

## Article

# Fault Classification of Disconnect Switches Based on Multi-Scale RGB Image Representation and SE-CNN-Attention

Xiaofei Kang <sup>1</sup>, Su Xu <sup>1</sup>, Yuqi Liu <sup>1</sup>, Jianguo Du <sup>1</sup>, Chunqiao Fan <sup>1</sup>, Jie Hou <sup>1</sup>, Shuaidong Zhang <sup>1</sup>  
and Jingang Wang <sup>2,\*</sup> 

<sup>1</sup> Baotou Power Supply Branch of Inner Mongolia Electric Power (Group) Co., Ltd., Baotou 014030, China; kangxiaofei0329@163.com (X.K.); xusu0802@163.com (S.X.); yuqiliu0854@proton.me (Y.L.); jianguodu0447@proton.me (J.D.); chunqiaofan665@proton.me (C.F.); jiehou621@proton.me (J.H.); shuaidongzhang558@proton.me (S.Z.)

<sup>2</sup> State Key Laboratory of Power Transmission Equipment Technology, School of Electrical Engineering, Chongqing University, Chongqing 400044, China

\* Correspondence: jingang@cqu.edu.cn

## Abstract

To address the challenges of analyzing vibration signals, overlapping fault characteristic frequency bands, and difficulties in accurately distinguishing similar faults, this study proposes a diagnostic classification scheme for GIS disconnect switch mechanical fault detection. The scheme integrates multi-scale RGB image representation with a Squeeze-and-Excitation Convolutional Neural Network Attention (SE-CNN-Attention) mechanism. First, continuous wavelet transform is employed to decouple the one-dimensional vibration signal into low-, medium-, and high-frequency physical bands, which are then mapped to the red, green, and blue channels to generate a  $64 \times 64 \times 3$  RGB time–frequency feature image, enabling three-dimensional encoding of time, frequency, and energy intensity. Based on this approach, a SE-CNN-Attention fusion model is developed. The two-dimensional CNN automatically extracts region-specific features from the time–frequency images. Concurrently, the squeeze-and-excitation (SE) attention mechanism adaptively assigns weights to enhance fault-sensitive frequency bands and suppress noise interference. The model adopts global average pooling instead of traditional fully connected layers, combined with Dropout regularization and a Softmax classifier, to achieve efficient and robust fault classification. Experiments are conducted under four operating conditions with a total of 19,800 sample groups collected. The findings indicate that the proposed approach attains an overall classification accuracy exceeding 94%, outperforming traditional support vector machine (SVM), baseline convolutional neural network (CNN), Vision Transformer (ViT), and ResNet-18 models. The SE-Attention module significantly enhances the model's focus on key frequency band features while accelerating convergence speed and improving classification stability. This method provides reliable technical support for intelligent maintenance and fault warning of GIS disconnect switch equipment and holds significant engineering application value.

**Keywords:** convolutional neural network; SE-Attention mechanism; fault classification; vibration signal analysis



Academic Editor: George A. Tsihrintzis

Received: 1 June 2026

Revised: 25 June 2026

Accepted: 2 July 2026

Published: 28 July 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

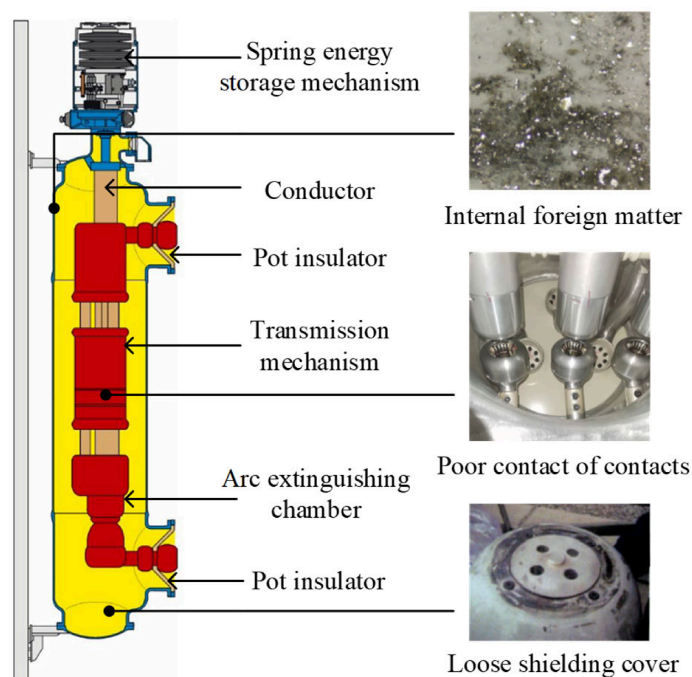
This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

## 1. Introduction

Gas-Insulated Switchgear (GIS) integrates primary equipment—including current transformers, voltage transformers, grounding switches, disconnect switches, busbars, and

surge arresters, but excluding transformers—into a metal enclosure filled with SF<sub>6</sub> gas, which serves as both an arc-extinguishing and insulating medium [1]. Owing to their compact structure, small footprint, high reliability, and long maintenance intervals, such devices have been widely adopted in power systems worldwide since the last century. However, after long-term operation, the failure rate of GIS gradually increases. Statistical data indicate that mechanical faults account for approximately 39.3% of GIS failures, and partial discharge faults often result from the accumulation of mechanical defects. Therefore, research on mechanical fault diagnosis of GIS disconnect switches is critically important to guarantee the safety and stability of electricity networks [2].

The GIS disconnect switch mainly consists of stationary arc contacts, nozzles, moving arc contacts, insulating rods, etc., as illustrated in Figure 1. Its basic function is to interrupt the circuit, with contacts opening and closing in an SF<sub>6</sub> atmosphere. Leveraging the excellent insulation and arc-extinguishing properties of SF<sub>6</sub>, the switch can safely interrupt high-voltage currents within a very short time [3]. The connection between the moving and stationary contacts is a critical part of the internal electrical joints. Typical mechanical faults fall into three main categories: poor contact of contacts, loosening of shielding covers, and internal foreign objects [4]. Poor contact usually occurs at the junction between moving and stationary contacts, caused by insufficient insertion depth, spring slack, or contact surface erosion and oxidation [5]. This leads to reduced contact pressure, increased contact resistance, localized heating, and intermittent arcing. Shielding cover looseness is often observed at the fixing points of the shielding cover on the high-voltage conductor, arising from bolt loosening, weld cracking, or long-term vibration-induced structural instability. It manifests as power-frequency vibration superimposed with high-frequency impacts, significantly increasing the risk of partial discharge. Internal foreign objects are distributed at the bottom of the gas chamber, on insulator surfaces, or in gaps of shielding covers. Their sources include residual metal shavings from installation, wear particles from operating mechanisms, and detached fragments. Under the influence of an electric field, free particles may jump, generating random impact signals.



**Figure 1.** Structure diagram of GIS disconnect switch.

GIS equipment encloses primary electrical devices inside a metal housing, making it difficult to directly sense the internal state from the outside [6]. Once a fault occurs, prolonged power outages are often required for maintenance, with high difficulty and cost, consuming substantial manpower and resources. Currently, scheduled outage maintenance remains the dominant practice on-site, which often leads to either over-maintenance or under-maintenance [7], failing to fully exploit equipment operational efficiency.

GIS disconnect switches exhibit specific vibration patterns under different operating conditions, with differences in time-domain, frequency-domain, and spectral characteristics [8]. Fault identification can be achieved by extracting vibration signal features, making vibration analysis an important tool for electrical equipment fault diagnosis [9]. Early studies on GIS mechanical fault diagnosis have investigated the frequency characteristics of vibration signals under different fault conditions, initially revealing the relationship between vibration features and fault states [10]. Previous studies have found that the mechanical vibration of the GIS enclosure is predominantly at a fundamental frequency of 100 Hz. When mechanical defects such as poor contact exist, the amplitudes of the 200 Hz and 300 Hz components increase significantly. Vibration signal measurement is immune to electromagnetic interference, and vibration-based identification of GIS mechanical states has achieved good results. However, the physical meanings of features extracted by different methods vary, and single-dimensional feature extraction methods often lack specificity, making it difficult to effectively characterize mechanical defects and potentially leading to missed or false diagnoses. Composite feature extraction methods combining time–frequency features and data sequence features have been proposed, achieving significantly better diagnostic accuracy than conventional single methods. Nevertheless, due to the complex structure of GIS and the diversity of defect types, achieving unified and reliable mechanical fault diagnosis remains challenging.

In the realm of intelligent fault diagnosis, support vector machines, ensemble learning, and deep belief networks have been explored for GIS equipment fault identification [10,11]. The second-order synchrosqueezing Fourier transform has been adopted to obtain time–frequency distributions of vibration signals and build a diagnostic model combined with deep transfer learning, validating the feasibility of integrating time–frequency transformation with deep learning. However, FSST2 is essentially an improvement on the short-time Fourier transform, and its time–frequency resolution remains limited by the fixed window length. Clustering contrastive learning frameworks with unsupervised domain adaptation have been proposed to enhance cross-scenario diagnostic robustness. In six cross-domain tasks, the average accuracy improved by 8.29% to 19.3% compared with six classical methods. Although this method performs excellently in GIS insulation defect diagnosis, it relies on partial discharge UHF signals rather than vibration signals and does not convert 1D signals into 2D representations for 2D-CNN processing. Some studies have attempted to convert 1D vibration signals from GIS mechanical defects into 2D image form using GAF and then used lightweight networks such as SqueezeNet for classification, thereby avoiding manual feature extraction and achieving diagnosis of multiple types of defects. However, GAF has limited time–frequency resolution, and SqueezeNet has relatively weak feature extraction capability, making it difficult to handle complex fault patterns.

Although significant progress has been made in the above studies, several common limitations remain: First, time–frequency analysis methods are mostly limited to fixed-resolution transformations and cannot simultaneously achieve high-frequency resolution at low frequencies and high time resolution at high frequencies. Second, intelligent models often adopt one-dimensional convolutions, lightweight networks, or traditional machine learning methods, which have limited receptive fields and insufficient feature extraction capability. Third, the signal sources are mostly UHF, VFTO, or acoustic signals, which

differ from conventional vibration monitoring scenarios. To address these issues, CNNs, with their local connectivity and weight-sharing properties, have demonstrated significant advantages in image recognition and fault diagnosis. Since the breakthrough of AlexNet in 2012, CNNs have become a core technology for tasks such as image recognition, detection, classification, and segmentation [12]. CNNs automatically extract image features through convolution operations, typically consisting of conv layers, pooling layers, and FC layers. Traditional machine learning methods rely on complex signal processing and analysis, require substantial manual effort, and mostly use shallow networks with limited ability to learn complex nonlinear relationships. In contrast, deep learning can automatically extract deep features from raw data and perform prediction and classification, and has gained widespread attention in areas such as analog circuit fault diagnosis [13].

Among one-dimensional sequential deep learning methods, deep belief networks suffer from large parameter sizes and high modeling costs due to their fully connected structures. Stacked autoencoders, although capable of fast diagnosis, have feature expression mechanisms that are difficult to interpret. One-dimensional convolutional neural networks have convolution kernels that slide only in one direction, resulting in a limited receptive field and potential information loss. The GIS-SegFormer model validated the potential of the Transformer architecture for GIS condition monitoring but also pointed out that the self-attention mechanism tends to neglect local features. To overcome these shortcomings, this paper proposes converting the original one-dimensional time-domain signal into a two-dimensional feature map, thereby leveraging the widely used two-dimensional convolutional neural networks from image processing [14]. In disconnect switch fault diagnosis, converting 1D oscillatory signals into 2D CNNs based on time-series acceleration signals yields a visual representation with feature enhancement capabilities, providing high-quality input data for subsequent deep learning models.

Recent studies have demonstrated the effectiveness of attention mechanisms in enhancing the feature representation capability of deep learning models for fault diagnosis tasks. Ye et al. [15] proposed a multiscale attentive convolutional neural network with squeeze-and-excitation attention for small-sample fault diagnosis, achieving superior diagnostic accuracy through adaptive channel-wise feature recalibration. Li et al. [16] developed a hybrid attention mechanism combined with improved convolutional neural network layers, demonstrating that attention mechanisms can effectively enhance the model's focus on fault-relevant features. Mahmoud et al. [17] designed a wavelet-enhanced multi-scale CNN with SE attention specifically for noise-resilient fault diagnosis, where the SE module adaptively recalibrates channel-wise feature responses to suppress noise interference while enhancing fault-sensitive components. These works collectively highlight the potential of SE attention for adaptive feature weighting in fault diagnosis applications, which motivates the introduction of the SE-Attention module in our framework.

The SE-Attention mechanism is introduced to selectively enhance important feature channels and suppress redundant information [18], effectively mitigating the interference of non-sensitive features during deep neural network training and thereby fully exploiting inter-feature correlations. The CNN is responsible for efficient local spatial feature extraction, while the attention mechanism enhances the influence of key feature factors on fault state discrimination. The advantage of the SE-CNN-Attention model lies in its ability to leverage the local feature extraction strengths of CNNs while employing the SE-Attention mechanism to adaptively recalibrate channel-wise feature responses [19], further improving the model's expressive power for complex fault features.

Compared with recent CNN-attention-based fault diagnosis methods, the novelty of the proposed framework is twofold. First, most existing approaches apply attention modules directly to single-channel time–frequency representations or raw signals, whereas

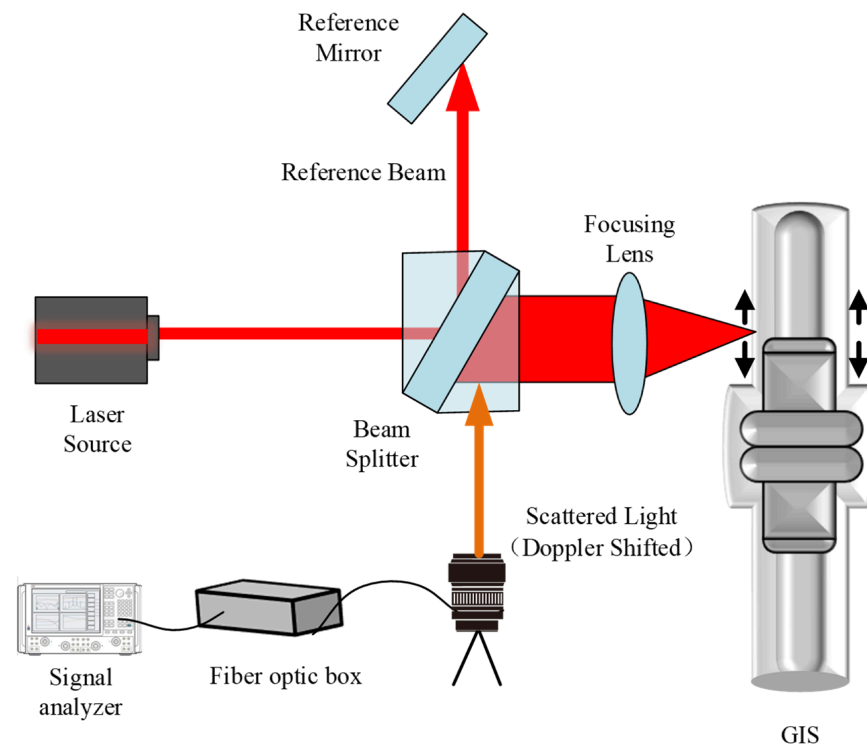
our method constructs a physically meaningful multi-scale RGB image that explicitly encodes low-, medium-, and high-frequency bands corresponding to distinct mechanical fault mechanisms. This frequency-band-specific decomposition provides the subsequent SE-Attention module with more discriminative input features than conventional single-scale or single-channel representations. Second, unlike generic applications of SE-Attention that treat channel recalibration as a purely data-driven process, our framework aligns the learned attention weights with the physical frequency-band allocation. This physical interpretability, validated by the SE weight distribution analysis, distinguishes our approach from purely data-driven attention mechanisms and provides insights into the model's decision-making process.

In summary, to achieve accurate diagnosis of three fault types—poor contact of disconnect switch contacts, internal foreign objects, and shielding cover looseness—this paper presents a fault detection approach leveraging multi-scale RGB image representation and SE-CNN-Attention fusion. This approach effectively overcomes the bottlenecks of conventional time–frequency analysis in feature expression and anti-interference capability [20]. Through multi-level, multi-perspective time–frequency information fusion and the synergistic effect of a deep attention mechanism, it achieves high-precision and high-stability mechanical fault diagnosis for GIS disconnect switches, providing a new technical pathway and theoretical support for the construction of intelligent maintenance systems for disconnect switches.

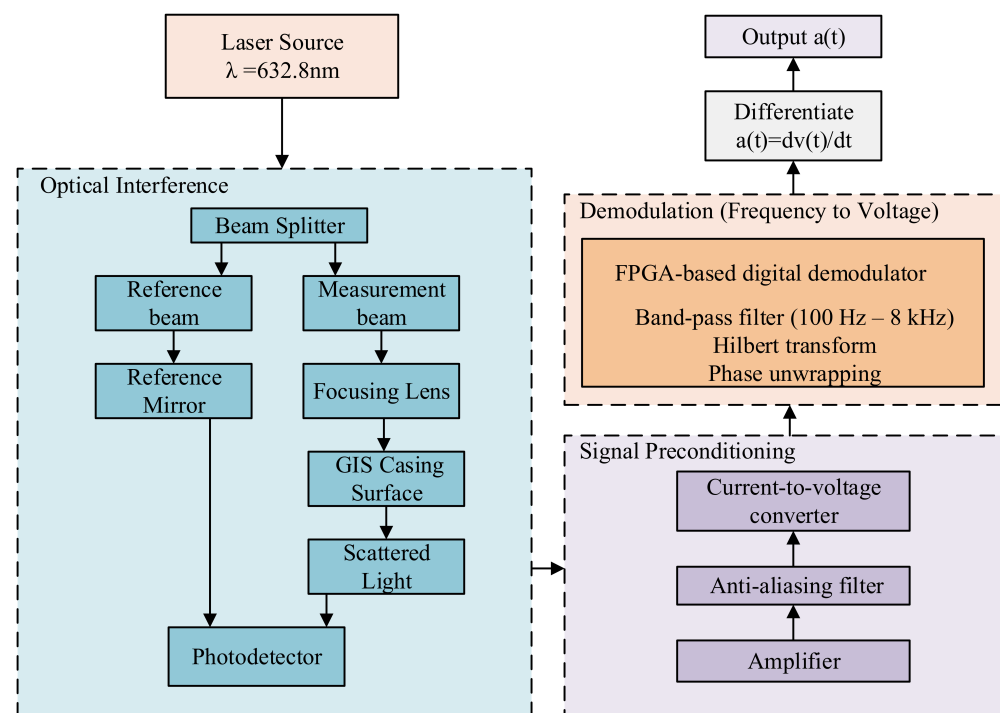
## 2. Fault Classification of Disconnect Switches Based on Multi-Scale RGB Image Representation and SE-CNN-Attention Fusion

### 2.1. Signal Acquisition

In this study, a non-contact photoelectric measurement technique, specifically laser Doppler vibrometry (LDV), is employed to acquire mechanical vibration signals from GIS disconnect switches. Photoelectric measurement technology integrates infrared sensing, image processing, laser vibrometry, microelectronic control, and automation techniques, offering strong environmental adaptability, fast response, and excellent immunity to electromagnetic interference [21]. Given the enclosed structure of GIS equipment and the difficulty of placing contact-based sensors in a high-voltage operating environment, non-contact LDV is selected to capture mechanical vibration signals from the surface of the disconnect switch enclosure. As shown in Figure 2, the laser beam emitted from the source is split by a beam splitter into a reference beam and a measurement beam. The reference beam is directed to a reference mirror and returns along the same path, while the measurement beam is focused onto the GIS enclosure surface. When the enclosure surface experiences small displacements due to internal mechanical vibrations, the frequency of the reflected measurement beam shifts due to the Doppler effect, generating scattered light that carries vibration information [22]. This scattered light is combined with the returned reference beam through a fiber optic coupler and sent to a photodetector and signal analyzer. After demodulation, the vibration velocity and acceleration signals on the enclosure surface are obtained. The corresponding signal processing workflow is illustrated in Figure 3. The heterodyne signal is conditioned, band-pass filtered, and digitally demodulated to obtain the vibration velocity  $v(t)$ , which can be further differentiated to yield the acceleration  $a(t)$ ; both signals are sampled at 20 kHz for subsequent processing.



**Figure 2.** GIS enclosure vibration signal acquisition system.

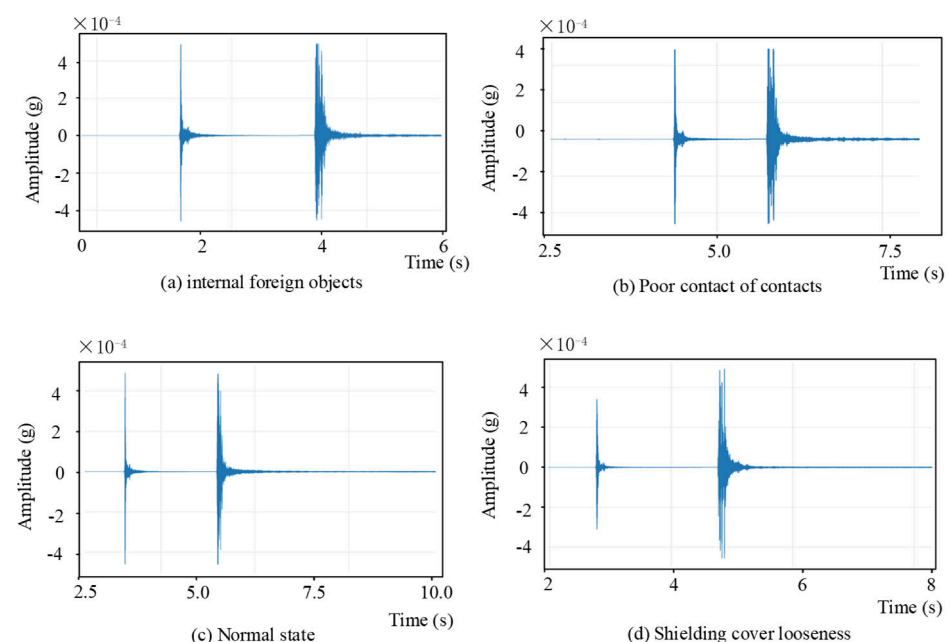


**Figure 3.** Signal processing flowchart.

All experiments were conducted on a 220 kV GIS disconnect switch test platform within a single coordinated experimental campaign from May to June 2024, lasting approximately three weeks. A Polytec laser Doppler vibrometer (OFV-5000 controller with OFV-505 sensor head; Polytec GmbH, Waldbronn, Germany) was used for non-contact vibration measurement. Consistent sensor placement, a fixed sampling rate of 20 kHz, and identical operational protocols were maintained throughout the data acquisition process to ensure experimental repeatability.

GIS structures are large and complex, and the locations of mechanical faults exhibit a certain degree of randomness. Although the impact vibration signals generated by disconnect switch operations are of relatively high frequency, the excitation energy attenuates with increasing transmission distance [23], and vibration signals propagate efficiently only when their frequency matches the intrinsic modal frequency of the equipment. The above characteristics determine the vibration signal acquisition strategy: sensors should be placed as close as possible to the operating mechanism or key components to reduce the impact of energy attenuation on signal strength; meanwhile, the sampling frequency does not need to be excessively high—covering the dominant modal frequencies of the GIS enclosure is sufficient to effectively capture fault features. Based on this, this paper selects a sampling frequency of 20 kHz, which satisfies the Nyquist theorem while balancing engineering economy.

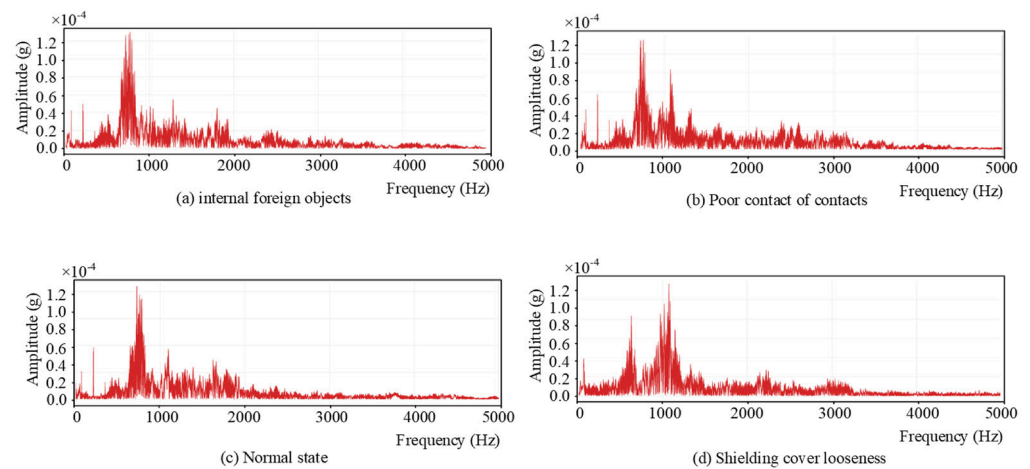
As shown in Figure 4, the raw vibration signals of the four operating states provide limited discriminative information. The normal state Figure 4c exhibits regular low-amplitude oscillations, while internal foreign objects Figure 4a show irregular low-amplitude fluctuations without obvious impact features. Poor contact Figure 4b features occasional high-amplitude impulsive spikes; however, their occurrence is intermittent and varies significantly in amplitude and interval, making time-domain indicators unreliable. Shielding cover looseness Figure 4d presents high-frequency oscillations superimposed on a low-frequency envelope, yet these features are easily masked by background noise. Overall, relying solely on the original time-domain waveforms is insufficient for reliable fault diagnosis, because the characteristics do not offer consistently distinguishable cues and are highly susceptible to environmental interference. Therefore, more informative representations are required to capture the intrinsic fault signatures.



**Figure 4.** Raw time-domain vibration signals of the four operating states.

The original time-domain waveforms shown in Figure 4 provide only limited information for distinguishing the four operating states. To analyze the frequency characteristics, spectral analysis is employed. By applying the fast Fourier transform or continuous wavelet transform, the one-dimensional vibration signal is converted into a frequency-domain representation, where the horizontal axis represents frequency and the vertical axis represents amplitude (or energy magnitude). Such a representation reveals the energy distribution across different frequency bands, offering more discriminative features than the raw wave-

form alone [5]. Figure 5 shows the frequency spectra of the four states, from which more distinct patterns can be observed.



**Figure 5.** Frequency spectra of the vibration signals under the four operating states.

### 2.1.1. Internal Foreign Objects

Foreign object sources primarily include metal debris left during installation, wear particles generated by operating mechanisms, and intentionally placed short metal wires. When metallic foreign objects exist inside the GIS gas chamber, the vibration impact during opening and closing operations can cause the objects to jump, significantly enhancing partial discharge signals [24]. Conductive micron-scale metal particles are placed on the shielding cover surface to simulate random impact signals under the influence of an electric field.

As shown in Figure 5a, the energy of internal foreign objects is mainly distributed below 1000 Hz within the overall 0–5000 Hz range. The spectral structure is highly similar to that of the normal state, with good energy concentration. However, slight energy dispersion or weak components above 1000 Hz may appear in certain local frequency bands. This is because, although internal foreign objects may jump randomly within the gas chamber, the vibration energy is limited by the size and motion trajectory of the foreign objects, making it difficult to excite sustained high-frequency components. Consequently, the spectral characteristics of internal foreign objects are highly confusable with those of the normal state, and it is difficult to distinguish them based solely on spectral morphology. Due to the lack of obvious impact or broadband features extending above 2000 Hz, internal foreign objects are easily confused with the normal state.

### 2.1.2. Poor Contact of Disconnect Switch Contacts

During the design of GIS disconnect switch moving contacts, a certain overtravel is provided between the moving contact and the contact fingers to further ensure good contact. Therefore, there is a significant nonlinear relationship between the closing travel of the moving contact and the contact state. The poor contact condition is simulated by manually adjusting the closing travel of the moving contact to change the contact state.

As shown in Figure 5b, when poor contact occurs, energy is distributed across the entire 0–5000 Hz range, but is mainly concentrated below 2000 Hz. Compared with the normal state, the energy distribution is broader, accompanied by intermittent instantaneous impact characteristics. This is because poor contact leads to increased contact resistance and reduced contact pressure, resulting in unstable electrical contact between the moving and stationary contacts during closing, which in turn generates partial discharge and micro-arc, exciting high-frequency vibration components. The energy distribution within the

2000–5000 Hz band reflects the intensity and frequency of impact signals and is the most distinguishing feature from other states. Therefore, poor contact can be identified by the richness of high-frequency components and impact characteristics.

### 2.1.3. Normal State

As shown in Figure 5c, under normal operating conditions, the energy of the disconnect switch is mainly concentrated in the low-frequency band, particularly near the dominant frequency of approximately 1000 Hz, with harmonic components showing a clear decaying trend. The spectral structure is clear, and the energy distribution is concentrated and regular, without significant components above 1000 Hz or impact characteristics. This indicates that during normal operation, the motion of the operating mechanism is smooth, and no abnormal impact or friction occurs during the contact closure process, with vibration energy mainly originating from normal movements of mechanism components. The spectral characteristics of the normal state serve as a reference baseline for other fault states.

### 2.1.4. Loosening of Conductor Shielding Cover

The loosening defect of the conductor shielding cover is simulated by loosening the fixing bolts on the moving-side or stationary-side shielding cover with a uniform torque, mimicking the decline in fastening force after long-term operational vibration. When shielding cover loosening occurs inside the GIS, the dominant frequency of the vibration signal shifts.

As shown in Figure 5d, under the shielding cover looseness state, energy is relatively densely distributed around the 2000 Hz frequency band within the 0–5000 Hz range, shifting toward the mid-to-high-frequency range compared with the normal state. This is because shielding cover looseness leads to unbalanced forces on the three-phase conductors, generating higher harmonics under the alternating electric field, thereby exciting additional vibration components in the mid-to-high-frequency range. The amplitude of components in the 2000–5000 Hz range increases significantly, but the energy distribution remains relatively concentrated, without the widespread distribution observed in poor contact. Therefore, shielding cover looseness can be identified by the upward shift in energy concentration toward the mid-to-high-frequency range and the increase in harmonic components.

Synthesizing the spectral analyses presented in Figure 5, the spectral characteristics of the four operating states are summarized in Table 1. However, the above spectral analysis only provides a preliminary estimation of the operating state and cannot accurately distinguish between similar faults, especially the normal state and internal foreign objects.

**Table 1.** Time–frequency domain characteristics.

| GIS Disconnect Switch State | Energy Concentration Band | Typical Time–Frequency Characteristics                               |
|-----------------------------|---------------------------|----------------------------------------------------------------------|
| Internal foreign objects    | Approx. 1000 Hz           | Highly similar to normal state; weak local high-frequency components |
| Poor contact of contacts    | Within 2000 Hz            | Broadband distribution; intermittent instantaneous impacts           |
| Normal state                | Approx. 1000 Hz           | Clear spectrum; sequentially decaying harmonics; no impacts          |
| Shielding cover looseness   | Around 2000 Hz            | Upward energy shift; significant high-order harmonics                |

## 2.2. Data Preprocessing

Vibration signals are inherently non-stationary transient signals, and their fault characteristics manifest as localized energy bursts in both the time and frequency domains. Traditional Fourier analysis can only provide global frequency domain information and cannot capture the temporal evolution of features [25]. Although the short-time Fourier transform can obtain time–frequency distributions, its time–frequency resolution is constrained by the fixed length of the window function, making it difficult to simultaneously achieve high-frequency resolution in the low-frequency band and high time resolution in the high-frequency band. In contrast, continuous wavelet transform employs scalable basis functions, automatically achieving high-frequency resolution in the low-frequency band and high time resolution in the high-frequency band, making it more suitable for analyzing the transient impact signals generated during disconnect switch closing operations [26]. Furthermore, although the Hilbert–Huang transform has adaptive decomposition capabilities, its empirical mode decomposition process is prone to mode mixing, and the lack of a unified basis function hinders consistency in batch signal processing. Based on a comprehensive comparison, this study selects CWT as the time–frequency analysis tool.

Although the discrete wavelet transform (DWT) is computationally more efficient and widely used for standard sub-band energy extraction in industrial fault diagnosis, it is not optimal for the proposed visual feature extraction pipeline. As a non-redundant transform implemented via a filter bank structure, the DWT produces a compact set of coefficients with a fixed grid-like representation, which is primarily suitable for signal compression and denoising rather than for generating high-resolution, smooth time–frequency images for visual analysis. In contrast, the CWT is a redundant transform that can operate on any scale, producing an infinitely redundant, translation-invariant, and visually rich scalogram. This translation invariance ensures that time–frequency features remain consistent regardless of when the transient impact occurs within the sliding window, which is critical for training a robust 2D CNN that does not overfit to specific temporal positions. The ability of the CWT to simultaneously capture both steady-state and transient behavior in a continuous and detailed manner makes it particularly robust when paired with a 2D convolutional neural network, as it provides the deep learning model with a physically meaningful and visually interpretable input representation. This high-resolution time–frequency representation is strictly necessary for the subsequent multi-scale RGB image encoding and SE-CNN-Attention feature extraction, where the detailed texture and energy distribution patterns are critical for distinguishing overlapping fault characteristics.

The sliding window method segments the time series using a fixed-length window with a fixed step size. This approach allows the extraction of statistical information—such as the mean and variance—from each windowed data segment. On one hand, this approach divides long-sequence signals into multiple local segments, effectively augmenting the sample size. On the other hand, the overlapping design ensures that no matter where a fault feature appears in the signal [27], it is fully covered by at least one window, preventing feature loss. This method has been widely applied in fault diagnosis of rotating machinery and high-voltage switchgear.

A complete closing operation generates a vibration signal that can be up to several hundred milliseconds in length, but significant fault information is present only during certain periods, such as the moment of contact impact and key stages of mechanism operation. The modal frequencies of GIS disconnect switches are primarily concentrated within 5 kHz. According to the Nyquist sampling theorem, the sampling frequency must exceed 10 kHz. Considering the transition band margin of anti-aliasing filters, 20 kHz is chosen as the ultimate sampling rate. A window size of 1024 samples is equivalent to a duration of 51.2 ms, which can cover multiple vibration cycles during the disconnect

switch closing process. A stride of 512 samples is a common setting for sliding windows, ensuring a sufficient number of samples while avoiding abrupt feature changes between adjacent windows.

The duration of 51.2 ms corresponds to approximately 5.12 fundamental cycles at the dominant 100 Hz frequency of GIS enclosure vibration, which is sufficient to encompass the entire energy decay process of mechanical impacts induced by contact closure and structural faults. A shorter window would risk truncating the transient envelope and losing critical fault-related information, whereas a longer window would introduce excessive steady-state background noise and stationary harmonic components, potentially diluting the transient signatures that are most discriminative for fault classification. The 50% overlap between consecutive windows is primarily adopted to mitigate the edge effects inherent to the CWT, which may otherwise distort time–frequency features near window boundaries; as a secondary benefit, it also serves as a data augmentation strategy to enrich the training set without altering the underlying signal characteristics. This parameter configuration ensures that the resulting scalograms preserve both the temporal localization and the spectral integrity of the vibration transients, providing reliable input representations for the subsequent deep learning model.

The low-frequency band primarily reflects the motion of large-mass components of the disconnect switch operating mechanism, foundation looseness, and overall enclosure swinging. The medium-frequency band corresponds to spring fatigue, shielding cover looseness, and certain harmonic resonances. The high-frequency band is associated with instantaneous impacts caused by poor contact, arc discharge, and metal particle collisions. By mapping the three frequency bands to the three RGB channels, the deep learning model can use color channels as cues to automatically learn the differences in frequency band sensitivity across fault types, thereby improving the interpretability of the classification.

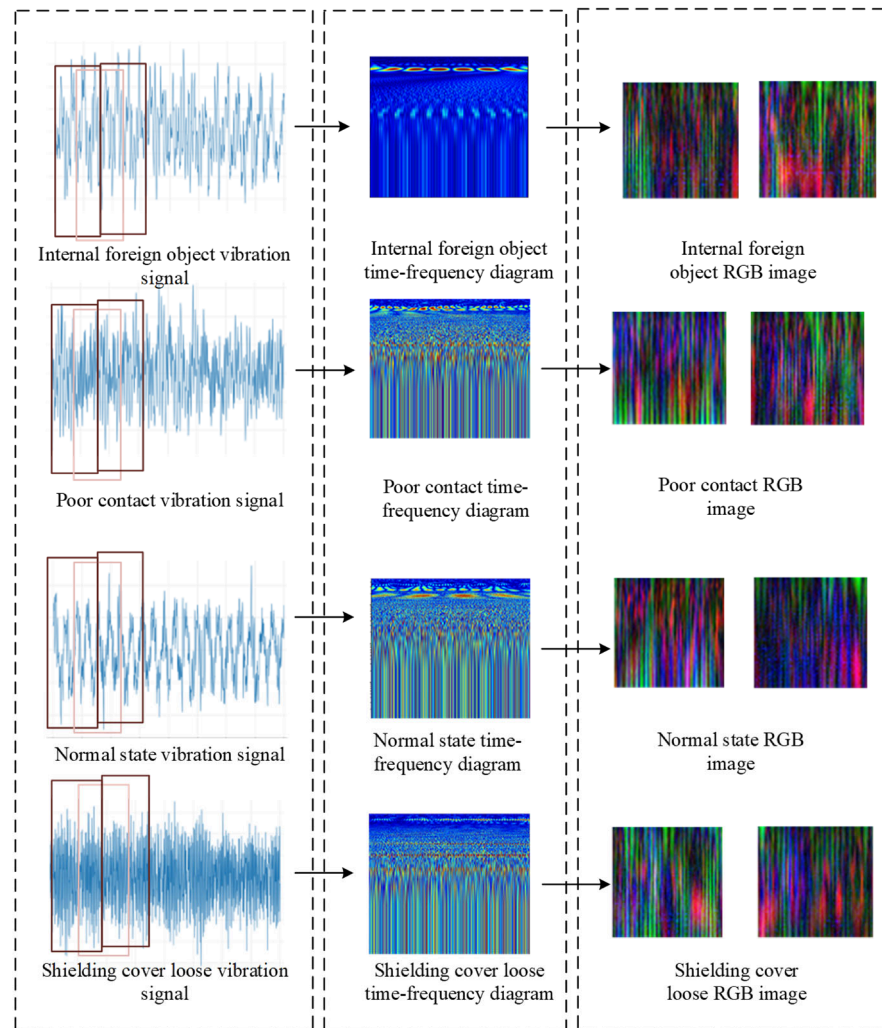
Under normal GIS conditions, energy is mainly concentrated near the dominant frequency of 1000 Hz, with harmonics decaying sequentially. For internal foreign objects, energy is also concentrated within the 1000 Hz band, similar to the normal state. Therefore, frequencies below 1000 Hz are designated as the low-frequency band, fully covering the dominant energy of the normal state and internal foreign objects.

Under shielding cover looseness, energy is relatively densely distributed in the 2000 Hz band, shifting toward higher frequencies compared with the normal state. The vibration spectral distribution of mechanical defects is generally within 4 kHz. Hence, the frequency band from 1000 to 4000 Hz is designated as the medium-frequency band to capture harmonics and resonance features caused by faults such as shielding cover looseness.

Under poor contact conditions, energy extends to 5000 Hz, accompanied by intermittent instantaneous impact characteristics. The energy distribution in the high-frequency band reflects the intensity and frequency of impact signals caused by poor contact. Therefore, the frequency band from 4000 to 5000 Hz is designated as the high-frequency band, focusing on high-frequency impacts and partial discharge features associated with poor contact.

The Morlet wavelet, due to its good localization properties in both the time and frequency domains and its waveform similarity to impact signals, is widely used for time–frequency analysis of mechanical vibration signals [28]. Based on the frequency analysis range and frequency resolution requirements, the number of scales is set to 128 to ensure sufficient frequency resolution in the low-frequency band. The CWT coefficients are divided into three physical frequency bands along the frequency axis: low-frequency band (0–1000 Hz), medium-frequency band (1000–4000 Hz), and high-frequency band (4000–5000 Hz). For each frequency band, the coefficient matrix is converted to its modulus and independently normalized to the [0, 1] interval, and then mapped to the red, green, and

blue channels, respectively. Finally, the three-channel matrices are uniformly scaled to  $64 \times 64$  pixels using bilinear interpolation, synthesizing an RGB three-channel time–frequency image. In this image, the horizontal axis represents temporal evolution, the vertical axis represents frequency distribution, and the color intensity reflects vibration energy. Through the above transformation, the vibration signal of a single closing operation is encoded into a  $64 \times 64 \times 3$  color image, serving as the input sample for the subsequent SE-CNN-Attention model. Figure 6 illustrates the process of converting vibration signals into images.



**Figure 6.** Sample segmentation and composition of three-band vibration signals into an image.

The frequency band boundaries are not empirically tuned but are physically determined by the distinct spectral distributions of the four operating states. As demonstrated in Figure 5 and summarized in Table 1, the normal state and internal foreign objects exhibit energy primarily concentrated near 1000 Hz, shielding cover looseness shows dominant energy shifting toward approximately 2000 Hz, and poor contact presents broadband features extending up to 5000 Hz. These well-separated spectral signatures provide a clear physical basis for defining the three frequency bands at 1000 Hz and 4000 Hz, with a 1000 Hz margin preserved between adjacent bands to accommodate individual variations. Consequently, the classification performance is inherently insensitive to minor shifts in the threshold values, as the SE-Attention mechanism further adaptively recalibrates channel-wise feature responses to compensate for potential boundary offsets.

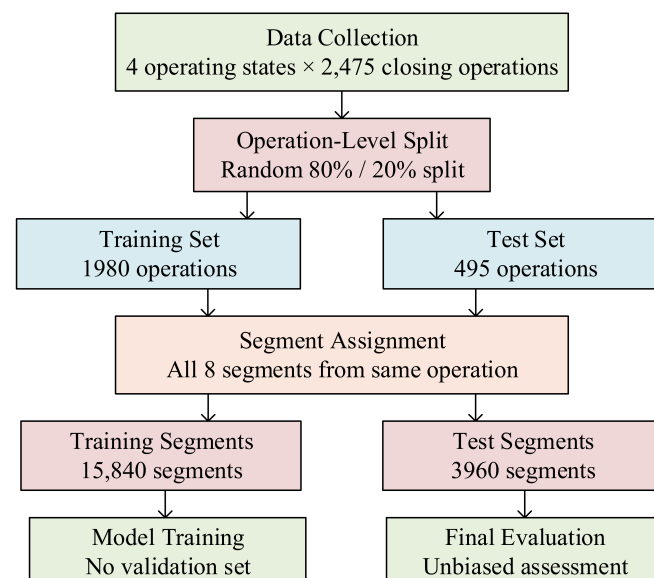
For each closing operation, 4608 sample points are recorded at a sampling rate of 20 kHz. Using a sliding window of length 1024 points and stride 512 points, each operation

is divided into 8 overlapping local segments. The split into training and test sets is performed at the operation level. A total of 2475 closing operations were collected across the four states, specifically 615 for internal foreign objects, 615 for poor contact, 622 for normal, and 622 for shielding cover looseness. These operations were randomly partitioned into a training set of 80%, which corresponds to 1980 operations, and a test set of 20%, which corresponds to 495 operations. All eight segments belonging to the same operation are assigned exclusively to the same set. Considering the slight class imbalance among the four operating states, with 615 samples for internal foreign objects and poor contact versus 622 samples for normal state and shielding cover looseness, a stratified random sampling strategy was adopted to prevent potential skewness in the class distribution of the test set. Specifically, the same 80% training and 20% test split applied at the operation level was also enforced independently within each fault category, ensuring that the proportion of each state remains consistent between the training and test subsets. This operation-wise stratified partitioning guarantees that the test set is a representative subset of every fault category and that the reported model performance provides a reliable estimate of its true generalization capability. Consequently, the training set contains 15,840 segments, calculated as 1980 multiplied by 8, and the test set contains 3960 segments, calculated as 495 multiplied by 8.

To provide a complete picture of the experimental design, the detailed sample distribution across the four operating states in both the training and test subsets is presented in Table 2. This explicit breakdown confirms that the operation-level partitioning preserves the proportion of each fault category between the two subsets, ensuring that the test set is representative of every state and that the reported performance metrics reliably reflect the model's generalization capability. The overall partitioning workflow is illustrated in Figure 7.

**Table 2.** Sample distribution of training and test sets across the four operating states.

| State                     | Training Segments | Test Segments | Number of Samples |
|---------------------------|-------------------|---------------|-------------------|
| Internal foreign objects  | 3940              | 985           | 4925              |
| Poor contact of contacts  | 3940              | 985           | 4925              |
| Normal state              | 3980              | 995           | 4975              |
| Shielding cover looseness | 3980              | 995           | 4975              |



**Figure 7.** Flowchart of the data partitioning and validation strategy.

No separate validation set was employed in this study. The 20% test set was reserved exclusively for final evaluation and was not used for model selection or hyperparameter tuning, ensuring an unbiased assessment of model generalization.

The feature images generated by the sliding window are then assembled into the training set and fed into the deep learning model enhanced with the SE-Attention mechanism for feature learning. The model outputs four GIS disconnect switch states—foreign matter inside, insufficient contact, normal state, and loose shielding cover—encoded in binary form, as shown in Table 3.

**Table 3.** Disconnect switch state classification.

| Label | GIS Disconnect Switch State | Corresponding Code |
|-------|-----------------------------|--------------------|
| 000   | Internal foreign objects    | 000–009            |
| 001   | Poor contact of contacts    | 010–016            |
| 010   | Normal state                | 017–026            |
| 011   | Shielding cover looseness   | 027–038            |

CNN is a feedforward neural network with strong spatial feature extraction capabilities [29]. It performs excellently in image feature extraction and nonlinear mapping and has been widely applied in image processing and fault diagnosis. A typical CNN comprises an input layer, convolution layers, dense layers, and an output layer. After the time–frequency feature images enter the input layer as two-dimensional tensors, the convolutional layers extract local features from the images, and the pooling layers enhance the stability of feature extraction. After aggregation through the fully connected layers of a conventional multilayer perceptron, the final image classification results are output using a Softmax output layer.

#### 2.2.1. Input Layer and Normalization

The input layer receives the preprocessed RGB time–frequency images with dimensions of  $64 \times 64 \times 3$ . To expedite model convergence and enhance training stability, the input data are normalized using the following formula to map pixel values to the range [0, 1]:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where  $X_{norm}$  is the normalized data,  $X$  is the original data,  $X_{min}$  is the minimum value of the original data, and  $X_{max}$  is the maximum value of the original data.

#### 2.2.2. Convolutional Layer

The convolutional layer, as the core component of a CNN, is primarily responsible for extracting local features from the input image and obtaining multi-level abstract feature representations through convolution operations. Since the input in this study is a three-channel time–frequency image, two-dimensional convolution is adopted. The input image size is fixed, and Same convolution is used to ensure sufficient extraction of time–frequency features while maintaining spatial dimensions. Two-dimensional convolution refers to the sliding inner product operation of a convolution kernel over a two-dimensional feature map to extract local spatial and channel features, calculated as follows:

$$Output = f(W * I + B) \quad (2)$$

where  $f$  denotes the activation function,  $W$  is the convolution kernel weight,  $*$  denotes the convolution operation,  $I$  is the feature input, and  $B$  is the bias term. This study constructs a two-layer convolutional structure with the following parameters: The Conv1 contains

32 filters of size  $3 \times 3$  employing a stride of 1 and Same padding mode, producing an output feature map of dimensions  $64 \times 64 \times 32$ . Conv2 has 64 of size  $3 \times 3$  using a stride of 1 and Same padding mode, outputting a feature map of size  $32 \times 32 \times 64$ .

The input time–frequency images are relatively small, and two convolutional layers are sufficient to extract effective local time–frequency features. Deeper networks would increase the number of parameters and may cause overfitting. Experiments show that the two-layer structure achieves satisfactory results for this task. Therefore, this study adopts a two-layer convolutional structure.

LeakyReLU is selected as the activation function, with a negative slope of 0.01, which preserves nonlinearity while avoiding neuron death, benefiting training stability and convergence speed.

### 2.2.3. Pooling Layer

A pooling layer is placed after each convolutional layer to compress the convolutional features and extract the most representative features, thereby reducing the risk of overfitting and decreasing feature dimensionality. Average pooling outputs the average value of the feature map within the receptive field, whereas max pooling outputs the maximum value within the receptive field of the feature map. Max pooling captures the position-independent maximum feature value within the receptive field and better preserves the texture information of the image. Since the time–frequency features in this study exhibit significant energy concentration within local regions, max pooling is adopted, and Same padding is used to maintain the feature map size. The max pooling expression is as follows:

$$P_{i,j} = \max_{(m,n) \in R} X_{i+m,j+n} \quad (3)$$

where  $R$  represents the pooling window region,  $X$  is the input feature map, and  $P$  is the pooling output. This study uses a  $2 \times 2$  pooling window with a stride of 2, which halves the feature map size.

### 2.2.4. Global Average Pooling Layer and Classification Decision Module

If one or more fully connected layers are added after the final convolutional and pooling layers in a traditional CNN, the number of parameters increases significantly, leading to higher model complexity, slower training, and reduced computational efficiency. This study adopts GAP as a substitute for the conventional flattening operation, reducing parameters and enhancing position robustness. The GAP calculation formula is as follows:

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W a_{i,j}^{(c)}, \quad c = 1, 2, \dots, C \quad (4)$$

where  $a_{i,j}^{(c)}$  is the activation value at position  $(i, j)$  of the  $c$ -th channel,  $z_c$  is the output of the  $c$ -th channel after GAP, and the final output vector  $z = [z_1, z_2, \dots, z_C]$  is sent to the subsequent classification layer. In the present study, the feature map dimensions are  $H = W = 32$ , and the number of convolutional channels is  $C = 64$ .

The classification decision module consists of a Dropout regularization strategy and a Softmax classifier.

The classification layer comprises a single dense layer and a Softmax function. The dense layer maps the 64-dimensional feature vector output by GAP to a 4-dimensional

output space, corresponding to the four operating states. The Softmax function converts the output into a probability distribution, satisfying  $\sum_{i=0}^3 \hat{y}_i = 1$ :

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^4 e^{z_j}}, i = 0, 1, 2, 3 \quad (5)$$

Considering that a moderate dropout rate balances regularization strength and feature retention, a dropout rate of 0.4 is set. During training, Dropout randomly discards neurons with probability  $p = 0.4$ . During testing, all neurons participate in prediction, but the activation values are multiplied by the retention probability of 0.6. This forces the network to learn redundant feature representations, enhancing generalization ability. The formula is:

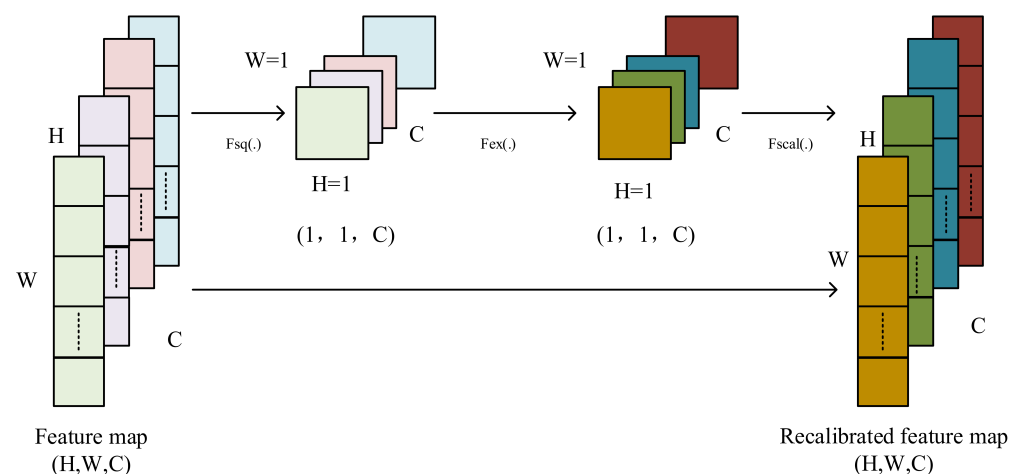
$$y_i^{(l)} = r_i^{(l)} \cdot a_i^{(l)}, r_i^{(l)} \sim \text{Bernoulli}(0.6) \quad (6)$$

### 2.3. Channel Attention Mechanism Module

In disconnect switch fault diagnosis, different fault types have different dependencies on the frequency components of vibration signals. Although CNNs can efficiently extract spatial hierarchical features, their receptive fields are limited by the convolution kernel size, making it difficult to capture long-range dependencies. Moreover, they lack the ability to adaptively weight the importance of different regions in the input. This equal treatment overlooks differences in the importance of information carried by different feature channels, limiting the representation capability and generalization performance of the model when facing engineering problems with complex signal components and overlapping fault frequency bands. To endow the model with the ability to focus on key information as needed, the channel attention mechanism has been developed.

The squeeze-and-excitation attention mechanism can establish global dependencies and assign different weights to each position in the input sequence, thereby screening and enhancing important information. The primary goal objective of the channel attention module is to allocate a weighting factor to every feature channel, allowing the network to adaptively amplify attention to critical channels while suppressing attention to redundant or unimportant channels.

Figure 8 illustrates the architecture of the SE-Attention module. This module comprises three core operations: squeeze, excitation, and scale, forming a complete channel attention mechanism.



**Figure 8.** SE-Attention mechanism module structure.

In the squeeze operation, global average pooling compresses the global information of the image, outputting a 32-dim vector that encodes the global importance of each channel. The compressed vector  $z$  encodes the statistical characteristics of each channel over the entire receptive field, as shown below:

$$Z = [Z_1, Z_2, \dots, Z_C] \in R^C \quad (7)$$

In the excitation phase, the first fully connected layer first reduces the 32-dimensional compressed vector to 2 dimensions and introduces nonlinearity through the ReLU activation function. Subsequently, the second fully connected layer restores the feature dimension to 32 dimensions and generates the importance weight coefficients for each channel through the Sigmoid activation function, producing weights in the  $[0, 1]$  interval:

$$s = F(z, W) = \sigma(W_2 \cdot \delta(W_1 \cdot z + b_1) + b_2) \quad (8)$$

where  $W_1$  represents the weights and biases of the first layer for channel reduction,  $W_2$  restores the channel dimension, the output weight  $s_c$  represents the importance weight of the  $c$ -th channel,  $\sigma$  denotes the ReLU activation, which injects nonlinearity, and  $\delta$  refers to Sigmoid activation function constraining the output to the  $[0, 1]$  interval. The formal expressions of the  $\sigma$  and ReLU activation functions are:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (9)$$

$$ReLU(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (10)$$

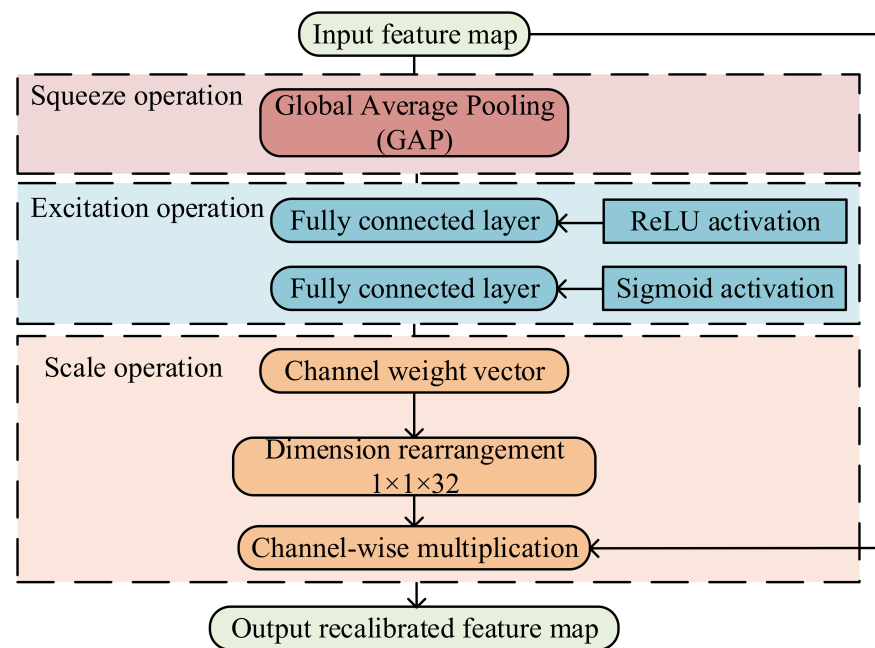
The scale operation broadcasts the  $1 \times 1 \times C$  weights output by the excitation module to the same spatial dimensions as the original feature map, performing channel-wise multiplication to enhance critical channels and suppress secondary channels. The calculation formula is as follows:

$$\tilde{X}_c = F_{scale}(X_c, s_c) = s_c \cdot X_c \quad (11)$$

The weights output by the excitation operation are regarded as the importance of each feature channel, weighted onto the original features through channel-wise multiplication, completing channel-wise recalibration. The recalibrated feature map  $\tilde{X}_c$  retains the details of key channels and suppresses interference from secondary channels.

The architectural diagram of the SE-Attention module is presented in Figure 9. The channel attention mechanism first performs global average pooling on the input feature map  $U$ , compressing the spatial features of each channel into a single value, obtaining the compressed vector  $Z$  with shape  $1 \times 1 \times C$ , where  $C$  denotes the quantity of feature channels. This operation compresses the  $H \times W \times C$  features into a  $1 \times 1 \times C$  feature vector, capturing the overall response magnitude for each channel to facilitate subsequent learning of channel-wise importance distributions.

Next, the compressed vector  $Z$  is fed into the excitation module. This module first passes through a fully connected layer FC1 that reduces the channel dimension from  $C$  to  $C/r$ , where the reduction ratio  $r$  is typically set to 16, followed by a ReLU activation function to introduce nonlinearity. Then, a second fully connected layer FC2 restores the channel dimension to  $C$ , and a Sigmoid activation function performs normalized mapping, outputting the importance weight  $S$  for each channel with shape  $1 \times 1 \times C$ , with weight values ranging between 0 and 1.



**Figure 9.** SE-Attention module flowchart.

Finally, the scale operation performs channel-wise multiplication between the learned channel attention weights  $S$  and the original input feature map  $U$ , producing the recalibrated feature map  $\tilde{X}$ . Through the scale operation, features of important channels are enhanced, while features of secondary or noisy channels are suppressed, enabling the model to adaptively focus on feature channels that are more critical to the current task.

Shallow feature maps have higher spatial resolution and retain rich local time–frequency details. Embedding the SE module after the first convolutional layer and before the first max pooling layer allows for more effective recalibration of these feature channels, enhancing fault-sensitive frequency bands while suppressing noise interference.

The additional parameters introduced by the SE-Attention mechanism amount to 128, representing an increase of less than 1% compared to the main CNN, and the computational overhead is negligible.

This study summarizes the distribution of SE weights under the four fault states, as presented in Table 4. As seen in Table 4, the spectral characteristics of internal foreign objects are highly similar to those of the normal state, with energy mainly concentrated in the low-frequency channel. The SE weights are high in the low-frequency channel and low in the medium- and high-frequency channels, reflecting its low-frequency, stable vibration characteristics. For poor contact faults, energy is distributed across the entire frequency band, exhibiting widespread, intermittent impact characteristics. The SE-Attention weights are relatively balanced across the three channels, with a slight increase in the high-frequency weight, indicating complex spectral features that require comprehensive information from multiple frequency bands. For shielding cover looseness, due to unbalanced forces on the three-phase conductors generating higher harmonics, energy shifts to the medium- and high-frequency ranges. The weights in the medium- and high-frequency channels are significantly higher than those in the low-frequency channel, indicating that the model pays more attention to medium- and high-frequency features. Under normal conditions, the energy distribution is uniform and concentrated, and the SE weights are most balanced across the three channels without a significant dominant frequency band, consistent with its stable, fault-free operating state.

**Table 4.** Average weights of different states of disconnect switch.

| No. | State                     | $\overline{s_R}$ | $\overline{s_G}$ | $\overline{s_B}$ | Dominant Band |
|-----|---------------------------|------------------|------------------|------------------|---------------|
| 0   | Internal foreign objects  | $0.74 \pm 0.08$  | $0.28 \pm 0.06$  | $0.22 \pm 0.05$  | Low           |
| 1   | Poor contact of contacts  | $0.51 \pm 0.09$  | $0.46 \pm 0.08$  | $0.53 \pm 0.11$  | Balanced      |
| 2   | Normal state              | $0.52 \pm 0.07$  | $0.48 \pm 0.06$  | $0.51 \pm 0.08$  | Balanced      |
| 3   | Shielding cover looseness | $0.32 \pm 0.06$  | $0.65 \pm 0.10$  | $0.78 \pm 0.12$  | High          |

The above distribution of SE weights is highly consistent with the physical mechanisms of the four states, validating the effectiveness and interpretability of the SE-Attention mechanism in feature selection and frequency band focusing. By combining traditional spatial feature extraction with channel importance learning, the SE-Attention module achieves multi-scale adaptive analysis of disconnect switch vibration signals.

The interpretability of the SE attention weights warrants further discussion regarding its scope of applicability, generalization capability, and potential limitations. The proposed interpretation is primarily applicable to mechanical fault diagnosis of GIS disconnect switches where vibration signals exhibit distinct spectral signatures across different fault types. The physical basis of the method—frequency-band-specific energy distribution—is rooted in the fundamental vibration mechanisms of the equipment, which remain consistent across similar GIS installations under comparable operating conditions.

Regarding generalization capability, the strength of the proposed approach lies in its physical interpretability rather than purely data-driven pattern matching. The SE weight distributions shown in Table 4 are not arbitrary learned patterns but reflect the underlying spectral characteristics of each fault type. This physical grounding provides two key advantages for generalization: First, because the method learns to focus on frequency bands that are physically relevant to specific fault mechanisms, it is expected to generalize to similar equipment where the same physical principles apply, without requiring retraining or manual recalibration. Second, the multi-scale RGB encoding preserves physical frequency information explicitly, rather than compressing it into black-box features, making the learned representations more transferable across different operating conditions where the spectral signatures of faults remain physically similar.

Regarding limitations, it should be noted that the current analysis is based on a specific set of operating conditions and fault types. In scenarios where vibration signals from different fault types exhibit largely overlapping spectral distributions, the discriminative information encoded in the frequency bands may become less distinct. Future work will extend the evaluation to multiple GIS units and long-term field data to further substantiate the cross-device and temporal generalization of the proposed interpretation framework.

#### 2.4. SE-CNN-Attention Fusion Model

By introducing the SE-Attention mechanism to optimize the traditional CNN, an SE-CNN-Attention fusion diagnostic model is constructed. The CNN is responsible for efficient local spatial feature extraction, while the SE-Attention mechanism adaptively enhances fault-sensitive frequency bands and suppresses noise interference. The synergy between the two improves the model's representation capability for complex fault patterns and effectively enhances diagnostic accuracy for different fault types and severities, forming a more powerful fault diagnosis system. The model mainly consists of four components: a multi-scale feature extraction module, a channel attention mechanism module, a deep feature learning module, and a classification decision module. The structure of the SE-CNN-Attention fusion module is shown in Figure 10.

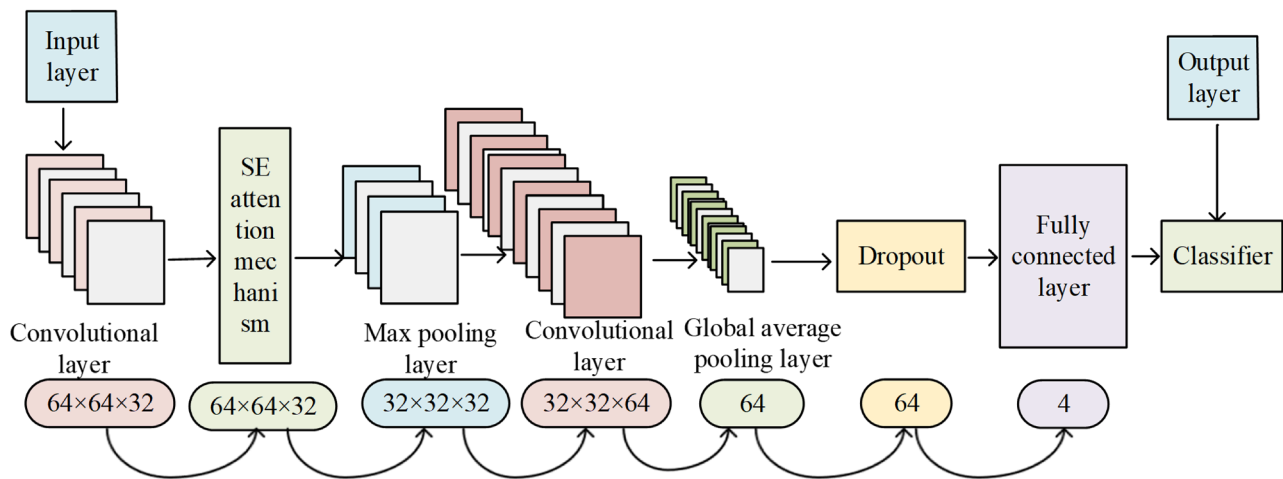


Figure 10. SE-CNN-Attention fusion model structure.

The original vibration signal is first converted into an RGB time–frequency image through image-based processing. Two-dimensional convolution is then used to extract local features. The SE-Attention mechanism is embedded to adaptively recalibrate the feature channels, achieving adaptive frequency band weighting, highlighting fault-sensitive frequency components, and suppressing noise. Spatial dimensionality reduction is achieved through max pooling. After the max pooling layer, global average pooling compresses the feature map into a 64-dimensional vector. Finally, after Dropout and a fully connected layer, the Softmax function outputs the probability distribution of the four states.

### 3. Results

#### 3.1. SE-CNN-Attention Model Parameter Configuration and Training Settings

The SE-CNN-Attention model is constructed using Python (version 3.8.10), and the model parameters are determined based on empirical optimization and grid search. The input layer receives three-channel time–frequency images of size  $64 \times 64 \times 3$ . Two-dimensional convolution kernels are used in the convolutional layers. The first convolutional layer has 32 kernels and the second has 64 kernels. Same convolution is applied, and the ReLU function is used as the activation function. The SE-Attention module is embedded after the first convolutional layer, learning channel weights with a reduction ratio of 16. The pooling layer uses two-dimensional max pooling with Same mode and a window size of  $2 \times 2$ . Global average pooling is used as the feature aggregation layer to replace the traditional flattening operation. The fully connected output layer has 4 neurons with the Softmax activation function, and L2 regularization ( $\lambda = 0.01$ ) is applied. The proposed approach adopts the categorical cross-entropy objective:

$$L = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^4 y_{i,c} \log(\hat{y}_{i,c}) \quad (12)$$

where  $N$  is the batch size,  $y_{i,c}$  is the true label and  $\hat{y}_{i,c}$  is the predicted probability.

After constructing the model, the preprocessed vibration signal dataset is used for iterative training. The training dataset consists of 15,840 samples randomly partitioned in an 8:2 ratio, and the test dataset consists of the remaining 3960 samples. The overall structure and the input–output parameters of each layer are shown in Table 5.

**Table 5.** Main parameter configuration of the SE-CNN-Attention model.

| Parameter          | Value                   | Description                                      |
|--------------------|-------------------------|--------------------------------------------------|
| Input size         | $64 \times 64 \times 3$ | RGB time–frequency image                         |
| Kernel size        | $3 \times 3$            | Two-dimensional convolution                      |
| Conv1 channels     | 32                      | Primary feature extraction                       |
| Conv2 channels     | 64                      | Deep feature extraction                          |
| SE reduction ratio | 16                      | Channel attention dimensionality reduction ratio |
| Dropout rate       | 0.4                     | Prevents overfitting                             |
| Number of epochs   | 500                     | Training rounds                                  |
| Batch size         | 16                      | Number of samples per training batch             |

To ensure a fair comparison, all benchmark models—including SVM, DBN, traditional CNN, ViT, and ResNet-18—were trained and evaluated under identical experimental conditions. Specifically, all deep learning models adopted the same optimizer (Adam), initial learning rate (0.001), batch size (16), and training epochs (500). The SVM baseline was optimized using grid search to identify the optimal hyperparameters. This unified configuration ensures that the observed performance differences primarily reflect the architectural advantages of each model rather than variations in training settings.

### 3.2. Model Evaluation Metrics

Once the SE-CNN-Attention model is trained, the model’s effectiveness is assessed via classification accuracy, precision, recall, and F1-score [30]. The corresponding calculation formulas are provided below:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

$$P = \frac{TP}{TP + FP} \quad (14)$$

$$R = \frac{TP}{TP + FN} \quad (15)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (16)$$

where  $TP$  denotes the number of faulty samples correctly identified as faulty;  $TN$  represents the number of samples that are truly normal and correctly predicted as normal;  $FP$  represents the number of samples that are normal samples mistakenly identified as faulty; and  $FN$  denotes the number of samples that are faulty samples mistakenly identified as normal.

For the multi-category classification, this paper employs both macro-average and weighted-average methods to compute the comprehensive metrics, comprehensively evaluating the model’s classification capability across the four disconnect switch states.

### 3.3. Experimental Results and Analysis

The same 19,800 groups of disconnect switch vibration sample data are fed into different models. The preprocessed dataset is randomly partitioned into training and testing subsets with an 8:2 ratio, with the training set comprising 15,840 samples and the testing subset consists of 3960 samples.

To provide a comprehensive comparison, Table 6 reports the parameter counts, training time, and inference time for all benchmark models alongside the proposed SE-CNN-Attention model. All models were trained and evaluated on the same hardware platform to ensure a fair comparison. Training time is measured as the total duration for 500 epochs. Inference time is calculated as the average processing time per sample over the entire test

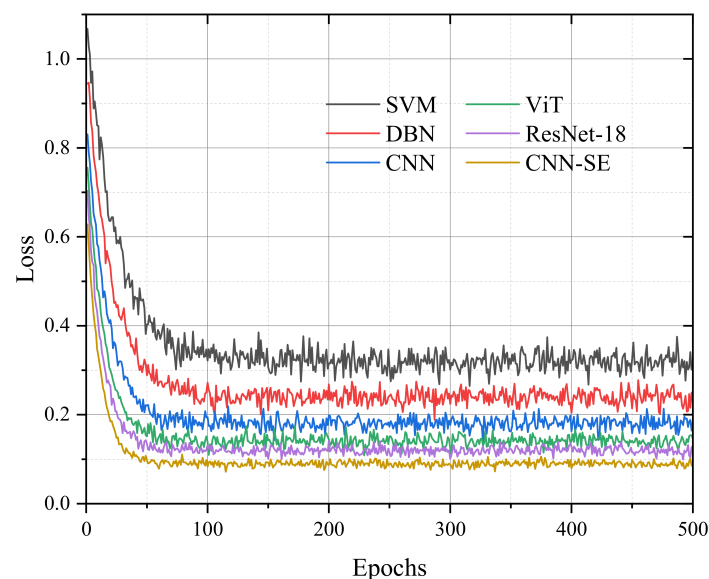
set. SVM is excluded from parameter reporting as its parameters depend on the training set size and are not directly comparable to deep learning models.

**Table 6.** Configuration and computational cost comparison of benchmark models.

| Model            | Parameters (M) | Training Time (min/500 Epochs) | Inference Time (ms/Sample) |
|------------------|----------------|--------------------------------|----------------------------|
| SVM              | N/A            | ~17                            | ~0.05                      |
| DBN              | ~0.4           | ~150                           | ~0.8                       |
| Traditional CNN  | 0.03           | ~200                           | ~1.2                       |
| ViT              | 22             | ~600                           | ~8.5                       |
| ResNet-18        | 11.7           | ~400                           | ~4.2                       |
| SE-CNN-Attention | 0.07           | ~207                           | ~1.3                       |

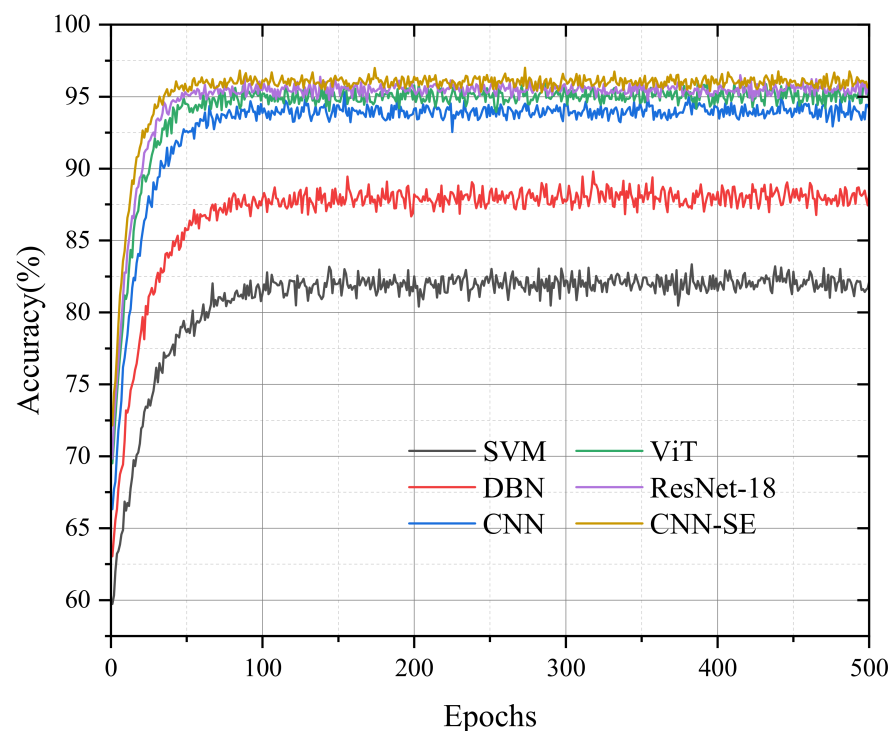
To validate the superiority of the proposed SE-CNN-Attention model, its classification performance for faults is compared with five benchmark models: SVM, DBN, traditional CNN, ViT, and ResNet-18. For the SVM baseline, a set of manually extracted time-domain and frequency-domain statistical features—including mean, variance, peak-to-peak value, and band energy ratios across the low-, medium- and high-frequency bands—was extracted from each sliding window segment and normalized as input. An RBF kernel was adopted, with hyperparameters optimized via grid search to ensure optimal performance.

As presented in Figure 11, the loss trajectories of DBN, traditional CNN, ViT, and ResNet-18 decrease relatively slowly, requiring more training epochs to reach a stable state. As a shallow model, SVM achieves a higher classification error rate. Among these, ViT and ResNet-18 converge faster than DBN and CNN but still lag behind the proposed model. In contrast, the loss curve of the proposed model decreases most rapidly in the early training stage, achieving a low loss level with fewer iterations. In terms of curve stability, the proposed model exhibits the smallest loss fluctuation, without significant oscillation or rebound, while DBN and traditional CNN show some fluctuation during the mid-training stage, and ViT also exhibits minor oscillations due to its self-attention mechanism. This indicates that the introduction of the SE-Attention mechanism enhances training stability. At final convergence, the proposed model achieves the lowest loss value among all compared models, demonstrating that the SE module effectively improves the model's ability to fit fault features through adaptive channel recalibration.



**Figure 11.** Training loss function curves.

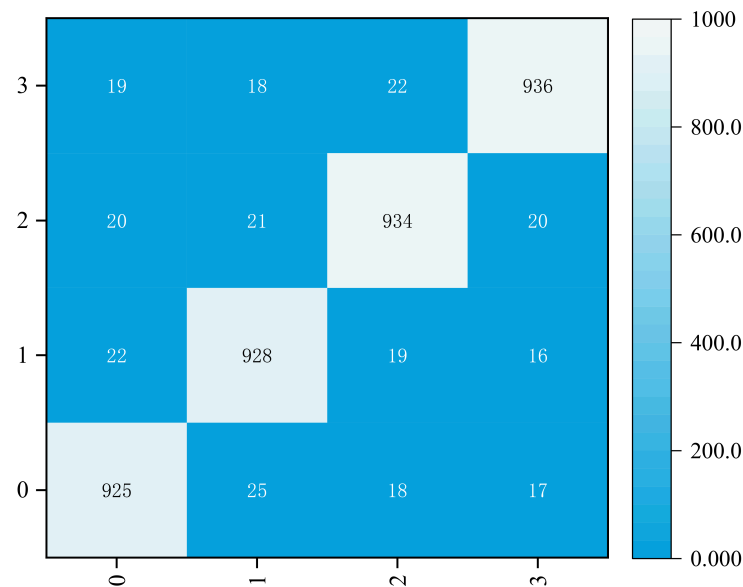
As displayed in Figure 12, the accuracy trends of DBN and traditional CNN rise relatively slowly, requiring more training epochs to achieve high accuracy. ViT and ResNet-18 show faster improvement, with ResNet-18 reaching 90% accuracy earlier than ViT due to its residual connections. Limited by its shallow structure, SVM has relatively low overall accuracy with limited room for improvement. The accuracy curve of the proposed model rises most steeply, achieving high accuracy more quickly and consistently outperforming other models in subsequent training. In terms of final performance, the proposed model achieves significantly higher maximum accuracy than SVM, DBN, traditional CNN, ViT, and ResNet-18, verifying the effectiveness of the synergy between CNN and the SE-Attention mechanism. Furthermore, the gap between training accuracy and validation accuracy remains consistently small for the proposed model, whereas the gap is larger for DBN, traditional CNN, and ViT, indicating that the SE module indirectly plays a regularizing role by suppressing noisy channels, reducing the risk of overfitting and enhancing model generalization.



**Figure 12.** Training accuracy curves.

In summary, the proposed SE-CNN-Attention model outperforms SVM, DBN, traditional CNN, ViT, and ResNet-18 in terms of convergence speed, final accuracy, and training stability, fully validating the effectiveness and superiority of the SE-Attention mechanism in the GIS disconnect switch fault classification task.

When the model is in the final prediction stage, if the data output predicted by the fully connected layer matches the true category of the disconnect switch fault, the sample is considered correctly classified. Figure 13 presents the confusion matrix of the SE-CNN-Attention model on the testing subset, where the categories include four operating states: internal foreign objects, poor contact of contacts, normal state, and shielding cover looseness.



**Figure 13.** Confusion matrix.

To deeply evaluate the model's recognition performance for each specific category, the internal foreign objects state is taken as an example to calculate its evaluation metrics in detail based on the confusion matrix results. Among a total of 985 test data points for internal foreign objects, 925 are correctly predicted as internal foreign objects, 25 are misclassified as poor contact of contacts, 18 are misclassified as normal state, and 17 are misclassified as shielding cover looseness. Therefore, for the internal foreign objects category,  $TP = 925$ ,  $FN = 60$ , and  $FP = 61$ . Through calculation, the accuracy for this category is 93.91%, precision is 93.81%, recall is 93.91%, and F1-score is 93.86%.

Using this method, the accuracy, precision, recall, and F1-score for the four states—internal foreign objects, poor contact of contacts, normal state, and shielding cover looseness—are obtained using Equations (13)–(16) based on the confusion matrix, and the results are presented in Table 7. From Table 7, it can be seen the disconnect switch fault classification model leveraging SE-CNN-Attention achieves high accuracy, precision, and recall for each state. Among them, the recognition performance for poor contact of contacts is the best, while that for internal foreign objects is relatively weaker, which is related to the lower distinguishability of internal foreign object features in vibration signals. Overall, the SE-CNN-Attention model provides effective technical support for intelligent maintenance of disconnect switch equipment and demonstrates reliable classification capability and high engineering practical value for mechanical faults of disconnect switches.

**Table 7.** Classification evaluation metrics.

| State                    | Accuracy | Precision | Recall | F1-Score |
|--------------------------|----------|-----------|--------|----------|
| Internal foreign matter  | 93.91%   | 93.81%    | 93.91% | 93.86%   |
| Poor contact of contacts | 94.21%   | 93.55%    | 94.21% | 93.88%   |
| Normal state             | 93.87%   | 94.06%    | 93.87% | 93.96%   |
| Loose shielding cover    | 94.07%   | 94.64%    | 94.07% | 94.35%   |

## 4. Conclusions

This study presents a fault identification scheme for gas-insulated switchgear disconnect switches by fusing multi-scale RGB image representation with a SE-CNN-Attention mechanism. First, continuous wavelet transform is used to decouple the one-dimensional vibration acceleration signal into three physical dimensions—low, medium, and high

frequencies—which are then mapped to the red, green, and blue channels of an RGB time–frequency feature image, achieving three-dimensional encoding of time, frequency, and energy intensity. Subsequently, a SE-CNN-Attention fusion network fault classification model is constructed leveraging the prepared image dataset, where the SE-Attention mechanism adaptively weights and optimizes the feature channels of the convolutional neural network. The proposed approach is empirically validated by comparisons with existing methods and performance metric assessments. The key findings of this study are summarized below:

- (1) In terms of data preprocessing, the combination of sliding window segmentation and continuous wavelet transform converts the one-dimensional vibration signal into a  $64 \times 64 \times 3$  RGB time–frequency image, effectively preserving the local time–frequency characteristics of the signal and providing physically meaningful input data for subsequent deep learning models. The overlapping window strategy inherently enriches the training set with temporally consistent segments, while no synthetic data generation or external augmentation algorithms are required, as the sample sizes for the four disconnect switch states are balanced and the feature distributions are concentrated.
- (2) In terms of model architecture design, the proposed SE-CNN-Attention model introduces a channel attention mechanism to adaptively weight the RGB three-channel features. Experimental results show that the distribution patterns of SE weights under different fault states are highly consistent with the physical mechanisms: internal foreign objects are dominated by low frequencies, poor contact of contacts shows increased high-frequency weights, shielding cover looseness exhibits significantly enhanced medium- and high-frequency features, and the normal state presents balanced weights across the three channels. This confirms the capability and interpretability of the mechanism in feature selection and frequency band focusing.
- (3) Throughout the training phase, the model achieves stable convergence without significant overfitting. Experimental results demonstrate that the proposed SE-CNN-Attention model achieves an overall test classification accuracy of over 94% in the four-state disconnect switch classification task, outperforming SVM, DBN, traditional CNN, ViT, and ResNet-18.

The proposed method, which integrates multi-scale time–frequency representations with an attention mechanism, can effectively identify typical mechanical faults of disconnect switches. The model exhibits favorable convergence behavior during training, and its adaptive weighting scheme helps enhance the interpretability of classification results. This work may provide a useful reference for condition awareness and intelligent maintenance of high-voltage electrical equipment.

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

**Funding:** This research was funded by Inner Mongolia Power (Group) Co., Ltd. Science and Technology Project, No. nmgdlkxm20250403.

**Data Availability Statement:** Data supporting this paper’s analysis are available upon author approval.

**Acknowledgments:** The authors sincerely thank the Baotou Power Supply Branch of Inner Mongolia Electric Power (Group) Co., Ltd. for providing the experimental platform and field data support.

**Conflicts of Interest:** Authors Xiaofei Kang, Su Xu, Yuqi Liu, Jianguo Du, Chunqiao Fan, Jie Hou and Shuaidong Zhang were employed by the company Baotou Power Supply Branch of Inner Mongolia Electric Power (Group) Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

1. Zhao, X.; Wan, L.; Li, K. Online Detection of Hydrogen Fluoride under Corona Discharge in Gas-Insulated Switchgear Based on Photoacoustic Spectroscopy. *Sensors* **2024**, *24*, 2806. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Yuan, Y.; Ma, S.; Wu, J.; Jia, B.; Li, W.; Luo, X. Frequency Feature Learning from Vibration Information of GIS for Mechanical Fault Detection. *Sensors* **2019**, *19*, 1949. [\[CrossRef\]](#)
3. Kaneko, S.; Okabe, S.; Yoshimura, M.; Muto, H.; Nishida, C.; Kamei, M. Detecting characteristics of various type antennas on partial discharge electromagnetic wave radiating through insulating spacer in gas insulated switchgear. *IEEE Trans. Dielectr. Electr. Insul.* **2009**, *16*, 1462–1472. [\[CrossRef\]](#)
4. Ma, H.; Liu, B.; Xu, H.; Chen, B.; Ju, P.; Zhang, L.; Qu, B. GIS mechanical state identification and defect diagnosis technology based on self-excited vibration of assembled circuit breaker. *IET Sci. Meas. Technol.* **2019**, *14*, 56–63. [\[CrossRef\]](#)
5. Zhong, Y.; Hao, J.; Lliu, Q.; Li, Y.; Li, X.; Liao, R.; Jiang, X. Novel Diagnosis Method for GIS Mechanical Defects Based on an Improved Lightweight CNN Model With Load Adaptive Matching. *IEEE Trans. Ind. Inform.* **2023**, *19*, 11041–11051. [\[CrossRef\]](#)
6. Wu, Z.; Zhang, Q.; Ma, J.; Li, X.; Wen, T. Effectiveness of on-site dielectric test of GIS equipment. *IEEE Trans. Dielectr. Electr. Insul.* **2018**, *25*, 1454–1460. [\[CrossRef\]](#)
7. Zhou, N.; Luo, L.; Sheng, G.; Jiang, X. Scheduling the Imperfect Maintenance and Replacement of Power Substation Equipment: A Risk-Based Optimization Model. *IEEE Trans. Power Deliv.* **2025**, *40*, 2154–2166. [\[CrossRef\]](#)
8. Niu, B.; Bai, T.; Shi, T.; Wu, X.; Han, X.; Liu, W. Initial Characteristics of Submillimeter Metal Particles in GIS under Impact Vibration. *Energies* **2024**, *17*, 5100. [\[CrossRef\]](#)
9. Yu, Y.; Zhao, H. Research on GIS Circuit Breaker Fault Diagnosis Based on Closing Transient Vibration Signals. *Machines* **2025**, *13*, 335. [\[CrossRef\]](#)
10. Cai, Y.Q.; Jin, L. Reviews of Research on Mechanical Fault Diagnosis in GIS. *E3S Web Conf.* **2020**, *155*, 01015. [\[CrossRef\]](#)
11. Li, X.; Hao, J.; Liao, R.; Zhong, Y.; Feng, Y.; Gong, R. Mechanical vibration state and its defect severity development trend prediction for gas-insulated switchgear equipment: Attention-bidirectional gated recurrent unit model construction and experimental verification. *High Volt.* **2025**, *10*, 831–844. [\[CrossRef\]](#)
12. Miao, Y. Review of Image Classification Algorithms Based on Convolutional Neural Networks. *Remote Sens.* **2021**, *13*, 4712. [\[CrossRef\]](#)
13. Zhang, J.; Kong, X.; Cheng, L.; Qi, H.; Yu, M. Intelligent fault diagnosis of rolling bearings based on continuous wavelet transform-multiscale feature fusion and improved channel attention mechanism. *Eksplot. Niezawodn. Maint. Reliab.* **2023**, *25*, 16. [\[CrossRef\]](#)
14. Yang, C.L.; Chen, Z.X.; Yang, C.Y. Sensor Classification Using Convolutional Neural Network by Encoding Multivariate Time Series as Two-Dimensional Colored Images. *Sensors* **2020**, *20*, 168. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Ye, L.; Zhang, K.; Jiang, B.; Simani, S.; Li, X. Small Sample Fault Diagnosis Using Gap-Regularized Loss and Multiscale Attention CNN. *IEEE Trans. Reliab.* **2026**, *75*, 1599–1610. [\[CrossRef\]](#)
16. Li, X.; Xiao, S.; Zhang, F.; Huang, J.; Xie, Z.; Kong, X. A fault diagnosis method with AT-ICNN based on a hybrid attention mechanism and improved convolutional layers. *Appl. Acoust.* **2024**, 225. [\[CrossRef\]](#)
17. Mahmoud, A.G.; Liu, F.; Lv, X.; Jiang, M.; Zhang, F.; Geng, X. Robust bearing fault diagnosis using wavelet-enhanced multi-scale CNN with SE attention. *Eng. Res. Express* **2025**, *7*, 0455b3. [\[CrossRef\]](#)
18. Zhou, X.; Tu, G.; Zhu, X.; Zhao, D.; Zhang, L. A Three-Channel Improved SE-Attention Mechanism Network Based on SVD for High-Order Signal Modulation Recognition. *Electronics* **2025**, *14*, 2233. [\[CrossRef\]](#)
19. Hu, J.; Shen, L.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *42*, 2011–2023. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Wang, M.; Cheng, C.; Chu, M. Mscon-C: A Model to Integrate Multiscale Convolution and CBAM Attention Mechanism for Bearing Fault Diagnosis. *Electronics* **2026**, *15*, 1335. [\[CrossRef\]](#)
21. Feng, Q.; Li, J.; He, Q. Photoelectric Measurement and Sensing: New Technology and Applications. *Sensors* **2023**, *23*, 8584. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Rothberg, S.J.; Allen, M.S.; Castellini, P.; Di Maio, D.; Dirckx, J.J.J.; Ewins, D.J.; Halkon, B.J.; Muyshondt, P.; Paone, N.; Ryan, T.; et al. An international review of laser Doppler vibrometry: Making light work of vibration measurement. *Opt. Lasers Eng.* **2016**, *99*, 11–22. [\[CrossRef\]](#)

23. Debeleac, C.N. Vibration-Based Fault Identification in Compaction Equipment Using Feature Extraction Techniques. *Appl. Sci.* **2026**, *16*, 4517. [[CrossRef](#)]
24. Wang, J.S.; Qiu, Y.C.; Wu, X.H.; Sun, X.B. Frequency response characteristics of the UHF sensors for partial discharge detection in GIS. *Proc. CSEE* **2000**, *20*, 42–45.
25. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [[CrossRef](#)]
26. Kim, G.I.; Lee, J. DFAN: Single Image Super-Resolution Using Stationary Wavelet-Based Dual Frequency Adaptation Network. *Symmetry* **2025**, *17*, 1175. [[CrossRef](#)]
27. Ma, C.; Zhang, W.; Meng, L.; Yang, M.; Zhang, K.; Xu, Y. A dual-objective optimized reweighted overlapping group sparse framework integrating frequency slice function for robust bearing fault diagnosis. *Mech. Syst. Signal Process.* **2026**, *242*, 113678. [[CrossRef](#)]
28. Gao, S.; Jia, M.; Yao, D.; Wang, J. Multi-Source Localization Using Optimized Time-Frequency Representation and Sparsity Component Analysis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 3564–3578. [[CrossRef](#)]
29. Shang, X.; Li, W.; Yuan, F.; Zhi, H.; Gao, Z.; Guo, M.; Xin, B. Research on Fault Diagnosis of UAV Rotor Motor Bearings Based on WPT-CEEMD-CNN-LSTM. *Machines* **2025**, *13*, 287. [[CrossRef](#)]
30. Wang, J.; Chen, Y.; Wang, Z.; Gu, Z.; Zhang, B.; Qiu, H.; Zhang, K.; Zhao, W.; Zhang, Y. A Heterogeneous System With Computing in Memory Processing Elements to Accelerate CNN Inference. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2025**, *72*, 3411–3424. [[CrossRef](#)]

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