Article

A Novel Feature Selection Based on VMD and Information Gain for Pipe Blockages

Xuefeng Zhu ^{1,2}, Zao Feng ^{1,2,*}, Jiande Wu ^{1,2} and Wei-quan Deng ^{1,2}

¹ Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650000, China; zhuxuefeng@stu.kust.edu.cn (X.Z.); wjiande@kust.edu.cn (J.W.); weiquan.deng@kust.edu.cn (W.D.)

² Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China

* Correspondence: zaofeng@kust.edu.cn; Tel.: +86-139-1187-5098

Abstract: Targeting the challenge of determining the degree of blockage in buried pipelines and the difficulty of effectively extracting blockage features, a blockage detection method integrating variational mode decomposition (VMD) and information gain is proposed. Acoustic impulse response signals were obtained by deconvolving the output signals of the system, which were then subjected to VMD to obtain 12 components in different frequency ranges. Next, information gain (IG) was introduced to characterize the 12 components quantitatively, through which the components containing rich information about the pipe conditions were selected out. Meanwhile, sound pressure level conversion was performed on the selected components to amplify any changes in the sound field. Finally, the root mean square entropy (RMSE) was calculated to constitute the feature eigenvectors, which were input into Random Forests (RF) classifier for defect identification of pipeline. As the experimental results demonstrate, the proposed method is capable of determining the degree of blockage effectively in the running state. Meanwhile, it can also eliminate the interference of functional parts such as lateral connections during the identification process, thereby improving the identification accuracy. The present study has shown both theoretical significance and application value in the field of defect detection and recognition.

Keywords: sewage blockage identification; VMD; information gain; feature selection

Citation: Zhu, X.; Feng, Z.; Wu, J.; Deng, W. A Novel Feature Selection Based on VMD and Information Gain for Pipe Blockages. *Appl. Sci.* **2021**, *11*, 10824. <https://doi.org/10.3390/app112210824>

Academic Editor: José A. F. O. Correia

Received: 30 September 2021

Accepted: 8 November 2021

Published: 16 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 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

Sewage are the lifelines of urban construction and social development. During the sewage operation, factors such as overload, fatigue, and environmental pollution result in cracks, blockages, leakages, and other functional defects inside it, thereby lowering its service life [1]. A blockage is a ubiquitous phenomenon during pipeline operations. In case of slight blockage of the pipe, the blocked area will enlarge continuously over time if timely detection and management are not given, which eventually leads to severe blockage [2]. The severe blockage will compromise the carrying capacity of the pipe and the reliability of the system; will increase the possibility of environmental pollution and the redundancy of the system; and will cause over-pressure of partial pipes in the system to increase the possibility of leakage, thereby ultimately resulting in a serious waste of water resources and environmental pollution. The losses resulting from severe and multiple blockages can be minimized as long as the minor blockages can be detected and handled in time [3]. Given that the pipelines are buried deep underground, the evaluation of their operational conditions is complicated and challenging [4]. Hence, non-destructive testing of buried pipeline conditions is profoundly meaningful for ensuring the high efficiency and reliability of their normal operation, which is the focus and challenge of urban infrastructure maintenance [5,6].

To date, a variety of detection methods have emerged [7], such as ultrasonic method [8], closed-circuit television (CCTV) [9,10], and sewer scanner evaluation technology (SSET) [11]. Acoustic detection, as a non-destructive testing technique, has been used widely in pipe detection owing to its advantages of simple operation, long detection range, low cost, and no heavy reliance on the subjectivity of testing personnel [12,13]. However, the acoustic signal transmission path of pipeline blockage is complex. The sound waves undergo varying degrees of physical phenomena such as reflection, refraction, and diffraction during passage through the discontinuous interface of acoustic impedance [14]. Acoustic impulse response signals are weak, and it is easy to be interfered with by the external environment noise, thus the signal-to-noise ratio is low [15]. On the other hand, due to the coupling of excitation and response, the acoustic signal of pipeline blockage presents obvious non-linearity and non-stationarity [16]. Therefore, it is necessary to pre-process the original acoustic signal to enhance the feature information to obtain better recognition results.

Feature extraction and selection is a key problem in pipeline blockage recognition [17,18]. Some methods, such as power spectral density function (PSD) [19], short-time Fourier transform (STFT) [20], empirical mode decomposition (EMD) [21], and local mean decomposition (LMD) [22] are used to decompose signals. VMD decomposes each mode from low frequency to high frequency and rebuilds the original signal by selecting the effective mode [23,24]. At present, K-L divergence [25], the difference of energy distribution [26], kurtosis value [27], and other methods are used to screen the effective components of decomposition. Nevertheless, within different frequency ranges, the sensitivity level of acoustic signals to blockages is affected by factors such as the inner diameter, length, and embedding conditions of pipes, which is also correlated with the size and severity of blockages. Hence, a detailed analysis is necessary regarding the amount of feature information contained in different frequencies of sound signals under the blockage condition [28]. The information gain in filter feature selection is introduced to select effective Intrinsic Mode Functions (IMF) components, extract features with large contributions and remove redundant feature information, and construct the best combination of differential features to represent blockage [29]. The advantage is that the screening process does not carry out traversal trial, the combination effect of characteristic factors is less considered and the calculation speed is faster [30].

In summary, this paper proposes a novel approach based on VMD, information gain, sound pressure level, and root mean square entropy (RMSE) for identifying blockages in pipelines. First, Butterworth filtering is performed on the collected acoustic signals initially in this study. Then variational mode decomposition (VMD) is implemented on the filtered signals, followed by the simplification and filtering of components based on information gain. Finally, conversion of sound pressure level is carried out on the filtered components, and the RMSE are extracted as the eigenvectors, which are input into four different classifiers k-Nearest Neighbors (KNN), Support Vector Machines (SVM), Extreme Learning Machine (ELM), and Random Forests (RF), thereby achieving the classification and identification of the pipe condition. The present study has great practical significance for the detection of pipe blockage severity.

The paper is organized as follows: Section 2 introduces the VMD and information gain methodologies and provides the basic framework of the proposed hybrid model. Section 3 Brief Experimental setup and experiment conditions. Section 4 is pipe blockage methodology which includes feature extraction, selection, and fault identification. Section 4 verifies the diagnostic performance of the proposed method. Finally, Section 5 concludes the work.

2. Materials and Methods

2.1. VMD

VMD, which was proposed by Dragomiretskiy et al. in 2014 [31], is an adaptive decomposition method for multicomponent signals. Instead of cyclic screening and stripping of the signals, the method transfers the process of signal decomposition into the variational

framework and determines the frequency center and bandwidth of various components by iteratively searching for the optimal solution of the variational model [32]. As it decomposes an actual signal $x(t)$ into k numbers of discrete mode components $u(t)$, it can adaptively accomplish the effective separation of frequency-domain part of the signals, as well as various components, which highlights the local features of data and exhibits better noise robustness.

Construction of variational problem is a process in which the sum of k mode components is made equal to the original signal $x(t)$. The steps for estimating the frequency bandwidth of various mode signals are as follows: (1) Hilbert transform is performed on various mode functions $u_k(t)$ to obtain their respective marginal spectra. (2) The functions $u_k(t)$ are mixed by an exponential modification to estimate their center frequency $\omega_k(t)$, and the one-sided spectra are modulated to the corresponding basebands. (3) The bandwidths of various mode signals $u_k(t)$ are estimated through Gaussian smoothing of demodulated signals to minimize the sum of their bandwidths. The constrained variational problem arising above is expressed as:

$$\begin{cases} \min_{\{u_k\}, \{\omega_k\}} \left\{ \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) \cdot u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \\ s.t. \sum_k u_k = x(t) \end{cases} \quad (1)$$

where $\delta(t)$ represents the impulse function, $\{u_k\} = \{u_1, u_2, \dots, u_k\}$ represents the set of various mode function components, and $\{\omega_k\} = \{\omega_1, \omega_2, \dots, \omega_k\}$ denotes the center frequency of various mode components.

To solve the optimal solution of the foregoing variational model, the Lagrange multiplication operator $\lambda(t)$ and the quadratic penalty factor α are used to transform the constrained variational problem into an unconstrained one. Relevant expression is:

$$L(\{u_k\}, \{\omega_k\}, \lambda) = \alpha \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) \cdot u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 + \|x(t) - \sum_k u_k(t)\|^2 + \left\langle \lambda(t), x(t) - \sum_k u_k(t) \right\rangle \quad (2)$$

During the solving process, the alternate direction method of multipliers (ADMM) is employed, where u_k^{n+1} , ω_k^{n+1} and λ_k^{n+1} are updated alternately. The saddle point ζ of Equation (3) is searched, which is precisely the optimal solution of the variational problem in Equation (2). Accordingly, the signal $x(t)$ is decomposed into k numbers of discrete mode components $u(t)$.

2.2. Information Gain-Based Selection of Effective IMF Components

The method operates on the following principle: Given a sample set D and continuous attributes a , assuming that a have n different values on D , which are sorted from small to large and are represented by $\{a^1, a^2, \dots, a^n\}$. The values of adjacent attributes are set as a^i and a^{i+1} , while the division results generated by t are identical when any value in an $[a^i, a^{i+1}]$ interval is taken. Thus, for continuous attributes a , the median point $a^i + a^{i+1} / 2$ of the interval $[a^i, a^{i+1}]$ is assigned as the candidate partition point. The partition point t can divide the given sample set D into subset sums $D_t^- D_t^+$, of which D_t^- represents the samples with the value range of attributes a not greater than t , while D_t^+ denotes those samples with the value range of attributes a greater than t . The relevant computational formula for information gain is [33]:

$$Gain(D, a) = \max_{t \in T_a} Gain(D, a, t) = \max_{t \in T_a} Ent(D) - \sum_{\lambda \in \{-, +\}} \frac{|D_t^\lambda|}{|D|} Ent(D_t^\lambda) \quad (3)$$

where $Gain(D, a, t)$ denotes the information gain of the given sample set D following the dichotomy based on the partition point t . According to the following steps, the partition point that makes $Gain(D, a, t)$ the maximum is selected:

Step 1: The given sample set D is computed.

Step 2: For each attribute a , i.e., component, the information gain $\text{Gain}(D, a, t)$ and partition point $T_a = \left\{ \frac{a^i + a^{i+1}}{2} \mid 1 \leq i \leq n-1 \right\}$ are computed.

Step 3: After selecting out the maximum value $\text{Gain}(D, a, t)$, the corresponding component is chosen as the root node of the decision tree. The sample-set is split into two parts according to the computed partition point, where the samples greater than t are denoted by D_t^+ , and the samples less than or equal to t are denoted by D_t^- .

Step 4: The remaining components are deemed as the dataset of the previous node, and the component with the largest information gain is chosen as the non-leaf node split by the root node.

Step 5: For each component, Steps 3 and 4 are repeated if the information gain value is greater than the given threshold.

Step 6: The component selection is terminated and completed if the maximum information gain is less than a given threshold. The information gain-based component selection method allows accurate extraction of the components that contain the majority of blockage features.

2.3. Sound Pressure Level Conversion

The conversion of sound pressure level amplifies the content of acoustic signals and enhances the distinction among various severities of blockages, thus that the characteristic information about different pipe conditions is more easily extractable in the subsequent decomposition [34]. The computational formula for the sound pressure level is as follows:

$$L_p = 20 \lg \frac{p_e}{p_0} \quad (4)$$

where p_e denotes the effective sound pressure value of the original acoustic signal and p_0 denotes the effective value of reference sound pressure, whose value is assigned as 2×10^{-5} Pa herein.

2.4. RMSE

For vibration signals, RMS value indicates the change in instantaneous signal amplitude within their sampling period, which can reflect their respective vibration energies. Meanwhile, information entropy represents the system complexity resulting from multiple uncertain factors [35]. The RMSE E_{RMS} is obtained by integrating the RMS into the information entropy, which combines the advantages of the two. Different fault types can be represented with different RMSE values [36].

The computational procedure for the RMS entropy is as follows:

Step 1: The RMS values of various IMF components are computed following the sound pressure level conversion:

$$R_i = \sqrt{\frac{1}{N} \sum_{n=1}^N IMF_i^2(n)} \quad (5)$$

where R_i denotes the RMS value of the i -th component and N is the number of sample points.

Step2: The RMS values are constituted into an eigenvector A:

$$A = [R_1, R_2, \dots, R_k] \quad (6)$$

Step3: The RMS values are homogenized:

$$P = R_1 + R_2 + \dots + R_k, P_i = \frac{R_i}{P}, \sum_{i=1}^K P_i = 1 \quad (7)$$

Step4: The RMSE can be derived from the definition of information entropy as:

$$E_{RMS} = -\sum_{i=1}^K P_i \log_2 P_i \quad (8)$$

where K represents the number of IMF components; and E_{RMS} denotes the RMS entropy.

2.5. Proposed Feature Extraction Method

The flowchart of VMD-Information Gain for the pipe blockage detection process is illustrated in Figure 1.

Figure 1. The flow diagram of the proposed method.

Step 1: Each 300 sets of sampling are performed on the 9 statuses of buried pipelines to obtain 2700 sets of sample data in total.

Step 2: The original acoustic signals in various statuses are denoised through the Butterworth filter, thereby obtaining acoustic signals with a [0–6000] Hz frequency range.

Step 3: The denoised signals in various statuses are decomposed by the VMD, and the number of optimal decomposition layers K is decided by determining whether over-decomposition is produced or not.

Step 4: VMD operation is performed on the 100 sets of sampled data in 9 statuses, thereby deriving K components. By utilizing information gain, M effective components are filtered from the K components. Information gain values of the components are selected as per the principle of decision tree selection. A threshold is set up, which is then compared with the filtered information gain values. Filtering is terminated if the information gain value of a component signal is less than the threshold.

Step 5: Sound pressure level conversion is performed on the filtered M effective components, and then the RMS entropy values are computed.

Step 6: The RMS entropy in various statuses are constituted separately into eigenvectors, which are then input into four different classifiers SVM, KNN, ELM, and RF, for effective identification of the pipeline operating conditions.

3. Experimental Setup and Experiment Conditions

An experimental platform was built in the laboratory [37], as shown in Figure 2, for acoustic detection of pipe blockages. In this experiment, the blocked substances were simulated with clay semi-cylinders, and sinusoidal sweep signals were used as the excitation signals were given their adjustable frequency bands as per demands. The use of sinusoidal sweep signals for exciting multi-DOF systems allows energy concentration of acoustic signals within the frequency range sensitive to the pipe thus that the required information in such frequency band can be easily elicited. They are often used as excitation signals for outdoor detection.

Figure 2. Experimental platform for pipe blockage detection.

The clay pipeline had a diameter of 150 mm and a length of 14.4 m. During detection, a computer with LabVIEW software was used to control the virtual instrument for generating sinusoidal sweep signals within a 100–6000 Hz frequency range. Then, the analog output port of NI PXIe-6363 was controlled via the DAQ assistant of LabVIEW to output analog voltage signals. After amplification of the signals with a power amplifier, the loudspeaker was driven to generate audio signals, which were then transmitted into the pipeline to serve as an excitation signal source. Such sound wave signals undergo complex interactions with the discontinuous interface of acoustic impedance in the pipe interior. The echo signals were received by the microphone placed at the pipe head end, which was then uploaded to the computer for storage. A sampling frequency of 44,100 Hz. Changes in the acoustic performance of pipe interior were identified by analyzing the received signals, during which the LM4950 power amplifier (Texas Instruments in Dallas, TX, USA) was used, as well as the FR874OHM loudspeaker (Visaton in Germany) and the SPM0208HE5 microphone (Knowles Acoustics in Illinois, USA).

For the sake of practical simulation, experiments were conducted at varying degrees of blockages on the experimental platform shown in Figure 2. Water flow was provided in the pipe for simulating the pipe condition, whose rate was dependent on the water pump. In the present experiments, the maximum rate of simulated water flow was 7 L/s, and other simulated water flow rates were 0.42 L/s, 1 L/s, 1.8 L/s, 4.25 L/s, and 6.1 L/s, respectively,

which were used to form intra-pipe water levels of different heights. The severity of pipe blockage was defined by the laboratory as the percentage of blockage height in the cross-sectional pipe height. The simulated blockages with heights of 20 mm, 40 mm, and 55 mm were placed separately in the pipe of 150 mm in diameter. The heights of these rigid, non-porous blockages accounted for 13%, 26%, and 37% of the cross-sectional pipe area, respectively, which were approximately considered to be slightly blocked, moderately blocked, and moderately to severely blocked. The detailed data about the experiments are shown in Table 1.

Table 1. Pipe conditions information.

No	State	Pipe Condition	Training Samples	Testing Samples
1	normally	Clean	200	100
2		Lateral Connection (LC)	200	100
3		a 20 mm blockage	200	100
4	single blockage	a 40 mm blockage	200	100
5		a 50 mm blockage	200	100
6		a 40 mm blockage and a LC	200	100
7	multiple blockages	a 55 mm blockage and a LC	200	100
8		a 40 mm blockage and a 55 mm blockage	200	100
9		a 40 mm blockage, a 55 mm blockage, and LC	200	100

In this study, the pipe condition was set as a normal running empty state. There were conventional parts, i.e., lateral connections (LC), inside the empty pipeline that was in a normal operating state. The pipe conditions with the presence of single blockages included a 20 mm blockage; a 40 mm blockage; and a 55 mm blockage. Meanwhile, the pipe conditions with multiple blockages included coexistence of a 40 mm blockage and a 55 mm blockage that were placed in different positions inside the pipeline; coexistence of a 40 mm blockage and a lateral connection; coexistence of a 55 mm blockage and a lateral connection; and coexistence of a 40 mm blockage, a 55 mm blockage, and a lateral connection. This comes to a total of 9 pipe conditions. There were 100 sets of samples for each condition, with a total of 900 sets.

4. Experimental Results and Analyses

4.1. Data Pre-Processing

In signal processing, the aim is to maximize the extraction of valuable information from the signals. For physical systems such as buried pipelines, the sound pulse or frequency response is a characteristic quantity that contains detailed information, including the system geometry, sound velocity, and boundary conditions. Any changes in the properties of the physical system are reflected in the variations of sound pulse or relevant frequency response. During propagation in the pipe, acoustic signals collide with the pipe wall, blockages, and lateral connections to cause reflection, refraction, and diffraction. To investigate the frequency content of measured sound pressure signals, these time-domain signals were transformed from the time domain to the frequency domain via the Fourier transform. In Figure 3, the time-domain and frequency-domain waveforms of the acoustic signals are illustrated in the typical pipe conditions.

Figure 3. Time domain and frequency domain diagrams of pipe conditions (normally, single blockage, multiple blockages): (a) time domain; (b) frequency domain.

In Figure 3a, the “Distance” of abscissa is the propagation distance of sound waves in the pipe, which facilitates the positional identification of pipe tail end, blockages, and lateral connections in the time-frequency domain diagrams. In the present paper, the propagation distance equals the propagation time multiplied by the sound propagation velocity in the air (approximately 340 m/s). The reasons are that in the drainage pipe, the water flow accounts for 20% of the cross-sectional pipeline area, and the sounds are transmitted mostly in the air. Figure 3a depicts the time-domain waveforms of original response signals. With the increasing propagation distance of sound waves in the pipeline, the energy of sound wave propagation was attenuated continuously, which was manifested as the decreasing signal amplitude with increasing distance in the sound pressure graphs. Nevertheless, the sound pressure waveforms in the three pipe conditions exhibit no distinct differences in the time domain, thus that it is difficult to distinguish the position and size of blockages in the time-domain waveforms, which are even hardly distinguishable from the pipe fittings (lateral connections). This is attributed mainly to the presence of environmental noise and the varying responses among objects to the frequency bands in different ranges. As is clear from Figure 3b, the frequencies in the three pipe conditions are concentrated primarily in [0–6000 Hz] in the spectrograms, while the components in other frequency bands are weak. Hence, the Butterworth filter was utilized to denoise the acoustic signals.

Given the difficulty of identifying the pipe operational status based on the analysis of conventional time-domain waveforms and spectrograms for sound pressure signals, further analysis, and processing of these signals were needed. The time-frequency map reflects the information that the frequency of acoustic signal changes with time. The time-frequency map reflects the energy carried by each frequency component of the signal through the cold and warm color. The warmer the color, the greater the energy. The common time-frequency transformation methods include short-time Fourier transform, Wigner distribution, and wavelet transform. Compared with the first two, wavelet transform has adaptive time-frequency resolution and a faster algorithm. Therefore, the sound signal was generated into a time-frequency map by continuous wavelet transform. As shown in Figure 4, the operational condition of a typical pipe is selected for time-frequency map analysis.

Figure 4. Time-frequency map: (a) clean; (b) lateral connection; (c) 55 mm blockage; (d) 55 mm + 40 mm blockage.

As shown in Figure 4a, lighter color indicates higher energy at the corresponding site. According to the time-frequency map of the normal pipeline. The energy concentrations are present only at the pipeline head and tail ends, while other locations show no energy concentration. For single blockages and lateral connections, as can be seen from Figure 4b,c, the energies have a good time-frequency concentration at the locations of lateral connections and blockages in the pipe, which conform to the acoustic theory. The energies gather at the blockage site while diverging slightly at the lateral connection site. In the case of multiple blockages, as can be seen from Figure 4d, with the increase in blockages, gradual attenuation was observed in the energies between various blockages and lateral connections. Moreover, the frequency bands, where the energies in different operating conditions appear, remain relatively stable. However, energy overlapping occurs when there are two blockages and a lateral connection in the pipeline. Although the energy spectrogram can locate the positions of the pipe head and tail ends, blockages, and lateral connections accurately, it is unable to determine the pipe operating condition properly.

Hence, further identification and research of the pipe condition were necessary. Since the sound waves of different frequencies were reflected in varying degrees at different rates, the intensities of signals and their sensitivities to the pipe condition vary in different frequency ranges. The characteristic components irrelevant to the pipeline blockage will interfere with the classification, thereby resulting in lowered classification accuracy.

4.2. Sensitive IMF Selection

VMD is applied to the acoustic signals for buried pipe conditions. The center frequency values of the components of the inherent mode function obtained by VMD decomposition are distributed from low to high, and the number of IMF components K is evaluated from 1. If the center frequency of the last IMF reaches its maximum value for the first time, it means

that no insufficient decomposition occurs, and the value of K increases gradually until the maximum center frequency remains relatively stable. The parameter of decomposing mode number K was decided in advance, where it was determined according to a previous study [38]. Therefore, the K value in the paper was set as 12.

As played in Figure 5, a set of acoustic signals with the coexistence of 40 mm, 55 mm blockages, and lateral connections in the pipe were selected for VMD processing, the K value of 9 pipe operating conditions was finalized at 12 after the pre-setting process and the analysis of maximum center frequency for various components. Thereby deriving a total of 12 IMF components. The 12 IMF components after VMD decomposition represent the signal characteristics at different frequency scales, but the effective components were different for different operational conditions in the pipeline. The goals of component filtering were to simplify the feature space, eliminate complexity, and enhance the system performance. A component is considered more important if it can bring richer information to the classification model. Its presence or absence in the classification model leads to a larger change in the amount of information, and the difference in information amount before and after its addition is precisely the information gain it brings to the model. Utilizing the information gain, M effective components were selected from the 12 original components, which were used for the identification of different pipe conditions.

Figure 5. VMD component decomposition diagram.

The major steps for information gain based filtration of effective components are as follows:

Step 1: Computation of the given sample set D . Nine types of pipe conditions are selected, each of which has 300 samples, totaling 2700 samples. For each sample, 12 components form a dataset D .

Step 2: Computation of information gain. The information gains of the 2700 samples are calculated according to formula (3), as well as the corresponding partition points.

Step 3: Selection of root node. As shown in Figure 6, the information gain value of “IMF1” is the largest among the 12 components, which is thus selected as the root node of the decision tree. The data set D is split into two parts based on the calculated partition points.

Figure 6. Component diagrams from information gain filtering.

Step 4: Selection of subnodes. For the remaining 11 components, the information gains are calculated as per Step 2, and the component with the largest information gain is selected as the leaf node of the decision tree. As shown in Figure 6, “IMF3” is the largest component making the dataset D less than or equal to the information gain in IMF1, which is thus selected as the subnode of the left subtree at the second layer. Meanwhile, “IMF11” is the maximum value making the dataset D greater than the information gain in IMF1, which is selected as the subnode of the right subtree at the second layer. After passing through the second layer, the data are divided into four parts.

Step 5: For the remaining nine discrete wavelet packet components, their information gains are calculated as per Step 2, where the given threshold for information gain is 0.5. As shown in Figure 6, the IMF10 component is selected as the left subtree of the left subtree at the third layer; the IMF6 component is selected as the right subtree of the left subtree at the third layer; the IMF4 component is selected as the left subtree of the right subtree at the third layer, and the IMF2 component is selected as the right subtree of the right subtree at the third layer. The selection process of IMF is shown in Table 2.

Table 2. Sensitive IMF selection based on information gain.

Number of Layers	Screening of Sensitive IMF
The first layer selection IMF	IMF1;
The second layer selection IMF	IMF3; IMF11;
The second layer selection IMF	IMF10; IMF6; IMF4; IMF2;

Step 6: After the third layer selection, the maximum information gain is less than the given threshold of 0.5, and the component filtering is terminated and completed. To sum up, the root node of the decision tree is component 1, which has the largest information gain at the first layer. After the second layer selection, component 3 has the largest information gain in the left node, whereas component 11 is the largest in the right node. Meanwhile, the components in the third layer are IMF10, IMF6, IMF4, and IMF2. According to Figure 6, the components selected following the information gain filtering are IMF1, IMF3, IMF11, IMF10, IMF6, IMF4, and IMF2.

4.3. Sound Pressure Level Conversion

The components filtered based on information gain are subject to the conversion of sound pressure level to enhance the discrimination between components for easier feature extraction. In Figure 7, a comparison is made between the sound pressure signals and the signals following sound pressure level conversion.

Figure 7. Conversion diagrams of sound pressure level: (a) sound pressure signal; (b) sound pressure level conversion.

As is clear from Figure 7a, for a normally operating pipeline, the inflection points of signals are not easily distinguishable regardless of whether there is a lateral connection or not. As shown in Figure 7b, the inflection points of signals that underwent sound pressure level conversion are distinguishable. The conversion of sound pressure level can better reflect the local features of signals, which can enhance the differentiation between pipe conditions and improve the sensitivity of acoustic signals.

4.4. RMS Entropy Features Extraction and Blockage Recognition

For 9 different pipe conditions, 300 sets of acoustic signals were collected as samples to perform Butterworth filtering, and then VMD was applied to derive 12 different IMF components. Seven effective components were selected through information gain filtering, which was then subjected to sound pressure level conversion for calculating the RMS of each set of samples. For vibration signals, RMS indicates the change in instantaneous signal amplitude within their sampling period, which can reflect their respective vibration energies. Meanwhile, information entropy represents the system complexity resulting from multiple uncertain factors. The RMSE is obtained by integrating the RMS into the information entropy, which combines the advantages of the two. As shown in Figure 8, the RMSE values of 100 samples of IMF1 components were selected.

As shown in Figure 8, it is clear that for different fault types, their ERMS values fluctuate in different numerical ranges. A rough classification of pipe conditions can be achieved after arrangement and comparison. ERMS can distinguish the blockage state from the normal state, and the effective component ERMS value of the normal pipe is lower than that of the blockage pipe. The analysis shows that when the pipe is blocked, the incident wave propagates along the axial direction of the pipe, and the physical phenomena such as reflection, refraction, and diffraction occur when the blockage occurs. Therefore, the signal is more disordered, which is manifested in the larger ERMS entropy value of the blocked pipe than that of the pipe in a normal operation state. For single blockages of different degrees (20 mm, 40 mm, 55 mm), the three types of pipe blockages are only different in the height of blockage, which leads to the crossover of ERMS values, affecting the final classification effect.

Figure 8. ERMS distributions of nine pipe operating conditions.

From each data sample in the above data set, 200 samples were randomly selected as training samples and 100 testing samples. To verify the performance of VMD-IG-RMSE-RF method, KNN, SVM, ELM, and RF were used to identify the running state of the pipeline. The identification accuracy of 20 cross-verifications is shown in Table 3.

Table 3. Identification accuracies of pipe condition with various classifiers.

Classifier	Max Accuracy (%)	Min Accuracy (%)	Average Accuracy (%)
KNN	90.44	86.67	87.56
SVM	91.11	90.56	90.78
ELM	94.44	90.44	93.68
RF	100	98.77	99.58

It can be seen from Table 3 that the average recognition accuracy of the four classifiers (KNN, SVM, ELM, RF) reaches 87.56%. In addition, it can be seen from Table 3 that the average recognition accuracy of the RF classification is 99.58%, which is higher than the other three classifiers, and the maximum recognition accuracy reaches 100%. The reason for this phenomenon is due to the relatively ideal test environment and the good performance of the data acquisition system. It makes the characteristics of signal data easier to identify. The results show that the method has good performance. Therefore, the classifier selection of the subsequent comparative experiments is RF.

To verify the effectiveness of the methods based on VMD and IG, the proposed method (VMD-IG-RMSE-RF) and some similar available methods (EMD-IG-RMSE, LMD-IG-RMSE, WT-IG-RMSE) were used to analyze the same experimental data mentioned above. The identification results of the four methods are shown in Table 4.

Table 4. Identification accuracy with VMD method and without VMD method.

Feature Extraction Method	Identification Accuracy (%)				Average Identification Accuracy (%)	Standard Deviation
	1	2	3	4		
EMD-IG-RMSE-RF	91.11	90.33	89.11	91.78	90.58	0.54
LMD-IG-RMSE-RF	80.14	82.13	80.31	82.52	81.28	0.21
WT-IG-RMSE-RF	86.13	84.13	85.13	86.78	85.32	0.14
VMD-IG-RMSE-RF	98.56	98.88	100	100	99.58	0.07

It can be seen from Table 4 that the average recognition accuracy of the proposed method is 99.58%, which is significantly higher than that of the other three methods. It shows that VMD can effectively extract the characteristics of pipeline blockage. At the

same time, compared with other methods, the standard deviation of this method is the smallest, which verifies the stability of this method and the effectiveness of VMD and Ig in this method.

In order to demonstrate the effectiveness of IG feature extraction, another set of acoustically detected signals are selected randomly as the testing samples, which includes nine types of pipe conditions. There are 100 sets of samples for each state, totaling 900 samples. Every 12 components are extracted and then subjected to Principal Component Analysis (PCA), Kullback-Leibler Divergence(K-L), and information gain filtering. Among them, PCA is reduced to three dimensions, the K-L divergence chooses four components, and seven IMF components, IMF1, IMF3, IMF11, IMF10, IMF6, IMF4 and IMF2, were selected for information gain. Afterward, the ERMS values of feature parameter sets for components are input to the KNN, SVM, ELM, and EF models, respectively, for examining the accuracy of pipe condition identification.

The average diagnostic accuracy of different methods is shown in Figure 9. The identification accuracy of IG method is superior to other methods. Compared with other methods, the accuracy of information gain filtering is higher. This indicates that the component filtering method based on information gain retains the data feature information to the maximum extent, reduces the interference of redundant features and noise features to components, and improves the model recognition accuracy. Too many or too few sensitive features will reduce the accuracy of blockage identification. If the number of feature sensitivities is too small, there will be less blocking feature information. On the contrary, too many sensitive features will lead to the redundancy of blocking feature information and reduce the accuracy of blocking identification. For different types of classifiers, the accuracy of VMD-IG method is better than other methods, and the proposed feature extraction method is better than the other three methods.

Figure 9. Identification accuracy of different IMF components screening methods.

Through the above analysis, it is found that the accuracy of acoustic signal identification for pipe conditions is improved markedly after component filtering, suggesting its effectiveness. The amount of information about feature parameters varies among components, with the components selected based on information gain containing more features about the pipe condition. Component filtering is effective in feature extraction, which is capable of reducing the data size substantially and can achieve high-accuracy identification with less time. The above results suggest that the present method is not only effective in identifying the pipeline blockage severity during operation but can also eliminate the influence of conventional parts such as lateral connections on the blockage identification, which improves the identification accuracy of pipe condition.

In order to further investigate the ability of VMD-IG-RMSE algorithm to identify pipe operational conditions and the details of blockage misjudgment, a multi-classification confusion matrix is introduced to quantitatively analyze the results of pipe operational conditions recognition in detail. The confusion matrix comprehensively reflects the recognition accuracy and number of misjudgments of pipelines at different blockage levels, as well as the misjudgment type of real blockage type. The quantization diagram of the confusion matrix of four classifiers KNN, SVM, ELM, and RF is shown in Figure 10.

Figure 10. Confusion matrices for four classifiers: (a) KNN; (b) SVM; (c) ELM; (d) sound pressure level conversion.

It is obvious from Figure 10: Class 1 and Class 2, respectively, represent the normal operating condition of the pipe clean and lateral connection, and the recognition accuracy of both the normal operation of the pipe on the test set reaches 100%. Therefore, the algorithm achieves 100% recognition accuracy between the normal operating conditions and blocked. Class 3, Class 4, and Class 5 represent a 20 mm blockage, a 40 mm blockage, and a 50 mm blockage for a single blockage in the pipe, respectively. By analyzing the types of misjudgment of blockage, it can be seen that the above misjudgment is the error between single blockage categories, which belongs to the misjudgment of different degrees of blockage of single blockage, and there is no misjudgment to multiple blockage. Class 6, Class 7, Class 8, and Class 9 represent a 40 mm blockage and a LC, a 55 mm blockage and a LC of multiple blocked pipes, a 40 mm blockage and a 55 mm blockage and a 40 mm blockage, a 55 mm blockage, and LC respectively. By analyzing the types of multiple blockage misjudgments, it can be seen that the above misjudgments are the misjudgments of multiple blockage categories with different degrees of blockage, and there is no misjudgment to a single blockage.

It can be seen that the recognition rate of comprehensive pipeline operation status can reach 99.56%. Through experimental verification, the improved VMD-IG-RMSE-RF algorithm has superior recognition ability and high diagnosis accuracy for pipeline blockage.

5. Conclusions

Targeting the problem that only a few components from VMD contain useful information for blockage identification, and an information gain-based selection technique is proposed. It is not only effective in selecting the feature components containing substantial blockage information but also plays a crucial role in the deep mining of the information. Sound pressure level conversion is performed on the acoustic pressure signals collected from the standard pipe and the pipe installed with a lateral connection to make a comparison. It is found that the sound pressure level is capable of reflecting the local characteristics of pipeline conditions through the mixture of information while enhancing the discrimination between various operating conditions. The RMS entropy is proven to be responsive to blockage changes from the noise-containing acoustic signals, which can thus be used as input to the classifiers for achieving effective identification.

Given the complexity of pipeline topology and detection environment, as well as the diversity of the pipe defects, further exploration and research are needed, which should cover the following aspects: the sound propagation within pipelines with multiple defects should be explored to discover the pattern thus that a condition prediction model for varying defects can be developed. Advanced active acoustical detection technology requires further study. Such research should involve the development of sensitive acoustic sensors with long-distance availability, selection of valid excitation signal type, wavelength, and frequency, the study of the distribution of sound fields, investigation of the sound attenuation corresponding to various new materials used in civil infrastructures, as well as more advanced data processing techniques.

Author Contributions: Conceptualization, X.Z. and Z.F.; methodology, X.Z.; software, W.D.; validation, X.Z. and Z.F.; formal analysis, X.Z. and W.D.; investigation, X.Z.; resources, J.W.; data curation, Z.F.; writing—original draft preparation, X.Z. and Z.F.; writing—review and editing, J.W.; visualization, X.Z. and W.D.; supervision, J.W.; project administration, Z.F. and J.W.; funding acquisition, Z.F. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China under Grant No. 61563024, 51765022 and 61663017.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data included in this study are all owned by the research group and will not be transmitted.

Acknowledgments: The author would like to thank the financial support provided by the National Nature Science Foundation of China. Thanks to my mentor and team for this research.

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

References

1. Zischg, J.; Rauch, W.; Sitzenfrei, R. Morphogenesis of urban water distribution networks: A spatiotemporal planning approach for cost-efficient and reliable supply. *Entropy* **2018**, *20*, 708. [\[CrossRef\]](#)
2. Fathy, I.; Abdel-Aal, G.M.; Fahmy, M.R.; Fathy, A.; Zelenáková, M. The Negative Impact of Blockage on Storm Water Drainage Network. *Water* **2020**, *12*, 1974. [\[CrossRef\]](#)
3. Tu, M.-C.; Traver, R. Clogging Impacts on Distribution Pipe Delivery of Street Runoff to an Infiltration Bed. *Water* **2018**, *10*, 1045. [\[CrossRef\]](#)
4. Pan, G.; Zheng, Y.; Guo, S.; Lv, Y. Automatic sewer pipe defect semantic segmentation based on improved U-Net. *Autom. Constr.* **2020**, *119*, 103383. [\[CrossRef\]](#)
5. Sen, D.; Aghazadeh, A.; Mousavi, A.; Nagarajaiah, S.; Baraniuk, R.; Dabak, A. Data-driven semi-supervised and supervised learning algorithms for health monitoring of pipes. *Mech. Syst. Signal Process.* **2019**, *131*, 524–537. [\[CrossRef\]](#)

6. Chi, Y.M.; Panda, A.; Byun, G.; Smith, C.F.; Lowe, K.T. Non-intrusive optical measurements of gas turbine engine inlet condensation using machine learning. *Meas. Sci. Technol.* **2021**, *32*, 044001.
7. Datta, S.; Sarkar, S. A review on different pipeline fault detection methods. *J. Loss Prev. Process. Ind.* **2016**, *41*, 97–106. [\[CrossRef\]](#)
8. Guan, R.; Lu, Y.; Duan, W.; Wang, X. Guided waves for damage identification in pipeline structures: A review. *Struct. Control. Health Monit.* **2017**, *24*, e2007. [\[CrossRef\]](#)
9. Wang, M.; Luo, H.; Cheng, J.C. Towards an automated condition assessment framework of underground sewer pipes based on closed-circuit television (CCTV) images. *Tunn. Undergr. Space Technol.* **2021**, *110*, 103840. [\[CrossRef\]](#)
10. Son, B.J.; Cho, T. Modified Crack Detection of Sewer Conduit with Low-Resolution Images. *Appl. Sci.* **2021**, *11*, 2263. [\[CrossRef\]](#)
11. Haurum, J.B.; Moeslund, T.B. A Survey on Image-Based Automation of CCTV and SSET Sewer Inspections. *Autom. Constr.* **2020**, *111*, 103061. [\[CrossRef\]](#)
12. Bin Ali, M.T.; Horoshenkov, K.V.; Tait, S.J. Rapid detection of sewer defects and blockages using acoustic-based instrumentation. *Water Sci. Technol.* **2011**, *64*, 1700–1707. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Khan, M.S. Empirical Modeling of Acoustic Signal Attenuation in Municipal Sewer Pipes for Condition Monitoring Applications. In Proceedings of the 2018 IEEE Green Technologies Conference (GreenTech), Austin, TX, USA, 4–6 April 2018; pp. 137–143.
14. Kim, K.; Wang, S.; Ryu, H.; Lee, S.Q. Acoustic-Based Position Estimation of an Object and a Person Using Active Localization and Sound Field Analysis. *Appl. Sci.* **2020**, *10*, 9090. [\[CrossRef\]](#)
15. Che, T.C.; Duan, H.F.; Lee, P.J. Transient wave-based methods for anomaly detection in fluid pipes: A review. *Mech. Syst. Signal Process.* **2021**, *160*, 107874. [\[CrossRef\]](#)
16. Zeng, W.; Zecchin, A.C.; Gong, J.; Lambert, M.F.; Cazzolato, B.S. Inverse Wave Reflectometry Method for Hydraulic Transient-Based Pipeline Condition Assessment. *J. Hydraul. Eng.* **2020**, *146*, 04020056. [\[CrossRef\]](#)
17. Hawari, A.; Alkadour, F.; Elmasry, M.; Zayed, T. A state of the art review on condition assessment models developed for sewer pipelines. *Eng. Appl. Artif. Intell.* **2020**, *93*, 103721. [\[CrossRef\]](#)
18. Singh, D.; Singh, B. Hybridization of feature selection and feature weighting for high dimensional data. *Appl. Intell.* **2018**, *49*, 1580–1596. [\[CrossRef\]](#)
19. Bayat, M.; Ahmadi, H.R.; Mahdavi, N. Application of power spectral density function for damage diagnosis of bridge piers. *Struct. Eng. Mech.* **2019**, *71*, 57–63.
20. Ahmadi, H.R.; Mahdavi, N.; Bayat, M. A novel damage identification method based on short time Fourier transform and a new efficient index. *Structures* **2021**, *33*, 3605–3614. [\[CrossRef\]](#)
21. Xu, C.; Du, S.; Gong, P.; Li, Z.; Chen, G.; Song, G. An Improved Method for Pipeline Leakage Localization With a Single Sensor Based on Modal Acoustic Emission and Empirical Mode Decomposition With Hilbert Transform. *IEEE Sens. J.* **2020**, *20*, 5480–5491. [\[CrossRef\]](#)
22. Sun, J.; Peng, Z.; Wen, J. Leakage Aperture Recognition based on Ensemble Local Mean Decomposition and Sparse Representation for Classification of Natural Gas Pipeline. *Measurement* **2017**, *108*, 91–100. [\[CrossRef\]](#)
23. Jiang, L.; Ma, Z.; Zhang, J.; Khan, M.Y.A.; Cheng, M.; Wang, L. Chaotic Characteristic Analysis of Vibration Response of Pumping Station Pipeline Using Improved Variational Mode Decomposition Method. *Appl. Sci.* **2021**, *11*, 8864. [\[CrossRef\]](#)
24. Fu, L.; Zhu, T.; Pan, G.; Chen, S.; Wei, Y. Power Quality Disturbance Recognition Using VMD-Based Feature Extraction and Heuristic Feature Selection. *Appl. Sci.* **2019**, *9*, 4901. [\[CrossRef\]](#)
25. Sun, J.; Xiao, Q.; Wen, J.; Zhang, Y. Natural gas pipeline leak aperture identification and location based on local mean decomposition analysis. *Measurement* **2016**, *79*, 147–157. [\[CrossRef\]](#)
26. Lee, P.J.; Duan, H.; Tuck, J.; Ghidaoui, M. Numerical and Experimental Study on the Effect of Signal Bandwidth on Pipe Assessment Using Fluid Transients. *J. Hydraul. Eng.* **2015**, *141*, 04014074. [\[CrossRef\]](#)
27. Wenxuan, W.; Zhijian, W.; Jiping, Z.; Weijin, M.; Junyuan, W. Research of the Method of Determining k Value in VMD based on Kurtosis. *J. Mech. Transm.* **2018**, *42*, 153–160.
28. Wang, B.; Li, Y.; Zhao, W.; Zhang, Z.; Zhang, Y.; Wang, Z. Effective Crack Damage Detection Using Multilayer Sparse Feature Representation and Incremental Extreme Learning Machine. *Appl. Sci.* **2019**, *9*, 614. [\[CrossRef\]](#)
29. Azhagusundari, B.; Thanamani, A.S. Feature selection based on information gain. *Int. J. Innov. Technol. Explor. Eng. (IJITEE)* **2013**, *2*, 18–21.
30. Deng, H.; Diao, Y.; Wu, W.; Zhang, J.; Zhong, X. A high-speed D-CART online fault diagnosis algorithm for rotor systems. *Appl. Intell.* **2019**, *50*, 29–41. [\[CrossRef\]](#)
31. Dragomiretskiy, K.; Zosso, D. Variational Mode Decomposition. *IEEE Trans. Signal Process.* **2014**, *62*, 531–544. [\[CrossRef\]](#)
32. Shi, H.; Fu, W.; Li, B.; Shao, K.; Yang, D. Intelligent Fault Identification for Rolling Bearings Fusing Average Refined Composite Multiscale Dispersion Entropy-Assisted Feature Extraction and SVM with Multi-Strategy Enhanced Swarm Optimization. *Entropy* **2021**, *23*, 527. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Kent, J.T. Information gain and a general measure of correlation. *Biometrika* **1983**, *70*, 163–173. [\[CrossRef\]](#)
34. Younes, R.; Ouelaa, N.; Hamzaoui, N.; Djamaa, M.C.; Djebala, A. The Influence of the Sound Pressure Level on the Identification of the Defects Severity in Gear Transmission by the Sound Perception. *Acoust. Aust.* **2019**, *47*, 239–246. [\[CrossRef\]](#)
35. Santo, F.T.; Sattar, T.P.; Edwards, G. Validation of Acoustic Emission Waveform Entropy as a Damage Identification Feature. *Appl. Sci.* **2019**, *9*, 4070. [\[CrossRef\]](#)

-
36. Sharma, V.; Parey, A. Gearbox fault diagnosis using RMS based probability density function and entropy measures for fluctuating speed conditions. *Struct. Health Monit.* **2017**, *16*, 682–695. [[CrossRef](#)]
 37. Feng, Z. Condition Classification in Underground Pipes Based on Acoustical Characteristics. Ph.D. Thesis, University of Bradford, Bradford, UK, 2013.
 38. Zhu, X.; Huang, G.; Feng, Z.; Wu, J. Condition Classification of Water-Filled Underground Siphon Using Acoustic Sensors. *Sensors* **2019**, *20*, 186. [[CrossRef](#)]