A Diagnosis Method of Bearing and Stator Fault in Motor Using Rotating Sound Based on Deep Learning. *Energies* **2021**, *14*, 1319. [CrossRef]
21. Heittola, T.; Mesaros, A.; Eronen, A.; Virtanen, T. Context-dependent sound event detection. *EURASIP J. Audio Speech Music. Process.* **2013**, *2013*, 1. [CrossRef]

22. Mesaros, A.; Heittola, T.; Dikmen, O.; Virtanen, T. Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 151–155, ISSN: 2379-190X. [CrossRef]
23. Huang, G.; Heittola, T.; Virtanen, T. Using Sequential Information in Polyphonic Sound Event Detection. In Proceedings of the 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo, Japan, 17–20 September 2018; pp. 291–295. [CrossRef]
24. Adavanne, S.; Politis, A.; Virtanen, T. Multichannel Sound Event Detection Using 3D Convolutional Neural Networks for Learning Inter-channel Features. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–7, ISSN: 2161-4407. [CrossRef]
25. Li, Y.; Liu, M.; Drossos, K.; Virtanen, T. Sound Event Detection Via Dilated Convolutional Recurrent Neural Networks. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 286–290, ISSN: 2379-190X. [CrossRef]
26. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef]
27. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555. https://doi.org/10.48550/arXiv.1412.3555.
28. Wang, Y.; Zhao, G.; Xiong, K.; Shi, G.; Zhang, Y. Multi-Scale and Single-Scale Fully Convolutional Networks for Sound Event Detection. *Neurocomputing* **2021**, *421*, 51–65. [CrossRef]
29. Guirguis, K.; Schorn, C.; Guntoro, A.; Abdulatif, S.; Yang, B. SELD-TCN: Sound Event Localization & Detection via Temporal Convolutional Networks. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–21 January 2021; pp. 16–20, ISSN: 2076-1465. [CrossRef]
30. Frigieri, E.P.; Brito, T.G.; Ynoguti, C.A.; Paiva, A.P.; Ferreira, J.R.; Balestrassi, P.P. Pattern recognition in audible sound energy emissions of AISI 52100 hardened steel turning: A MFCC-based approach. *Int. J. Adv. Manuf. Technol.* **2017**, *88*, 1383–1392. [CrossRef]
31. Warren Liao, T. Feature extraction and selection from acoustic emission signals with an application in grinding wheel condition monitoring. *Eng. Appl. Artif. Intell.* **2010**, *23*, 74–84. [CrossRef]
32. Brahim, J.; Loubna, R.; Noureddine, F. Rnn-And Cnn-Based Weed Detection For Crop Improvement: An Overview. *Foods Raw Mater.* **2021**, *9*, 387–396.
33. Jabir, B.; Falih, N.; Rahmani, K. Accuracy and Efficiency Comparison of Object Detection Open-Source Models. *Int. J. Online Biomed. Eng. (iJOE)* **2021**, *17*, 165. [CrossRef]
34. Zinemanas, P.; Cancela, P.; Rocamora, M. End-to-end Convolutional Neural Networks for Sound Event Detection in Urban Environments. In Proceedings of the 2019 24th Conference of Open Innovations Association (FRUCT), Moscow, Russia, 8–12 April 2019; pp. 533–539, ISSN: 2305-7254. [CrossRef]
35. Wang, C.Y.; Tai, T.C.; Wang, J.C.; Santoso, A.; Mathulaprangsan, S.; Chiang, C.C.; Wu, C.H. Sound Events Recognition and Retrieval Using Multi-Convolutional-Channel Sparse Coding Convolutional Neural Networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1875–1887. [CrossRef]
36. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in Vision: A Survey. *ACM Comput. Surv.* **2022**, *54*, 200:1–200:41. [CrossRef]
37. Imoto, K.; Kyochi, S. Sound Event Detection Utilizing Graph Laplacian Regularization with Event Co-Occurrence. *Ieice Trans. Inf. Syst.* **2020**, *E103.D*, 1971–1977. [CrossRef]
38. Kim, S.J.; Chung, Y.J. Multi-Scale Features for Transformer Model to Improve the Performance of Sound Event Detection. *Appl. Sci.* **2022**, *12*, 2626. [CrossRef]
39. Kong, Q.; Xu, Y.; Wang, W.; Plumbley, M.D. Sound Event Detection of Weakly Labelled Data With CNN-Transformer and Automatic Threshold Optimization. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2450–2460. [CrossRef]
40. D'Ascoli, S.; Touvron, H.; Leavitt, M.L.; Morcos, A.S.; Biroli, G.; Sagun, L. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. In Proceedings of the 38th International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 2286–2296, ISSN: 2640-3498.
41. Wang, J.J.; Wang, C.; Fan, J.S.; Mo, Y.L. A deep learning framework for constitutive modeling based on temporal convolutional network. *J. Comput. Phys.* **2022**, *449*, 110784. [CrossRef]
42. Wang, Y.; Zhao, G.; Xiong, K.; Shi, G. MSFF-Net: Multi-scale feature fusing networks with dilated mixed convolution and cascaded parallel framework for sound event detection. *Digit. Signal Process.* **2022**, *122*, 103319. [CrossRef]
43. Kopparapu, S.K.; Laxminarayana, M. Choice of Mel filter bank in computing MFCC of a resampled speech. In Proceedings of the 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010), Kuala Lumpur, Malaysia, 10–13 May 2010; pp. 121–124. [CrossRef]
44. Allen, J.; Rabiner, L. A unified approach to short-time Fourier analysis and synthesis. *Proc. IEEE* **1977**, *65*, 1558–1564. [CrossRef]
45. Tak, R.N.; Agrawal, D.M.; Patil, H.A. Novel Phase Encoded Mel Filterbank Energies for Environmental Sound Classification. In Proceedings of the Pattern Recognition and Machine Intelligence; Shankar, B.U., Ghosh, K., Mandal, D.P., Ray, S.S., Zhang, D.,

Pal, S.K., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2017; pp. 317–325. [CrossRef]
46. Li, Y.; Du, X.; Wan, F.; Wang, X.; Yu, H. Rotating machinery fault diagnosis based on convolutional neural network and infrared thermal imaging. *Chin. J. Aeronaut.* **2020**, *33*, 427–438. [CrossRef]
47. Vafeiadis, A.; Votis, K.; Giakoumis, D.; Tzovaras, D.; Chen, L.; Hamzaoui, R. Audio content analysis for unobtrusive event detection in smart homes. *Eng. Appl. Artif. Intell.* **2020**, *89*, 103226. [CrossRef]
48. Hasan, M.J.; Islam, M.M.M.; Kim, J.M. Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions. *Measurement* **2019**, *138*, 620–631. [CrossRef]
49. Hasan, M.J.; Islam, M.M.M.; Kim, J.M. Multi-sensor fusion-based time-frequency imaging and transfer learning for spherical tank crack diagnosis under variable pressure conditions. *Measurement* **2021**, *168*, 108478. [CrossRef]
50. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv* **2015**. [CrossRef]
51. Jiao, J.; Zhao, M.; Lin, J.; Liang, K. A comprehensive review on convolutional neural network in machine fault diagnosis. *Neurocomputing* **2020**, *417*, 36–63. [CrossRef]
52. Biscione, V.; Bowers, J. Learning Translation Invariance in CNNs. *arXiv* **2020**. [CrossRef]
53. Sainath, T.N.; Vinyals, O.; Senior, A.; Sak, H. Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia , 19–24 April 2015; pp. 4580–4584, ISSN: 2379-190X. [CrossRef]
54. Oord, A.v.d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. *arXiv* **2016**. [CrossRef]
55. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778, ISSN: 1063-6919. [CrossRef]
56. Mesaros, A.; Heittola, T.; Virtanen, T. Metrics for Polyphonic Sound Event Detection. *Appl. Sci.* **2016**, *6*, 162. [CrossRef]
57. Gao, D.W.; Zhu, Y.S.; Yan, K.; Fu, H.; Ren, Z.J.; Kang, W.; Guedes Soares, C. Joint learning system based on semi–pseudo–label reliability assessment for weak–fault diagnosis with few labels. *Mech. Syst. Signal Process.* **2023**, *189*, 110089. [CrossRef]
58. Shaw, A.; Smith, J.; Hassanien, A. MVDR Beamformer Design by Imposing Unit Circle Roots Constraints for Uniform Linear Arrays. *IEEE Trans. Signal Process.* **2021**, *69*, 6116–6130. [CrossRef]