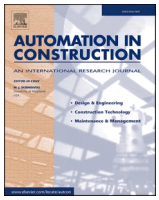
Automation in Construction 182 (2026) 106694

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon



Multimodal data fusion for welding defect detection using ensemble deep learning

Shiqiang Tang [a], Feilong Fei [b], Limao Zhang [a,c,*], Jinfeng Yu [d]

[a] School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China
[b] Hubei Yuntian Construction Technology Co., Ltd., No. 8, Huacheng Avenue, East Lake New Technology Development Zone, Wuhan, Hubei 430070, China
[c] National Center of Technology Innovation for Digital Construction, Huazhong University of Science and Technology, 1037 Luoyu Road, Hongshan District, Wuhan, Hubei 430074, China.
[d] Hubei Municipal Construction Group Co., Ltd., 999 Youyi Avenue, Hongshan District, Wuhan, Hubei 430080, China

ARTICLE INFO

Keywords:
Multimodal fusion
Deep learning
Defects detection
Resistance spot welding
Model interpretation

ABSTRACT

This study proposes a multimodal deep learning model for high-precision automated detection of resistance spot welding defects. A dual-input weight-sharing network is employed to process the surface images of the weld nugget, while infrared images and welding parameter data are processed by two additional base models. The outputs of these base models are fused using Dempster–Shafer theory, yielding the ensemble multimodal deep learning model (EMMDL). Validation on a welding dataset reveals that: (1) EMMDL achieves an accuracy of 91.6 %, significantly outperforming base models with single modality; (2) Dual-input and weight sharing increases classification accuracy by 7.87 % and enhances robustness in small sample scenarios; (3) The model uses more information from infrared images when identifying bad samples. By integrating complementary multimodal information, EMMDL overcomes blind spots of single-source methods and provides interpretable decision support for industrial quality control.

1. Introduction

The accelerated development of global urbanization has given rise to unprecedented demand for infrastructure construction, including high-rise buildings, subway tunnels, and transportation bridges [1]. As a critical joining process in infrastructure construction, welding utilizes a heat source to melt the base metal and filler metal, forming high-strength, highly stable permanent bonds upon solidification, thereby ensuring the integrity and reliability of the component [2,3]. This ensures the integrity and reliability of structural components. Given its pivotal role in load-bearing capacity, sealing performance, and corrosion resistance, welding quality directly affects the structural integrity, safety, durability, and service life of infrastructure. Consequently, precise monitoring and evaluation of the welding process and outcomes are of significant importance.

However, as a complex thermodynamic process, welding is prone to defects like cracks, porosity, slag inclusions, lack of fusion, incomplete penetration, and undercut due to material variations, parameter fluctuations, environmental factors, or operational errors. These defects cause geometric discontinuities and act as stress concentrators, weakening the weld's mechanical properties [4–6]. Under high-cycle fatigue or severe environmental corrosion, they can lead to catastrophic fractures [7,8]. To ensure reliable and controllable welding quality, traditional defect detection methods primarily include X-ray radiography, ultrasonic testing, magnetic particle inspection, and liquid penetrant testing [9–11]. These methods are widely used in scenarios such as bridge girder welding inspection, pipeline joint integrity evaluation, and steel structure connection quality control [12]. In actual construction, these methods generally rely on operator experience and suffer from limitations such as low detection efficiency and insufficient automation. Traditional machine learning methods analyze welding images and process data (e.g., current and voltage) for defect classification [13–15]. While these methods capture anomalies, they are inefficient, subject to cognitive bias, and struggle to address the complex nature of defects, hindered by shape, distribution, and noise interference.

In recent years, deep learning has gradually permeated the field of defect detection due to its outstanding performance in image recognition, signal processing, and multi-source data fusion. It has been widely applied in scenarios such as concrete crack detection, tunnel lining

* Corresponding author.
E-mail addresses: shiqiangtang@hust.edu.cn (S. Tang), zlm@hust.edu.cn (L. Zhang).

https://doi.org/10.1016/j.autcon.2025.106694
Received 1 August 2025; Received in revised form 21 November 2025; Accepted 28 November 2025
Available online 8 December 2025

defect identification, and reinforcement corrosion assessment [16]. Deep learning architectures can extract deep correlations between data features and welding defects from sensor data, demonstrating superior accuracy in defect localization, classification, and size quantification. Concurrently, diverse detection methods provide rich input data for deep learning models. X-ray [17], infrared thermal imaging [18], welding parameters such as current and voltage [19], and laser scanning [20] enable the identification and quantitative assessment of various defect types from external surfaces to internal structures, viewed from multiple perspectives. Deep learning models driven by high-precision sensor data significantly enhance defect detection rates, gradually establishing themselves as the mainstream research paradigm for intelligent welding quality monitoring.

The limitation of deep learning models based on single-modal data lies in the inability of a single data source to fully characterize the complex physical nature of welding defects. A single modality can only provide partial and localized characterization of welding defects, lacking global completeness and criterion redundancy. Therefore, multimodal deep learning that fuses complementary modal information is becoming essential for improving the accuracy and generalization ability of welding defect recognition. Common fusion strategies include feature-level, decision-level, and hybrid-level fusion [21]. Feature concatenation assumes equal modal reliability, while simple averaging or voting fails when modalities provide contradictory evidence. Material differences across construction scenarios (such as the combination of prefabricated steel frames and prestressed concrete [22–24]), diverse structural connections (such as welding versus bolted connections [25,26]), and varying temperatures with fatigue loads during construction (e.g., the fatigue loads on prefabricated airport pavements [27] and the non-uniform temperature effects in spatial truss structures [28]), increase the likelihood of conflicting modal predictions. In contrast, Dempster-Shafer (DS) theory provides a rigorous framework that explicitly considers epistemic uncertainty and inter-modal information conflict, enabling robust decision-making despite modal inconsistency.

The ensemble learning strategy can effectively improve the accuracy of defect recognition, but the “black-box” nature of deep learning models hinders understanding and trust of their internal decision-making processes [29]. In structural health monitoring, this is particularly critical. Engineers must verify that models genuinely focus on actual defects, such as concrete cracks and steel corrosion, to avoid safety hazards arising from misjudgment. To this end, the gradient-based visualization method, Gradient-weighted Class Activation Mapping (Grad-CAM) [30], is employed to reveal the image regions a model focuses on when making decisions, providing intuitive diagnostic basis. On the other hand, understanding the contribution of each modality data source to final predictions helps in modality importance analysis. Traditional methods often rely on performance variation [31], which can be influenced by model performance robustness. Recently, a performance-agnostic method based on Shapley value theory, called MultiModal SHapley Additive exPlanations (MM-SHAP), was proposed [32]. It can more reasonably measure the activity level of different input data in the final decision-making process. Particularly considering the diversity of civil engineering scenarios (such as automatic welding in aerial building machines [33], high-precision welding control in [34] structure assembly, and real-time detection of welding interfaces in tunnel boring machines [35,36]), the combination of high predictive accuracy and interpretability is crucial. Ensuring the adaptability and effectiveness of the model when considering different data sources and features is of great significance for enhancing the reliability and practicality of the welding defect detection system [37].

In summary, the blind spots of single-modality detection systems in capturing comprehensive defect characteristics coupled with the opacity of deep learning decision-making processes, present significant challenges for reliable welding quality assessment, where accuracy and trustworthiness are essential. Effective welding inspection systems must not only achieve high detection accuracy but also provide transparent, interpretable decisions while handling challenging conditions such as imbalanced defect distributions, varying data quality, and limited training samples. Therefore, the development of multimodal fusion approaches enhanced with interpretability mechanisms represents a crucial advancement toward addressing these technical challenges. Specifically, this research pursues the following objectives: (1) Achieve robust defect classification performance with high accuracy via effective utilization of data under conditions of limited and imbalanced training data; (2) Reduce computational complexity to enable efficient processing of multimodal data streams without sacrificing detection accuracy; and (3) Provide quantitative interpretability metrics that enable users to trust model decisions and optimize the sensor layout for quality control applications. Fig. 1 briefly presents the research gaps and research objectives.

This paper needs to address the following questions: (1) How to comprehensively utilize multimodal data from diverse sensors to reduce recognition errors during welding defect identification and classification. (2) How to design feature extraction networks to make full use of limited data for unique data types. (3) How to interpret model decisions, identify the specific roles of different modal data, and quantify the amount of information they contribute to classification decisions. To address these issues, this study proposes an ensemble deep learning method to enhance welding defect identification accuracy by leveraging complementary information from multimodal sources. Specifically, the information extracted by three sub-classifiers from three distinct data modalities is fused using DS theory. One of these sub-classifiers is a weight-sharing network based on ResNet that processes dual-image inputs from multi-view perspectives. Furthermore, the proposed model is interpreted based on Grad-CAM and MM-SHAP.

The innovative contributions of this study are as follows: First, defect classification accuracy is enhanced through a dual-input weight-sharing network and ensemble learning methods. Second, model interpretation is achieved using Grad-CAM and MM-SHAP to identify important modalities for subsequent improvements. Experimental results show that the proposed method overcomes the attention blind spots of single-modal defect recognition and provides valuable insights for rapid defect identification in welding and other manufacturing scenarios under an intelligent context. This research addresses the failure of single-modal detection caused by factors such as material variations and environmental interference in industrial inspection, offering a more robust defect identification framework with interpretability potential.

The rest of this article is organized as follows: Section 2 reviews relevant research on defect detection in welding scenarios. Section 3 provides a detailed description of the research framework and introduces specific technical details. Section 4 validates the effectiveness and feasibility of the proposed method via a case study on a welding dataset. Section 5 demonstrates the superiority of the proposed method over other deep learning models through comparative studies. Finally, Section 6 summarizes the main conclusions of this study, discusses the limitations of the current work, and explores potential future research directions.

2. Literature review on welding defect inspection

Welding quality inspection has gained widespread attention as a key element in ensuring the safety of engineering structures and product performance. Research in fields such as steel bridge engineering, high-rise building construction, and tunnel engineering has advanced significantly, with notable developments in both detection principles and technological applications. Overall, existing research revolves around multi-source data sensing, feature extraction algorithms, and intelligent decision models. The goal is to achieve precise recognition and classification of both surface and internal weld defects, offering reliable support for defect warning and quality evaluation during welding. The technological evolution demonstrates a notable transition

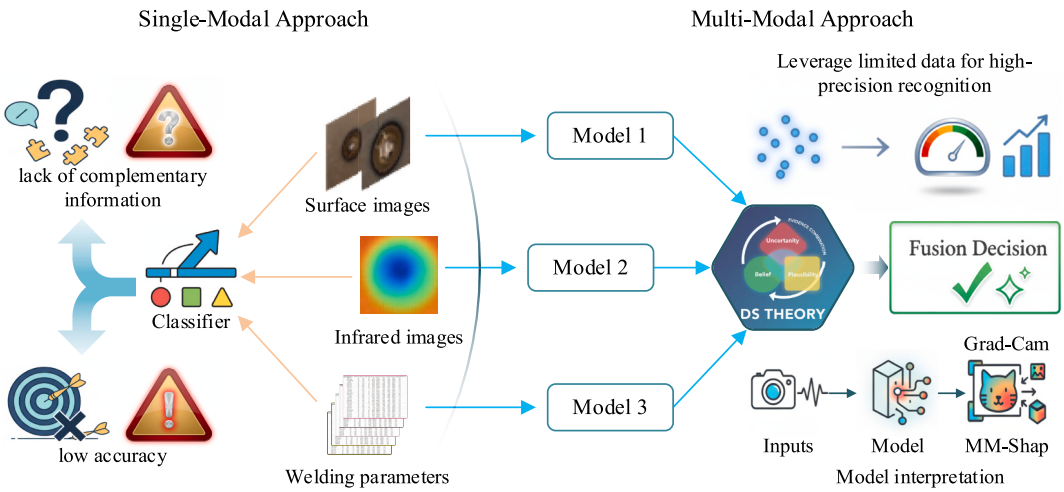


Fig. 1. Conceptual diagram of research gaps and goal.

from single-modal analysis to multimodal fusion, as well as from manual feature design to automatic feature learning.

Early technical frameworks for welding quality inspection relied on traditional machine learning coupled with manual feature extraction. These methods rely on expert experience to design feature extraction rules, extracting statistical features and morphological features from welding process data and weld images, and exhibit a high degree of subjectivity. Subsequently, classifiers such as Support Vector Machines (SVM) and Random Forests are used to identify defects [14,38]. Xia et al. [39] monitored RSW process signals, including dynamic resistance, electrode force, and electrode displacement, found a strong correlation between abrupt electrode force changes and spatter. Li et al. [40] employed a structured light vision detection system for weld profile measurement, extracting visual features to quantify weld dimensions and detect defects. Yu et al. [9] enhanced the detection sensitivity for small defects by performing modal analysis on infinitely large welded steel plates at high frequencies using the finite element method. Despite offering interpretability benefits in specific contexts, their performance is critically dependent on the quality of feature engineering and struggles to effectively capture the complex nonlinear relationships resulting from the multi-factor coupling in the welding process.

Increased computational resources and the accumulation of large-scale annotated data have fostered the widespread adoption of deep learning methods for defect detection [41,42], significantly improving the safety of processes such as tunnel excavation and high-rise building construction [43]. Vision-based deep learning extracts features directly from weld surface images, greatly improving the detection of visually apparent defects. Liu et al. [44] introduced a framework integrating YOLOv8, neural ensembles, and voting to enhance welding quality classification. Kumar et al. [45] developed a semi-supervised transfer learning-based multi-domain network for weld image segmentation and defect detection. These methods are driven by training data, and excessively high model complexity may lead to low computational efficiency and the risk of overfitting. On the other hand, the robust feature learning and expressive capacity of deep learning enable the generation of virtual samples. Dai et al. [46] used GANs with gradient penalty to generate diverse minority defect images and a transfer learning classifier, improving spot weld defect classification. However, the training process of GANs is unstable, and the generated virtual samples may contain artifacts, which affects the generalization ability. These methods offer the advantage of end-to-end training over traditional methods, effectively capturing subtle surface defect characteristics. Through training on large-scale datasets, they effectively overcome the subjectivity limitations inherent in feature design using traditional methods, but demand higher image quality and larger volumes of annotated data.

Non-vision-based deep learning inputs data from various sensors into deep learning models for welding quality assessment. The monitoring data used involves various modalities such as X-ray detection [4], infrared thermography [47], welding parameter curves [48], laser imaging [49], and ultrasonic detection [50]. These techniques represent welding conditions from various physical dimensions, and the semantic information related to welding quality is effectively captured through the powerful feature extraction capabilities of deep learning models like CNN. Zhang et al. [51] developed a series of functional modules with enhanced feature information and multi-scale fusion abilities, incorporating adaptive optimization strategies and defect size mapping algorithms to achieve high-accuracy detection and size measurement of defects in low-quality X-ray images. Zhou et al. [52] proposed an autonomous deep learning framework that analyzes infrared video from the RSW process to achieve rapid and highly accurate predictions of weld nugget shape and size. Zhou et al. [53] integrated denoising, spatiotemporal attention, and multiple residual modules to analyze vibration excitation response signals in RSW joints, enabling automatic and efficient detection of nugget quality. These technologies capture multi-dimensional welding states, allowing non-destructive detection of internal hidden defects. When integrated with process parameter modeling, they enable a thorough analysis of defect formation mechanisms.

Multimodal defect detection methods integrate multi-source data, creating cross-modal association models at either the feature or decision level [54]. For example, in yarn quality detection, taking fiber and process parameters as the main modality and yarn appearance images as the auxiliary modality, soft sensing of yarn quality could be achieved through a pre-trained network and a feature fusion mechanism [55]. Multimodal deep learning has also found certain applications in welding quality recognition. Wang et al. [56] input voltage signals processed by short-time Fourier transform along with molten pool images into a hybrid CNN model for feature extraction and concatenation, achieving online identification of welding defects. He et al. [57] used CNNs and Vision Transformers to extract features from magneto-optical imaging and infrared thermal data, and fused them in frequency channel attention networks, achieving efficient identification of defect-free, unfused, crater, and crack defects in resistance spot welds. Multi-modal systems utilize spatiotemporal alignment and feature complementarity mechanisms, not only suppressing noise interference from single modalities but also constructing a global representation of welding quality, driving welding quality detection toward end-to-end and fully automated processes, supporting the construction of infrastructure such as large-scale bridge projects, super high-rise buildings, and energy facilities.

In summary, existing literature has widely explored defect

recognition technologies in welding processes using various technical approaches. Rich semantic information related to defects during welding is recorded in various data modalities associated with physical phenomena, such as surface images and infrared thermographs. Feature extraction techniques based on deep learning can thoroughly investigate and differentiate various types of welding defects. However, current research mostly considers single-modal information as input to deep learning models, with few studies considering multimodal inputs that are limited to two modalities. Therefore, this paper aims to propose an ensemble multimodal deep learning method for welding defect recognition, enhancing classification accuracy via multisource data integration and targeted network architecture to advance intelligent welding processes.

3. Methodology

To accurately and automatically detect the welding quality, the EMMDL model proposed in this paper integrates multiple basic classification models. The EMMDL model accepts multi-modal data generated during the welding process as input, employs deep learning networks for feature extraction and classification, and outputs predictions for welding quality categories. This study comprises three principal workflows: (1) Acquisition and preprocessing of multimodal data during welding, including infrared images, RGB images, and welding parameters; (2) Pre-training and evaluation of base classifiers for each modality, followed by their fusion via Dempster-Shafer theory to establish a multimodal ensemble deep learning framework; (3) Interpretability analysis of the model using Grad-CAM and MM-SHAP methods, alongside quantify and comparative assessment of modality-specific contributions. The flowchart for this study is shown in Fig. 2, and a more detailed description will be provided below.

3.1. Description of multimodal data in welding

RSW is a representative resistance welding process in which precise regulation of electrode pressure, welding current, and duration achieves the metallurgical bonding of metal joints [58]. The procedure primarily comprises the following steps: First, the preprocessed workpieces are overlapped and positioned between the upper and lower electrodes of a spot welding machine, applying consistent pressure to eliminate interfacial gaps and establish a stable conductive pathway. Subsequently, a high-intensity current is applied for a brief duration, and the joule heating induced by contact resistance at the interface causes localized melting, forming a molten nugget. After the current cutoff, the electrode pressure is maintained to direct the solidification of the nugget under dynamic compaction, yielding a dense equiaxed grain weld. Concurrently, the electrode water-cooling system accelerates heat dissipation to prevent microstructural embrittlement. Finally, the pressure is released, and the workpiece is removed. A single spot welding cycle can be completed within 0.1 to 2 s.

Infrared images, surface digital images, and process parameter data are used as critical indicators for monitoring and quality assessment during resistance spot welding. During welding, the high-thermal-sensitivity infrared imaging system captures thermographic data that records the spatiotemporal evolution of the heat field, thereby reflecting the temperature distribution at the electrode-workpiece interface, providing essential insights for analyzing heat-input behaviors, identifying abnormal heating zones, and assessing nugget formation. Surface images of the welded component serve to detect surface defects such as spatter, burn-through, and indentations, and to some extent reveal issues arising from uneven electrode force or abnormal heat input, thereby supplying foundational information for evaluating the surface integrity and macroscopic defects of RSW. Cameras are used to capture images of both the upper and lower surfaces of the weld specimen, preventing the omission of defects confined to a single surface. Regarding welding process parameters, the input variables include welding current, weld time, electrode-applied force, and physical attributes of the materials, such as sheet thickness and material composition. These parameters directly influence heat input and pressure distribution, consequently determining nugget dimensions and joint strength. Output parameters primarily consist of post-weld mechanical performance metrics, such as tensile-shear strength and nugget diameter, which respectively characterize the joint strength and geometric dimensions, obtained via mechanical testing and measurement. As shown

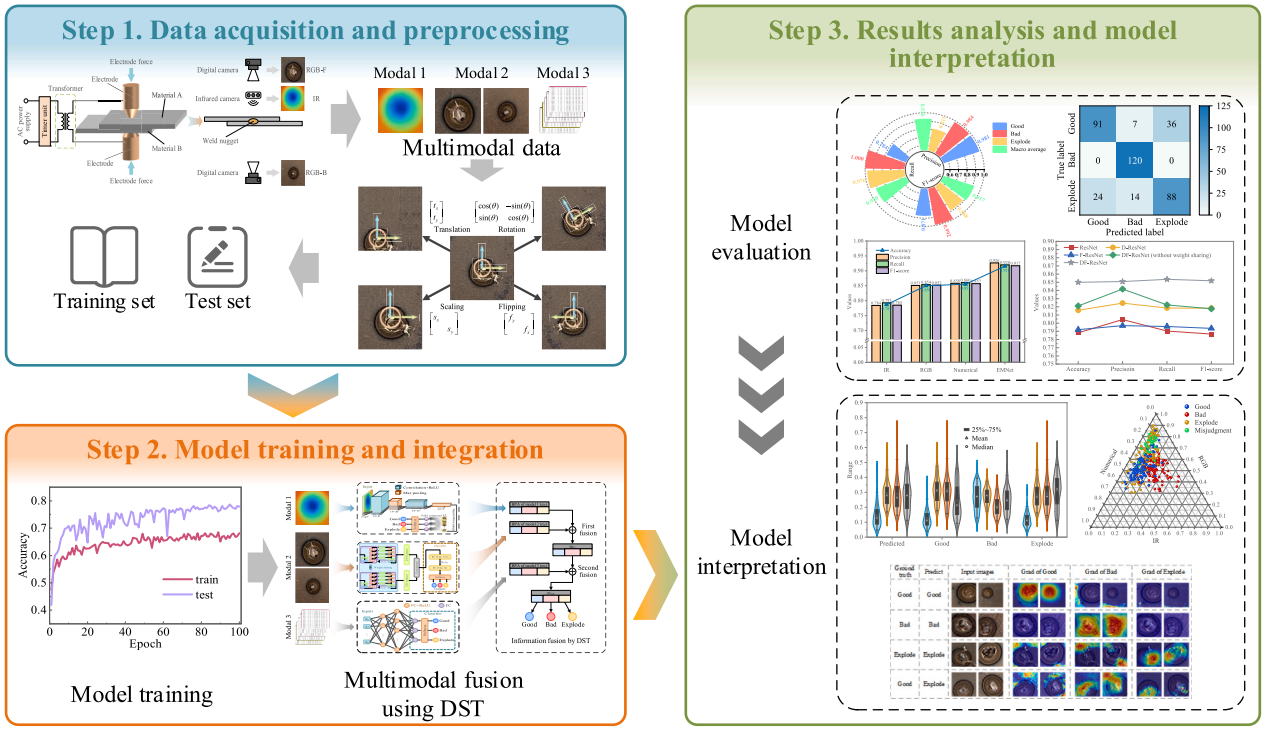


Fig. 2. Flowchart of the proposed multimodal deep learning framework.

in Fig. 3, these data enable comprehensive monitoring throughout the welding process.

In this study, weld quality is categorized into three classes: Good, Bad, and Explode, corresponding to normal nugget formation, incomplete fusion or cold soldering, and severe spatter events during welding, respectively. However, during actual data acquisition, a pronounced imbalance in distribution is observed across different categories. Specifically, the proportion of good weld samples substantially exceeds that of bad and exploded samples. Such class imbalance during supervised learning can bias the classifier toward the majority class, reducing its ability to recognize boundaries and minority classes and undermining overall predictive accuracy and robustness. To mitigate the modeling challenges posed by class imbalance, we introduce data augmentation strategies specifically designed for image data and numerical parameter data.

For the acquired infrared and bi-surface weld images, the affine transformation is employed for augmentation. Affine transformation is a linear spatial mapping that preserves the fundamental geometric structures of images and is extensively employed in computer vision and image processing [59,60]. Affine transformation possesses properties preserving collinearity and proportionality, meaning collinear points in the original image remain collinear, and parallel lines stay parallel following the transformation. Therefore, affine transformations can effectively generate geometrically varied yet semantically consistent image samples without introducing nonlinear distortions. This extends the spatial distribution of images and enhances the model's robustness to weld nugget pose and scale variations, and is suitable for training data expansion and generalization improvement. Especially for the "Explode" category with limited samples, affine augmentation aids in simulating additional potential spatter variations, enhancing recognition capability for this class. To preserve positional accuracy following the affine transformation, the geometric center of the image is typically used as the origin of transformation. In this study, four affine operations, translation, rotation, scaling, and flipping, are employed to augment the image data. This transformation can be represented by a linear transformation matrix combined with a translation vector in 2D space, as denoted by Eq. (1):

$$\boldsymbol{X}' = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{T} \tag{1}$$

where $\boldsymbol{X} = [x, y]^T$ denotes the pixel coordinates in the original image; $\boldsymbol{X}' = [x', y']^T$ represents the transformed coordinates; $\boldsymbol{T} = [t_x, t_y]^T$ is the translation vector, indicating a shift of t_x pixels along the x-axis and t_y pixels along the y-axis; $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$ constitutes the linear transformation matrix governing operations such as rotation, scaling, and flipping of the image, obtainable through the multiplicative composition of three operational matrices: rotation $\boldsymbol{R}$, scaling $\boldsymbol{S}$, and flipping $\boldsymbol{F}$. Their mathematical expressions are given by Eqs. (2) to (4):

$$\boldsymbol{R} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{2}$$

$$\boldsymbol{S} = \begin{bmatrix} s_x & \\ & s_y \end{bmatrix} \tag{3}$$

$$\boldsymbol{F} = \begin{bmatrix} f_y & \\ & f_x \end{bmatrix} \tag{4}$$

where θ represents the counter-clockwise rotation angle about the origin; s_x and s_y denote the scaling factors along the x- and y-axes, respectively; $f_x, f_y \in \{-1, 1\}$ control flipping along the x-axis and y-axis, with a value of 1 indicating no flip and -1 indicating flipping. The visual impact of different transformation methods is shown in Fig. 4.

Numerical input and output parameters of the welding process are augmented via a noise perturbation strategy. By constructing Gaussian noise neighborhoods within the parameter space of original minority class samples, physically plausible new samples are generated to enrich the feature distribution for model training [61,62]. Specifically, original values are multiplied by random numbers with a mean of 1 and a variance of 0.01, ensuring a 95 % confidence interval of [0.9, 1.1] by the 3σ rule. This operation avoids sample redundancy while preserving key features, and the multiplication operation ensures that the sign of the values remains unchanged.

Welding parameters may exhibit multicollinearity. For instance, higher welding current may correlate with shorter welding time, while thicker workpieces require extended welding times. When severe multicollinearity occurs, mutual interference among independent variables makes it impossible for the model to obtain the true relationship between independent variables and the dependent variable. Multicollinearity diagnostics can quantify the extent of multicollinearity among variables, thus providing a basis for subsequent processing. The Variance Inflation Factor (VIF) is a common metric for assessing multicollinearity severity in multiple linear regression and is calculated via Eq. (5):

$$\mathrm{VIF}(X_i) = \frac{1}{1 - R_i^2} \tag{5}$$

where R_i is the coefficient of determination obtained when X_i is regressed linearly against all other predictor variables. The VIF quantifies the extent to which multicollinearity among predictors inflates the variance of regression coefficients, with VIF $\geq$ 10 indicating severe multicollinearity. By iteratively removing the variable with the highest VIF and recalculating until all VIFs of all independent variables fall below 10, mutual independence can be ensured.

The entire dataset is randomly partitioned into training and test sets.

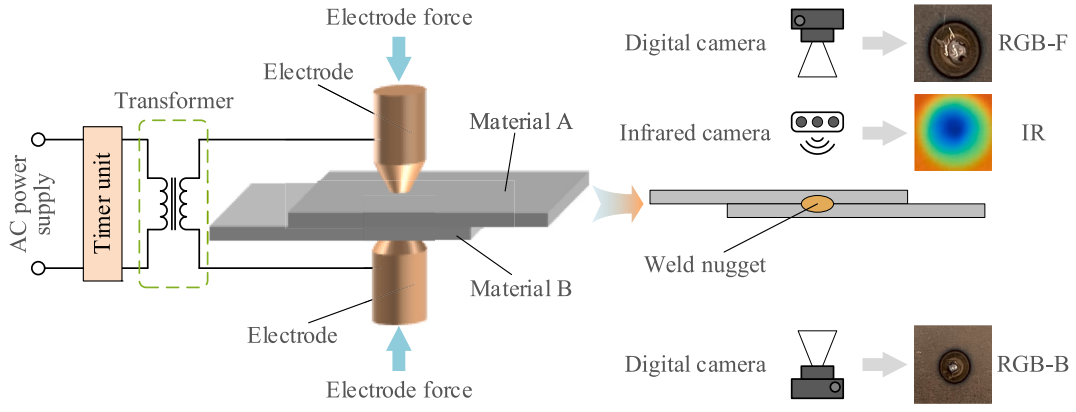


Fig. 3. Multimodal data acquisition in RSW.

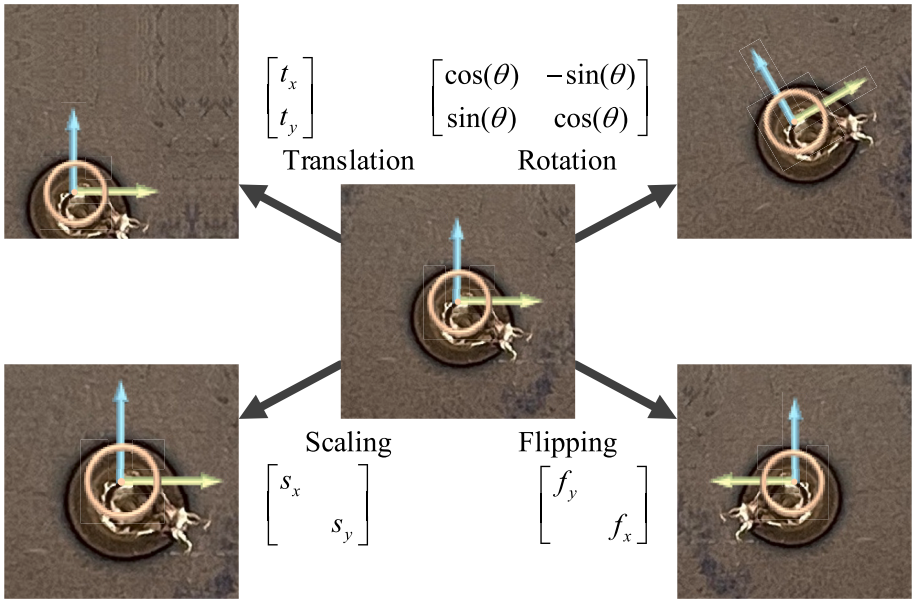


Fig. 4. The effects of different affine transformations.

It is important to note that allocating augmented samples derived from the same original data to both training and test sets induces train-test contamination. A contaminated model will exhibit spurious robustness and accuracy on the test set, biasing model evaluation metrics. A physical segregation mechanism is employed to prevent test set contamination of the training set. Specifically, after random division of the original dataset, augmentation is applied independently within each group, ensuring that augmented samples remain within the same group. Furthermore, stratified sampling based on class-label distribution is employed to maintain consistent category proportions across the training and test sets, thereby mitigating the potential bias of the model.

3.2. Image classification based on ResNet

Residual Networks (ResNet) is a seminal convolutional neural network architecture in the field of computer vision, introduced by the research team led by Kaiming He at Microsoft Research Asia in 2015 [63], and which won the 2015 ILSVRC (ImageNet Large Scale Visual Recognition Challenge). To address the prevalent challenges of gradient vanishing and degradation commonly encountered during the training of deep networks, ResNet innovatively introduced residual blocks. Through skip connections, shallow features are directly transmitted to deep layers, enhancing representational capacity without significantly increasing computational overhead. ResNet-18 stands as a quintessential and widely adopted model within the ResNet family, comprising 18 weighted layers. This research employs the ResNet-18 backbone to devise a deep neural network architecture tailored for the classification of bi-surface welding spot images, incorporating their domain-specific characteristics.

The core of ResNet-18 is comprised of four stages, each containing two residual blocks. As illustrated in Fig. 5 (a), Stage 1 employs two identical residual blocks as its fundamental building units, each

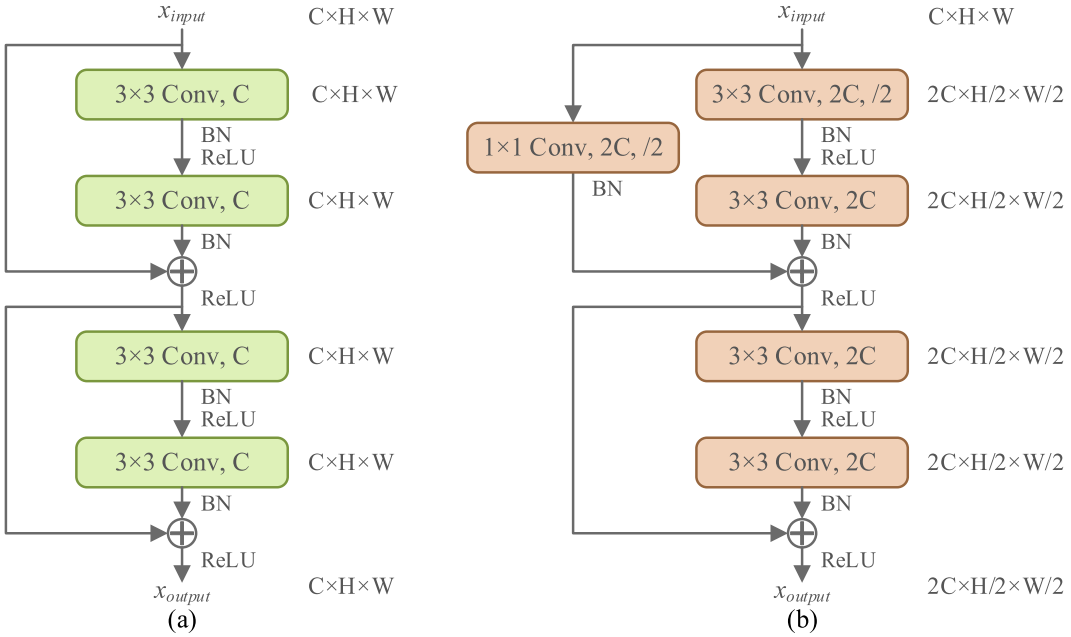


Fig. 5. Two types of residual structures in ResNet-18: (a) Maintaining feature map size, and (b) Changing feature map size.

comprising two sets of convolutional layers, batch normalization, and the ReLU activation function. For an input feature map, it first passes through the initial convolutional layer, followed by batch normalization and ReLU activation, then proceeds through the second convolutional layer and batch normalization. The feature map derived from the convolution is then added element-wise to the original input feature map, followed by a ReLU activation, yielding the output of a residual block. Stage 1 preserves the spatial dimensions of the feature map, outputting a feature map of size $64 \times 56 \times 56$. The structures of Stages 2, 3, and 4 are shown in Fig. 5 (b). Different from Stage 1, they consist of two residual blocks of distinct structures. In the first residual block, the initial convolution doubles the channel count of the feature map, while a stride-2 convolution concurrently halves its spatial dimensions. Simultaneously, a 1×1 convolution is applied to the input feature map for channel expansion and downsampling, ensuring compatibility with the output dimensions. The multi-stage architecture progressively reduces feature map resolution while increasing channel depth, with Stages 2, 3, and 4 yielding outputs of $128 \times 28 \times 28$, $256 \times 14 \times 14$, and $512 \times 7 \times 7$, respectively. The skip connections enable gradients to propagate directly from deeper to shallower layers during backpropagation, thereby enhancing network convergence and mitigating risks of gradient vanishing and overfitting.

Feature Pyramid Network (FPN) is an architecture that leverages multi-scale features to enhance object detection and classification performance. FPN employs top-down feature fusion and lateral connections to progressively refine feature maps across scales, achieving synergistic integration of high-level semantic information with low-level spatial resolution [64]. In this study, the four-stage outputs (C2, C3, C4, C5) of the ResNet18 backbone network serve as input features for FPN, with spatial resolution decreasing and semantic abstraction increasing across stages. As illustrated in Fig. 6, each stage output from ResNet-18 is first passed through a 1×1 convolution to standardize the channel dimension to 256. Subsequently, top-down upsampling is performed layer-wise, followed by element-wise summation with correspondingly scaled lower-level features for fusion. Finally, a 3×3 convolution is employed to smooth the fused features. Global average pooling is applied to the fused multi-scale feature maps to derive a 256-dimensional feature vector, which is then projected onto final category scores through a fully connected layer. The integration of ResNet and FPN preserves high-resolution shallow features while fully exploiting deep semantic information, yielding enriched feature representations for weld-defect patterns at varying scales.

To make full use of the defect information contained in the bi-surface images of solder joints, a dual-image input network architecture with weight sharing is constructed, as shown in Fig. 7. Specifically, the front and back solder joint images are fed in parallel into two structurally identical and parametrically shared F-ResNet networks devoid of fully connected layers, to extract multi-scale features from each view. After obtaining the two feature vectors, fused features are generated through channel-wise concatenation, followed sequentially by a fully connected layer, batch normalization, ReLU activation, another fully connected layer, and Softmax activation to yield class probabilities for three weld quality categories. Owing to structural symmetry and the same feature patterns between front and back solder joints, the weight-sharing mechanism ensures consistent feature learning spaces across both branches, capturing invariant quality characteristics across different perspectives. The dual-input weight-sharing architecture not only fully leverages the complementary information from both surfaces but also significantly reduces parameter count while enhancing the model's generalization capability.

3.3. Multimodal fusion framework design

Infrared image of the nugget and welding parameter data contain rich information about the welding process, serving as supplementary inputs for weld quality classification. To enhance the accuracy of welding defect identification, an ensemble deep learning approach integrating multiple classification models is proposed. Specifically, beyond the aforementioned image classification network, an Artificial Neural Network (ANN) and a convolutional neural network are constructed to classify numerical parameters and infrared images, respectively, with their architectures detailed in Fig. 8. The ANN comprises two fully connected layers with corresponding ReLU activation functions, containing 16 and 32 neurons, respectively. Input welding parameters are mapped to a 32-dimensional feature vector through feature extraction, which is then transformed into probability scores for three

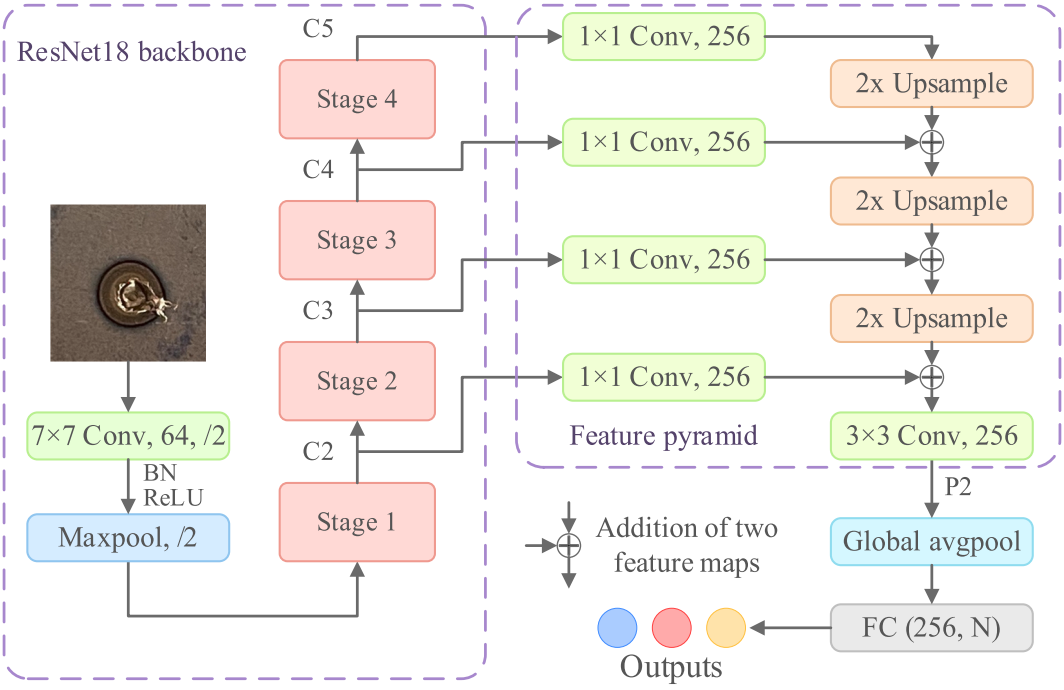


Fig. 6. FPN-enhanced Resnet (F-ResNet) network architecture.

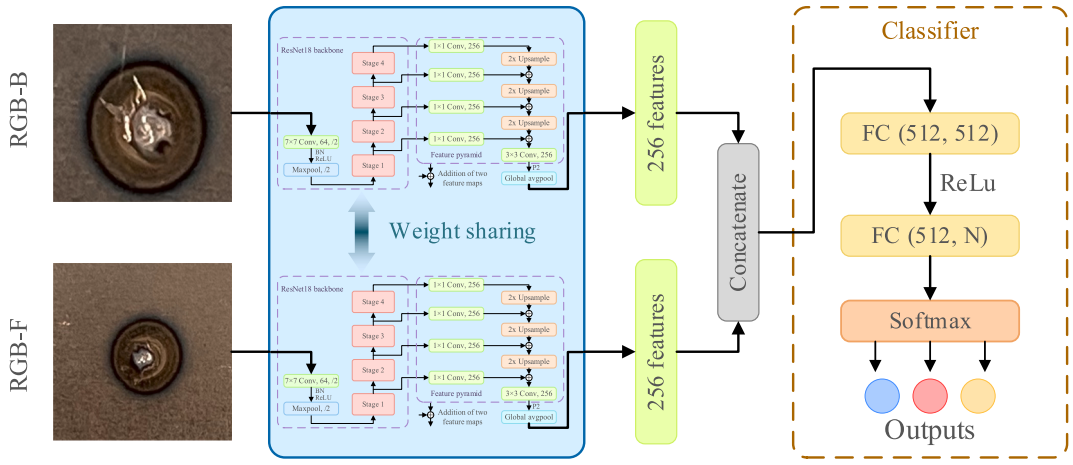


Fig. 7. Dual-input FPN-enhanced ResNet (DF-ResNet) network architecture.

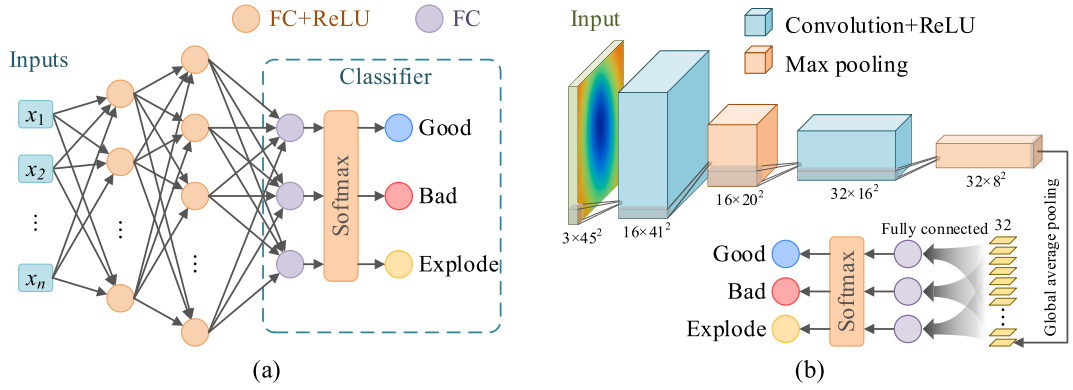


Fig. 8. Structure of ANN and IrNet: (a) ANN, and (b) IrNet.

weld quality classes via a fully connected layer and a Softmax function. The model designated for infrared image classification, named IrNet, processes the input image through two convolutional operations, ReLU activations, and max pooling layers to generate a feature map of size 32 × 8 × 8. Global average pooling is then applied to reduce this feature map to a 32-dimensional feature vector, which is finally mapped to class-specific probability scores via a fully connected layer and Softmax activation.

Although the probability scores produced by each model reflect discriminatory competence within singular data sources, they present inherent challenges for direct fusion into a unified decision. To effectively synthesize uncertainty information from heterogeneous models, this study introduces Dempster-Shafer theory to establish a multimodal classification framework, aiming to enhance the robustness and accuracy of weld quality classification. Dempster-Shafer theory, originally proposed by Dempster in 1967 and later refined by Shafer in 1976, is a mathematical framework for uncertainty reasoning under incomplete information. Its principal advantage lies in distinguishing "uncertainty" from "probability distribution," quantifying the credibility and plausibility of propositions via belief intervals rather than reliance upon singular probability estimates [65,66]. Firstly, the discernment frame is defined as $\Theta = \{\text{Good}, \text{Bad}, \text{Explode}\}$. For each classifier i, the output is a probability vector $\boldsymbol{P^i} = [p_1^i, p_2^i, p_3^i]$ over the three classes, satisfying $\sum_k p_k^i = 1, 0 \le p_k^i \le 1$. To convert these probabilities into basic probability assignments (BPAs), the mass committed by classifier i to any individual hypothesis is equated to its corresponding probability. In this study, we posit the absence of residual indeterminate mass allocated to Θ (thus $m_i(\Theta) = 0$). Consequently, each BPA m_i is defined over the power set of Θ, 2^Θ, allocating non-zero mass only to singleton sets, with all other subset masses equal to zero. As a result, the BPAs m_1, m_2 and m_3 from three independent sources are obtained, which satisfy Eq. (6):

$$\begin{cases} m_i(\varnothing) = 0 \\ \sum_{A\epsilon\Theta} m_i(A) = 1 \end{cases} \tag{6}$$

where $m_i(A)$ denotes the degree of support assigned by model i to proposition A, which may be a singleton hypothesis (e.g., {Good}) or a composite hypothesis (e.g., {Good, Bad}). To fuse the BPAs from three independent models, the Dempster combination rule is employed for sequential synthesis. Initially, any two BPAs (e.g., m_1 and m_2) are fused as follows:

$$(m_1 \bigoplus m_2)(C) = \frac{\sum_{A\cap B=C} m_1(A) m_2(B)}{1-K}, \forall C \subseteq \Theta, C \neq \varnothing \tag{7}$$

$$K = \sum_{A\cap B=\varnothing} m_1(A) m_2(B) \tag{8}$$

where $\bigoplus$ denotes the Dempster combination operator; K is the conflict coefficient, measuring the degree of contradiction between evidences; The denominator $1-K$ serves as a normalization factor, ensuring the fused result remains a valid BPA. The intermediate fusion result $m_{12} = m_1 \bigoplus m_2$ is then fused with the third BPA m_3 to yield the final fused BPA:

$$m_{123} = m_{12} \bigoplus m_3 \tag{9}$$

The fusion process in the DS theory is both associative and commutative, so the order of combination does not affect the final result. Once m_{123} is obtained, the DS theory further permits the computation of the "belief degree" (Bel) or "plausibility degree" (Pl) for any hypothesis

A. In this study, belief degree is used as the decision criterion, with the belief function defined as $Bel(C) = \sum_{D \subseteq C} m_{123}(D)$. For singleton hypotheses, the belief degree equals the corresponding mass assignment. Finally, the belief values for the three categories are compared, and the one with the highest belief degree is selected as the final decision. The DS theory substantially enhances the robustness of multi-source heterogeneous decision-making. The sequential synthesis strategy can readily accommodate the fusion of new information sources when additional modalities or classification models are introduced in the future. When evidence is highly conflicting, the value of K approaches 1, resulting in a very small $(1 - K)$ term. This may lead to unreasonable conclusions, and under such circumstances, the fusion outcome should be considered invalid. Fig. 9 illustrates the DS-based multimodal ensemble deep learning framework.

3.4. Model evaluation and interpretation

To comprehensively assess the effectiveness of the proposed framework, this section presents both quantitative evaluation metrics and interpretability analyses. While the former provides an objective measurement of the model's predictive performance, the latter elucidates the internal decision-making mechanisms and the relative influence of different modalities.

3.4.1. Model evaluation metrics

To evaluate the classification performance of the proposed method, quantitative metrics are necessary for calculation and validation. For multiclass tasks, a rigorous evaluation of a model relies on a comprehensive analysis of multiple metrics to avoid the limitations of a single perspective. This study employs four core metrics commonly used in classification tasks, including Accuracy, Precision, Recall, and F1-score.

As a fundamental metric, Accuracy reflects the proportion of correctly predicted samples, suitable for preliminary assessment of a classifier's global performance. However, it may obscure recognition deficiencies of minority classes in imbalanced datasets. To further dissect the model's fine-grained performance, Precision focuses on the proportion of actual positive instances among those predicted as positive, thereby quantifying the reliability of the model's predictions. Complementarily, Recall evaluates the model's coverage capability for positive instances by measuring the proportion of actual positives correctly identified. A higher Recall indicates a reduced risk of overlooking genuine positive instances. However, Precision and Recall often exhibit a trade-off relationship; unilateral optimization of one metric may degrade the other. To reconcile this contradiction, the F1-score is introduced as a composite evaluation metric. As the harmonic mean of Precision and Recall, the F1-score synthetically incorporates both the model's precision and coverage capability, mitigating biases associated with any single metric. A higher F1-score denotes superior balancing between precision and recall. In multiclass tasks, these metrics are typically extended through computation strategies such as Macro average, Micro average, or Weighted average. Macro average assigns equal weight to all classes for equitable evaluation; Micro average weights by sample size, emphasizing majority classes; Weighted average allocates weights proportional to the size of categories to accommodate varying data imbalance characteristics. Given that the preprocessing procedures have balanced sample quantities across categories, this study employs the Macro average for metric computation. These evaluation metrics are computed according to Eqs. (10) to (13):

$$Accuracy = \frac{\sum_{i=1}^{C} TP_i}{N} \tag{10}$$

$$Precision = \frac{1}{C}\sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i} \tag{11}$$

$$Recall = \frac{1}{C}\sum_{i=1}^{C} \frac{TP_i}{TP_i + FN_i} \tag{12}$$

$$F1 = \frac{1}{C}\sum_{i=1}^{C} 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{13}$$

where C denotes the total number of categories; N is the total sample count; TP_i, FP_i and FN_i represent the true positive, false positive, and false negative counts for category i, respectively.

Industrial defect detection focuses not only on the overall classification capability of algorithms but also emphasizes their performance in reliability and efficiency. Reliability concerns the system's ability to stably and accurately identify all defects, with key metrics including defect detection sensitivity and miss detection rate (MDR). The cost of a

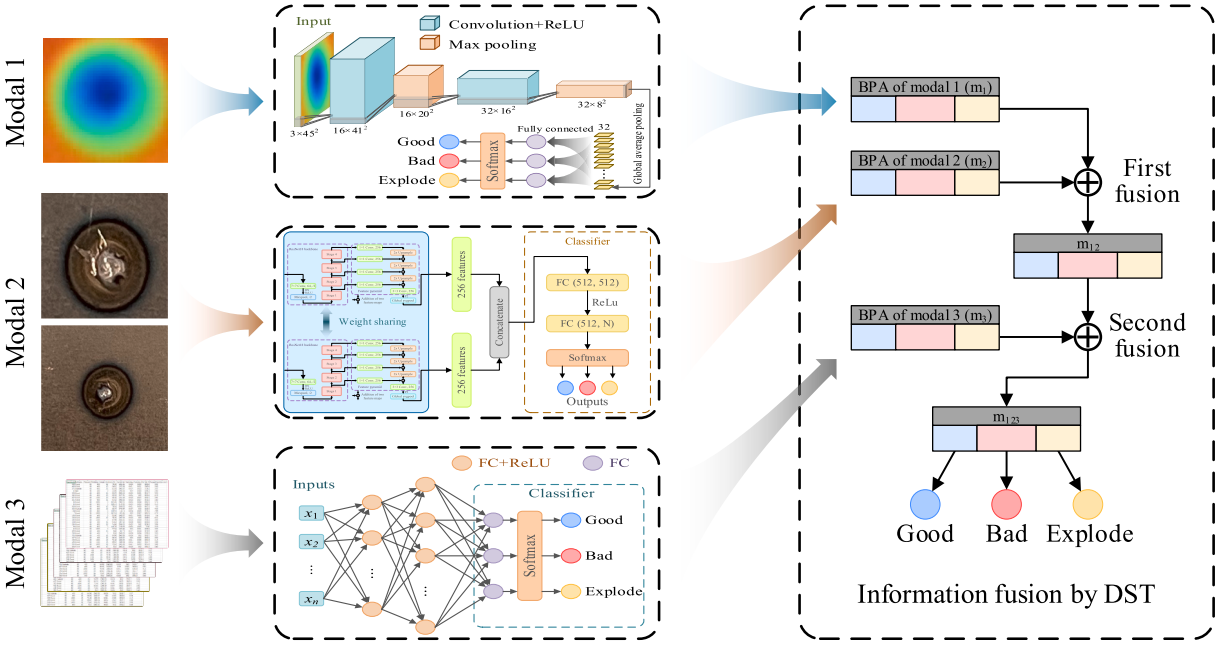


Fig. 9. Ensemble multimodal deep learning framework.

missed detection is typically far greater than that of a false alarm, creating a strong demand for high sensitivity. Efficiency pertains to the speed of detection. An excessively high false alarm rate (FAR) leads to numerous unnecessary re-inspections, increasing labor costs and production downtime. Categorizing all defects into a single class, the detection sensitivity, MDR, and FAR are calculated using Eqs. (14) to (16).

$$Sensitivity = \frac{TP}{TP + FN} \tag{14}$$

$$MDR = 1 - Sensitivity\# \tag{15}$$

$$FAR = \frac{FP}{FP + TN} \tag{16}$$

where TP, FP, TN, and FN represent the counts of true positive, false positive, true negative, and false negative samples for the defect category, respectively.

3.4.2. Model interpretation and modal contribution quantification

In multimodal tasks, model interpretability is critical for understanding the decision-making process and the extent to which different modalities influence those decisions. To thoroughly analyze model behavior when processing multimodal inputs such as images and numerical parameters, this study introduces two complementary interpretation methods: Grad-CAM and MM-SHAP. Firstly, MM-SHAP is used to quantify each modality's contribution to the final classification, and then Grad-CAM is applied to visualize the regions of interest in the DF-ResNet during image processing, thereby providing a comprehensive evaluation of the model's multimodal fusion capability.

Deep learning models are often regarded as “black boxes,” and their internal decision-making mechanisms are difficult to understand. Grad-CAM generates heatmaps that visually highlight the image regions attended to by the convolutional network during a specific class prediction, thus revealing the model's decision rationale [30]. The core concept involves weighting feature maps from the final convolutional layer using gradient information of target classes to obtain importance weights for each feature map relative to the target class, accentuating spatial regions that significantly influence classification. Specifically, given an input image and a target class c, let the output feature maps of the last convolutional layer be $A^k \in \mathbb{R}^{u\times v}$ $(k = 1, 2, \ldots, K)$, and denote the prediction score for class c as y^c. Grad-CAM first computes gradients $\frac{\partial y^c}{\partial A^k_{ij}}$ via backpropagation, which reflects the sensitivity of each position (i, j) in feature map A^k to class c. Subsequently, global average pooling is applied to derive the channel-wise weights:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{17}$$

where Z denotes the feature map size $(Z = u \times v)$; The weight α_k^c characterizes the importance of channel k to class c. A channel-weighted summation is then applied to the feature map, followed by a ReLU activation to retain positive contributions, producing the Grad-CAM heatmap for class c:

$$L^c_{Grad-CAM} = \mathrm{ReLU}\left(\sum\nolimits_k \alpha_k^c A^k\right) \tag{18}$$

The heatmap is upsampled to the input size and superimposed on the original image, visually displaying the regions that the model relies on for its decision. Warm colors indicate high-contribution regions, while cool colors denote low-contribution areas. Notably, Grad-CAM requires no architectural modifications or retraining and can be broadly applied across diverse CNN frameworks, including image classification and object detection tasks.

However, Grad-CAM primarily focuses on visual interpretation of the image modality, highlighting the importance of distinct features on images, yet it cannot quantify the model's reliance on different modalities. To address this limitation, we introduce the MM-SHAP method to further analyze the modality-dependent behaviors in multimodal data processing [32]. MM-SHAP is a performance-agnostic metric designed to quantify the proportional contribution of each modality to the final prediction in multimodal models, with its computational workflow illustrated in Fig. 10. MM-SHAP extends the classical Shapley value to support joint explanations of multimodal data. Grounded in game-theoretic Shapley value theory, MM-SHAP treats input tokens as participants in a cooperative coalition and evaluates modality importance by computing their marginal contributions to model outputs through combinatorial analysis. The purpose of token assignment is to balance feature quantities across modalities, ensuring comparability of Shapley values and thus enabling an objective comparison of each modality's contribution. The setting of tokens is based on the numerical parameter modality: if there are n_N numerical features, then n_N tokens are allocated to that modality, while each image modality is assigned $\lceil\sqrt{n_N}\rceil^2$ tokens.

MM-SHAP computes the Shapley value for each input token. For an input containing n tokens, the Shapley value ϕ_j of the j-th token is calculated by:

$$\phi_j = \sum\nolimits_{S\subseteq\{1,\ldots,n\}\setminus\{j\}} \frac{|S|!(n - |S| - 1)!}{n!}\left(\mathrm{val}(S\bigcup\{j\}) - \mathrm{val}(S)\right) \tag{19}$$

where S denotes a subset of tokens; $\mathrm{val}(S)$ represents the model output for subset S (non-subset tokens are masked); The baseline $\mathrm{val}(\varnothing)$ is the output when all tokens are masked. In practice, Shapley values are approximated by using a Monte Carlo procedure to generate random permutations of features or data points, computing each point's marginal contribution under each permutation, and averaging those contributions. Modality-level contributions are obtained by aggregating the Shapley value of tokens. The contribution Φ_i of modality i is defined as:

$$\Phi_i = \sum_{j=1}^{n_i} \left|\phi_j(i)\right| \tag{20}$$

where n_i denotes the number of tokens in modality i; $\phi_j(i)$ represents the Shapley value of the j-th token in modality i. The absolute value operation disregards contribution direction (enhancing or suppressing prediction), focusing on the activity strength of the modality in specific class predictions. Finally, the normalized contribution of modality i is calculated by Eq. (21):

$$\rho_i = \frac{\Phi_i}{\sum_k \Phi_k} \tag{21}$$

where ρ_i represents the proportion of information contributed by the i-th modality in the process of the model obtaining the output.

The workflow described above is implemented on a Windows 10 platform using Python 3.10.16, with PyCharm Professional 2023.3.4 serving as the IDE. Image enhancement processing is conducted using Opencv 4.11.0.86, while Pandas 2.2.3 supported numerical data input and augmentation operations. The deep learning network is then built and trained using Torch 2.0.0. Finally, model evaluation metrics are computed via interfaces provided by Scikit-learn 1.6.1, and model interpretation methods are implemented using Grad-cam 1.5.5 and Shap 0.42.0.

4. Experimental study

To validate the feasibility of the proposed approach for welding quality classification, this study conducts an experimental verification using a multimodal resistance spot welding dataset. Different welding types share certain similarities, particularly in the influence patterns of welding processes and joint design on weld quality [67]. Dual-surface images can be extended to images from different perspectives, while

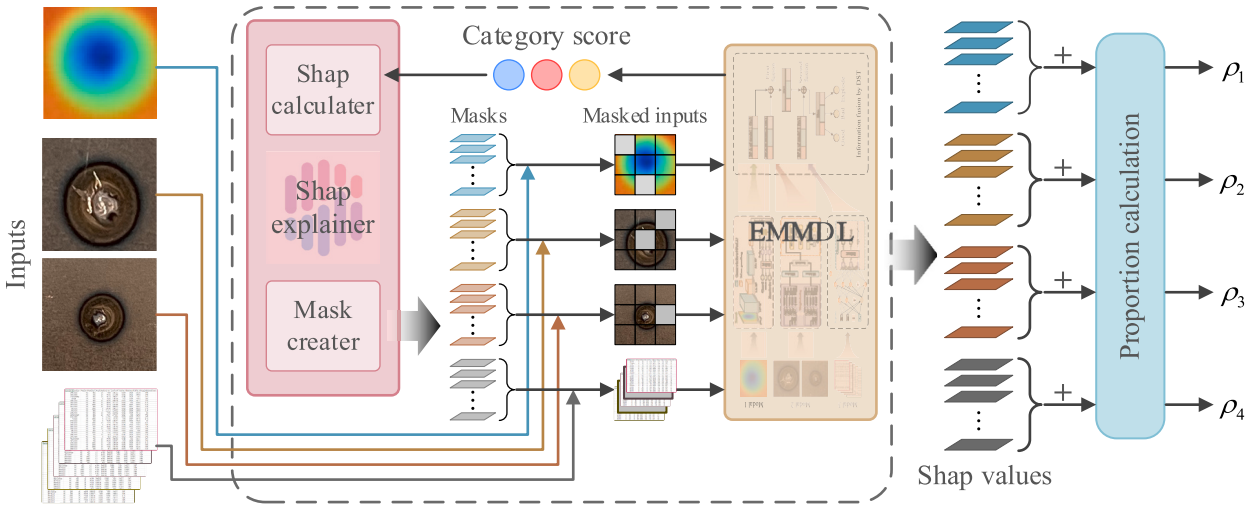


Fig. 10. MM-SHAP calculation process.

infrared images reflect the heat distribution across various welding scenarios. Meanwhile, different welding parameters also influence the formation of defects. This makes the resistance spot welding dataset highly representative for weld quality classification tasks. Details regarding the dataset's basic information, data preprocessing, model training and evaluation, and result analysis are presented in Section 4.1 Data and resources, 4.2 Model development, and 4.3 Results analysis, respectively.

4.1. Data and resources

The welding multimodal data used comes from a public dataset provided by Luis Alonso Dominguez Molina: Resistance Spot Welding Insights [68]. This dataset comprehensively records the RSW process, capturing seven categories of input parameters: Pressure, Welding time, Electrode angle, Electrode force, Welding current, Material thickness, and Material type, which pertain to the welding process parameters employed and material characteristics during nugget formation. Since all samples utilize an identical Material Type, this feature is excluded from this study. During the welding cooling phase, thermal images of the nugget and surrounding area were captured using an infrared camera from a 10 cm distance, while surface images were acquired from both the front and back sides of the weld spot at a 15 cm distance. Subsequently, tensile testing was conducted using a Tinus Olsen testing machine with a maximum capacity of 300 kN and a resolution of 0.1 N, followed by measurement of the nugget diameter using an electronic caliper of 0.01 mm resolution. Finally, weld quality was classified into three categories: Good, Bad, and Explode. By varying the electrode force, electrode angle, and welding time, 495 welds were performed, yielding a dataset of 495 samples. Table 1 provides statistical descriptions of the parameters, including inputs and outputs, while Fig. 11 presents representative examples of surface images and infrared images of nuggets across the different quality categories.

The original dataset exhibits a pronounced class imbalance. Specifically, among the 495 samples, 443 are labeled as “Good,” accounting for 89.5 %, while only 21 and 31 samples are labeled as “Bad” and “Explode,” respectively. Such an imbalance severely hinders the model's learning capability. Therefore, augmentation strategies based on affine transformations and noise injection are applied to both image data and numerical parameters to enhance and expand the dataset. Each “Bad” sample is augmented 19 times, and each “Explode” sample 13 times, with augmented data retaining the original class labels. For each augmentation of RGB and infrared images, the outer 10 % regions of all edges are first extracted and extended by mirror reflection, expanding the canvas to twice the original size. Next, the following random geometric transformations are applied in sequence: random horizontal and vertical flips; a rotation sampled uniformly from -180° to 180°; random scaling in the range 1.0–1.5×; and random translation up to ± 20 % of the image size. All performed on the enlarged canvas to avoid boundary loss, after which a center crop matching the original image size is extracted. For numerical data augmentation, each original value is multiplied by a Gaussian random variable drawn from $N(1, 0.1^2)$. Finally, the number of samples in the three categories became 443, 420, and 434, respectively. The implemented strategies of affine transformation and noise addition not only expanded the sample size but also effectively simulated potential morphological variations and parameter fluctuations in real-world scenarios. This approach enabled the augmented minority-class samples to cover a more comprehensive feature space, thereby mitigating the risk of limited generalization capability caused by insufficient minority samples and promoting more balanced feature learning across all defect categories by the model.

Table 1
Description of welding parameters.

Type	Variable	Description	Unit	Range
Input	Pressure	Pressure on the pneumatic cylinder	PSI	{35, 60, 80, 95}
Input	Welding time	Welding process time	ms	200–1500
Input	Electrode angle	Angle between the electrodes	Deg	{0, 15}
Input	Electrode force	Force applied to the electrodes	N	0–133.53
Input	Welding current	Current flow through the metal sheet	A	639.81–5009.43
Input	Material thickness A	Thickness of the material A	mm	0.61–1.057
Input	Material thickness B	Thickness of the material B	mm	0.608–1.01
Output	Pull test force	Mechanical resistance of the welding joint	N	1410.3–5806.5
Output	Nugget diameter	Diameter of the welding nugget	mm	1.9–4.72

For deep learning models, although multicollinearity does not directly undermine their predictive ability, it still has negative effects. Firstly, highly correlated features may cause severe weight oscillations, make optimization difficult to converge, and affect the reliability of the weight matrix. Additionally, correlated features tend to be redundantly encoded in hidden layers, making it difficult for the model to distinguish their contributions and reducing the interpretability of feature importance. Moreover, redundant features increase network complexity

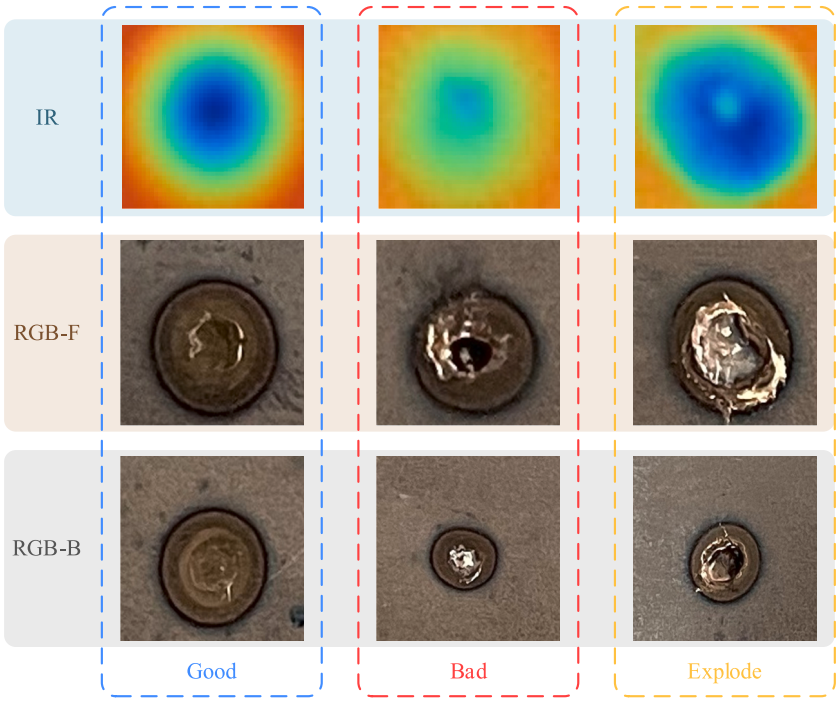


Fig. 11. Typical examples of images in the three categories.

without contributing new information, leading to wasted computational resources and slower training. Therefore, multicollinearity mitigation is performed on the nine variables listed in Table 1 based on Eq. (5). Firstly, we calculate the multicollinearity across all variables. The variable with the most severe multicollinearity (with the largest VIF) will be selected and removed. Then we calculate the VIF for the remaining variables and remove the one with the most severe multicollinearity. This process is repeated until all variables meet the multicollinearity threshold (VIF $<$ 10). The multicollinearity elimination process is detailed in Table 2. Once the variables are reduced to five: Welding time, Electrode angle, Electrode force, Welding current, and Material thickness, no severe multicollinearity remained, and these are used as the inputs for the ANN network.

4.2. Model development

The augmented dataset comprises 1297 samples, each consisting of three distinct data modalities and a categorical label (an infrared image of the weld nugget, two digital images of the front/back surfaces, and a set of welding parameters). These samples are split into training and test sets at a 7:3 ratio. To preserve the class distribution in the dataset, a label-stratified sampling is adopted, i.e., splitting the training and test sets within each group of identical labels. Moreover, to prevent contamination between training and test sets due to random sampling, all augmented samples and original samples corresponding to identical instances are grouped as indivisible units during sampling. The divided training set contains 917 samples, and the test set contains 380 samples. The specific distribution is listed in Table 3. Surface digital images are resized to 224 $\times$ 224 pixels and converted into numerical arrays with RGB channels to meet the input size requirements of the feature extractor. Subsequently, to optimize gradient propagation and accelerate model convergence, pixel values across all channels are normalized to the range [−1,1]. For infrared images, given their lower resolution (45 $\times$ 45) and customized design of their feature extraction network, it is sufficient to convert them into three-channel numerical arrays, followed by normalization. Numerical parameters are standardized using the training set's mean and standard deviation, rendering each feature approximately standard normal.

The multimodal classification model is trained using a staged training strategy. First, the IrNet, DF-ResNet, and ANN networks are trained separately on infrared images, surface images, and numerical parameter data, respectively. The ResNet backbone is initialized with PyTorch-provided weights pretrained on the ImageNet dataset, while the remaining network components are initialized with the default random initialization strategy. Upon completing the training of the three base models, the DS theory is used to combine their outputs, forming the final ensemble model. Computations are performed on a desktop equipped with 12th Gen Intel(R) Core (TM) i7-12700KF CPU, NVIDIA GeForce RTX 4060 GPU, and 32 GB RAM. The corresponding computational environment and software configurations are summarized in

Table 2
Process of multicollinearity treatment.

Variable	VIF				
	Step 1	Step 2	Step 3	Step 4	Step 5
Pressure	15.76	15.40	15.16	**15.15**	/
Welding time	7.76	7.76	7.61	5.75	5.45
Electrode angle	3.67	3.67	3.54	2.82	2.74
Electrode force	15.46	14.94	13.07	10.00	8.12
Welding current	13.60	13.54	10.64	9.85	9.71
Material thickness A	**603.48**	/	/	/	/
Material thickness B	602.69	30.95	28.03	13.45	4.93
Pull test force	59.84	59.84	**40.69**	/	/
Nugget diameter	76.83	**76.66**	/	/	/

Note: The bold numbers in the table represent the maximum VIF of the current step.

Table 3
Distribution of samples before and after augmentation.

Category	Original			Augmented		
	Train	Test	Total	Train	Test	Total
Good	309	134	443	309	134	443
Bad	15	6	21	300	120	420
Explode	22	9	31	308	126	434
Total	346	149	495	917	380	1297

Table 4
Computational environment and software configuration.

Component	Version
CPU	12th Gen Intel(R) Core (TM) i7-12700KF
GPU	NVIDIA GeForce RTX 4060
PyCharm	PyCharm Professional 2023.3.4
Python	3.10.16
Opencv	4.11.0.86
Torch	2.0.0
CUDA	11.8
Numpy	1.26.4
Pandas	2.2.3
Scipy	1.15.2
Scikit-learn	1.6.1
Grad-cam	1.5.5
Shap	0.42.0

Table 4. To prevent out-of-memory problems during training, the batch size is set to 64. Stochastic gradient descent (SGD) is used for parameter optimization, with momentum and weight decay empirically set to 0.9 and 5×10^{-5}, respectively, to ensure stability during training. Cross-entropy is adopted as the loss function, with Accuracy, Precision, Recall, and F1-score serving as evaluation metrics. Table 5 provides the detailed hyperparameter configurations. For reproducibility, all random seeds are set to 42. To further enhance the model's generalization, a 30 % random dropout is applied to the extracted features during training. Dropout and weight decay are common regularization techniques in deep learning: dropout randomly disables neurons during training to force the network to learn more robust features, whereas weight decay adds the squared weights as a penalty to the loss function to mitigate overfitting to noise in the training data. The three base models are trained for 100 epochs, with training durations of IrNet: 7.14 s; DF-ResNet: 748.98 s; ANN: 5.79 s. Fig. 12 illustrates the evolution of evaluation metrics for the three base classifiers during training.

The trained base models are used for the comprehensive classification of RSW quality. According to Eqs. (6) to (9), the model outputs are normalized via Softmax and can be directly converted into basic probability assignments for different categories of the three classifiers. The three sets of basic probability assignments are then sequentially fused using Dempster's combination rule to generate the prediction probability scores of the EMMDL model. The category with the highest predicted probability score is selected as the final prediction of the ensemble model. During evidence aggregation, the conflict coefficient K is computed by summing the products of probabilities for all pairs of focal elements whose intersection is not empty. If $1 - K < 0.001$, the evidence is considered fully conflicting and the fusion process is terminated. For cases that can be fused, a normalization factor $1/(1-K)$ is applied to reassign the belief mass of all non-conflicting focal elements. No filtering threshold is imposed on belief values during this process, so all non-zero beliefs participate in the fusion. In our testing, the average inference time for a single sample within the multimodal detection framework is 6.14 milliseconds, meeting the real-time requirements for industrial inspection scenarios. The complete training and fusion process is presented in Algorithm 1.

Algorithm 1. EMMDL model construction.

Input: Infrared images X_{ir}, Surface images X_{rgb}, Welding parameters X_p.

Output: Predicted class y $\in$ {Good, Bad, Explode}

Procedure:
- Data Preprocessing:
  - Augment images using affine transformations: Rotate, Scale, Flip.
  - Add Gaussian noise to welding parameters to enhance diversity.
  - Normalize images and standardize welding parameters.
- Train Base Classifiers:
  - Train $f_{ir}(X_{ir})$ using IrNet, output probabilities p_{ir}.
  - Train $f_{rgb}(X_{rgb})$ using DF-ResNet, output probabilities p_{rgb}.
  - Train $f_p(X_p)$ using ANN, output probabilities p_p.
- Fusion via Dempster-Shafer Theory:
  - Convert p_{ir}, p_{rgb}, p_p to basic probability assignments (BPA).
  - Fuse BPAs using Dempster's combination rule to get a fused BPA.
  - Compute belief degrees: Bel(A) for each class.
  - Select the class with the highest Bel(A) as the final prediction $\hat{y}$.
- Evaluation:
  - Calculate Accuracy, Precision, Recall, and F1-score using standard formulas.
  - Compute Macro averages for multi-class metrics.
- Interpretation:
  - Grad-CAM: Visualize important image regions affecting prediction.
  - MM-SHAP: Quantify contribution of each modality (image, parameter).
- Return: Predicted class $\hat{y}$

End Procedure

Table 5
Details of the hyperparameter settings.

Hyper Parameters	Value
Batch size	64
Learning rate	0.0005
Epochs	100
Optimizer	SGD
Weight decay	5×10^{-5}
Loss function	Cross-Entropy Loss
Early stopping	No early stopping
Evaluation metric	Accuracy、Precision、Recall and F1-score

4.3. Results analysis

This section provides a comprehensive performance analysis and interpretation of the proposed ensemble learning model on the test dataset. The model's performance is assessed using Accuracy, Precision, Recall, and F1-score, as shown in Fig. 13. To fully demonstrate the superiority of the ensemble model over individual models, Fig. 15

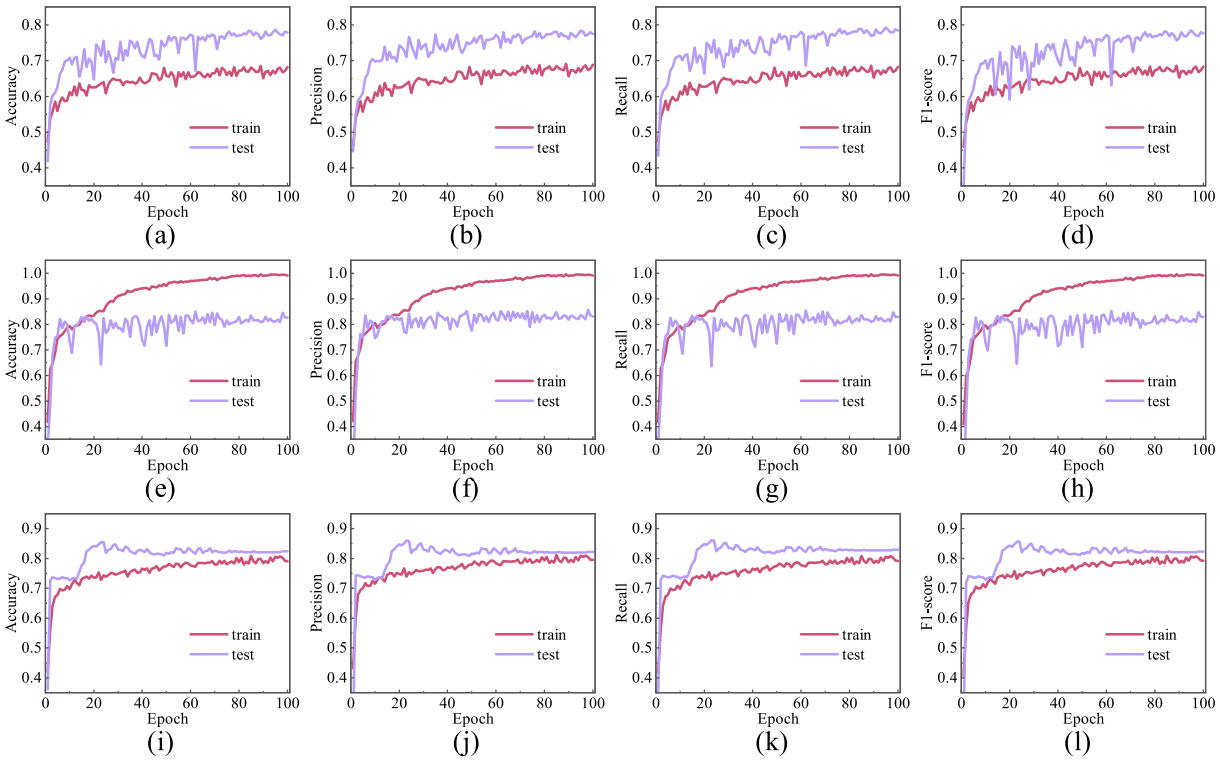


Fig. 12. Evolution of evaluation metrics: (a) IrNet accuracy; (b) IrNet precision; (c) IrNet recall; (d) IrNet F1-score; (e) DF-ResNet accuracy; (f) DF-ResNet precision; (g) DF-ResNet recall; (h) DF-ResNet F1-score; (i) ANN accuracy; (j) ANN precision; (k) ANN recall; (l) ANN F1-score.

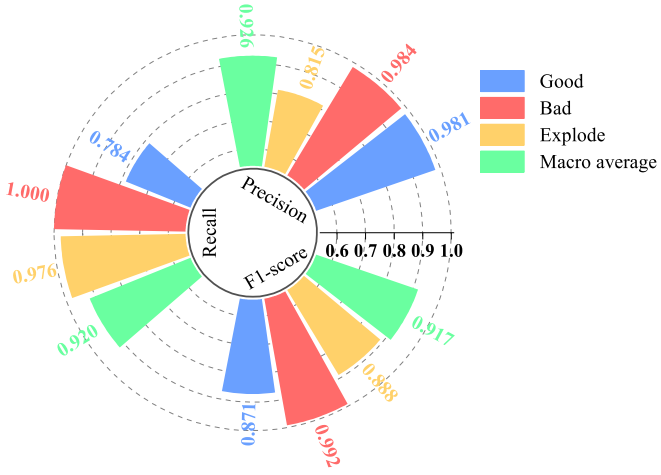


Fig. 13. Classification performance of the ensemble deep learning model.

Table 6
Comparison of evaluation metrics across models.

Model	Accuracy	Precisoin	Recall	F1-score
IrNet	0.787	0.784	0.793	0.785
DF-ResNet	0.850	0.851	0.854	0.852
ANN	0.855	0.858	0.860	0.857
EMMDL	0.916	0.926	0.920	0.917

compares the confusion matrices of three base models and the ensemble model on the test set, while Table 6 calculates the evaluation metrics for them and Fig. 16 provides a visual comparison. A series of ablation experiments is conducted to further verify the effectiveness of the proposed FPN-enhanced dual-input weight-sharing network, with results shown in Fig. 17. To evaluate the ensemble model's utilization of information from each modality, Fig. 18 and Fig. 19 employ MM-Shap to quantify modality importance. In this study, the numerical modality has dimensionality 5; consequently, MM-Shap assigns five tokens to the numerical modality and nine tokens per input image. Under these settings, computing the MM-Shap value for a single sample requires on average 2.35 s. Based on Grad-cam, the model's interest regions in the image during classification are visualized, as shown in Fig. 20. Detailed results analysis is presented below.

(1) The proposed method can detect defects in resistance spot welding with high accuracy. As shown in Fig. 13, although the model's classification performance varies across different categories, after macro averaging, the ensemble model achieves a

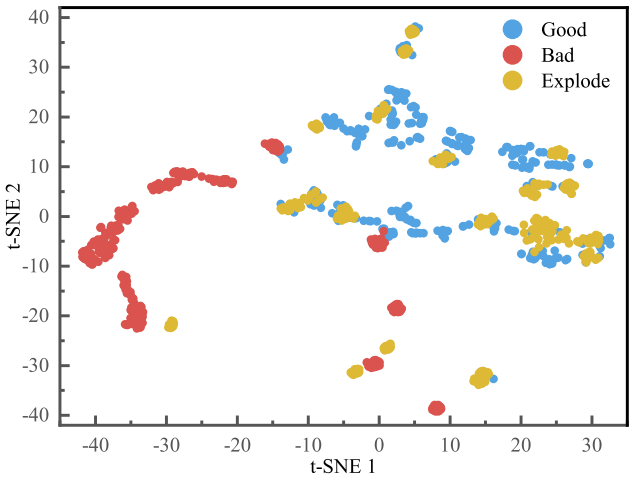


Fig. 14. Results of t-SNE analysis for numerical parameters.

precision, recall, and F1-score of 0.926, 0.920, and 0.917, respectively, and an overall accuracy of 0.916. Notably, for samples labeled "Bad", the ensemble model achieves a precision of 0.984, a recall of 1.000, and an F1-score of 0.992, indicating that the model accurately extracts and identifies the characteristics of these samples. On the other hand, for samples labeled "Good", while Precision is also high at 0.981, Recall is only 0.784, resulting in an F1-score of 0.871. Meanwhile, for samples labeled "Explode" in the test set, Precision is only 0.815, but Recall reaches 0.976. This indicates that some samples labeled "Good" are incorrectly predicted as "Explode", which is the main source of the model's classification errors. A comparison of the surface images of the three sample categories reveals that the "Bad" samples exhibit notably darker colors and coarser textures, demonstrating a clear distinction from the other two categories. However, the "Good" and "Explode" samples share similar color and textural characteristics, with the only difference lying in the presence of splattering. Fig. 14 presents an analysis of the numerical parameters using t-SNE, which also illustrates the characteristic that "Bad" samples form an independent cluster while the other two categories exhibit significant overlap. When confronted with ambiguous samples that are difficult to classify, the model evidently adopts a conservative strategy, leaning toward assigning them to "Explode". With respect to industrial metrics, Sensitivity, MDR, and FAR are 99.2 %, 0.8 %, and 21.6 % respectively, indicating high sensitivity for defect detection. In terms of runtime, single-sample inference takes only 6.14 ms, corresponding to a theoretical throughput of approximately 162 fps, which is sufficient for real-time detection. Overall, the model achieves 0.916 accuracy for RSW multimodal data, with Precision, Recall, and F1-score maintained at high levels, reflecting the effectiveness and reliability of the proposed ensemble multimodal deep learning method in the RSW quality recognition task.

(2) The classification performance of the EMMDL model shows significant improvement compared to individual modalities. As shown in Table 6 and Fig. 16, among the basic unimodal classification networks, IrNet, DF-ResNet, and ANN achieve Accuracies

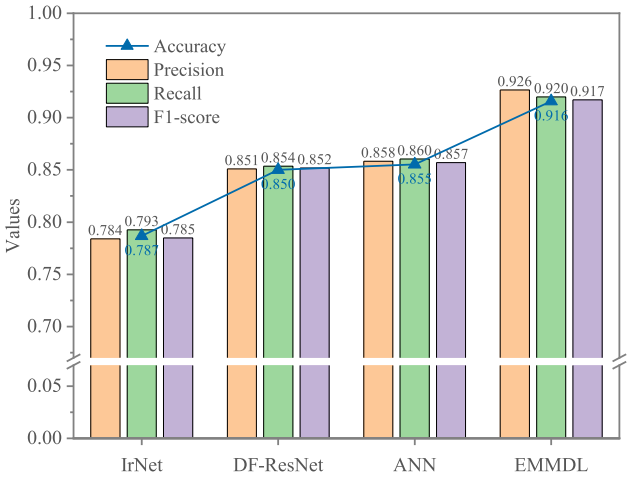


Fig. 16. Comparison of unimodal and multimodal metrics.

of 0.787, 0.850, and 0.855 on the test set, respectively. However, the accuracy of the ensemble classification model reaches 0.916 after using Dempster-Shafer evidence theory for information fusion. Fig. 15 shows the confusion matrices of the three base models and the ensemble model, providing a clearer view of the correct and incorrect predictions for each category. It is noteworthy that in the base models, there are numerous instances where samples labeled "Explode" are incorrectly identified as "Good", indicating that the base models exhibit missed detections. However, the ensemble model misclassifies only two "Explode" samples as "Good", indicating that the ensemble greatly improves the missed detection of welding defects. Meanwhile, the similarly frequent misclassification of "Good" samples as "Explode" does not lead to serious consequences in practical applications, and it effectively ensures that welding defects are accurately detected. The standalone IrNet network captures the distribution features of the thermal field during welding to identify anomalous heating regions and infer the

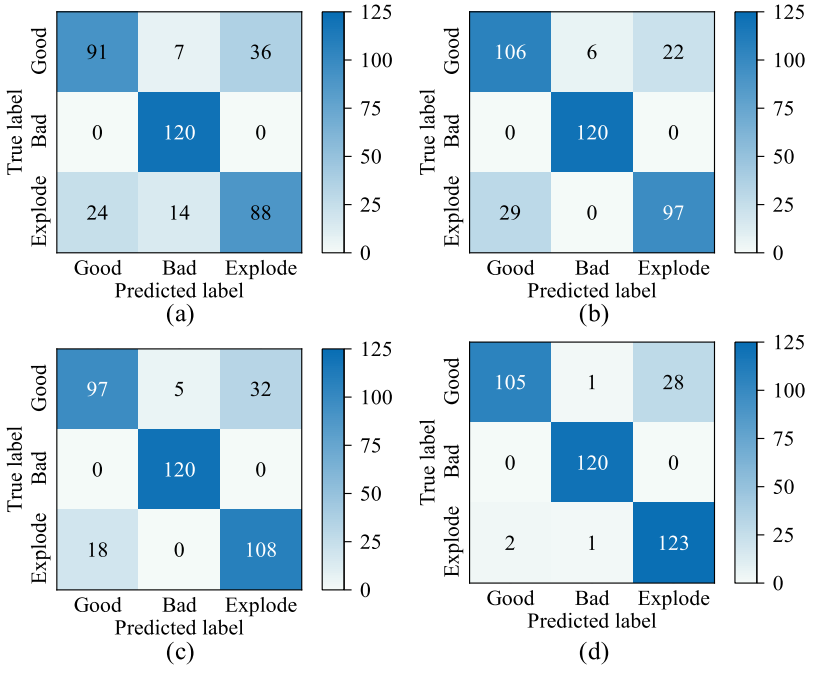


Fig. 15. Confusion matrices of (a) IrNet, (b) DF-ResNet, (c) ANN, and (d) EMMDL.

Table 7
Ablation experiments for model components.

Model	FPN	Double input	Weight sharing	Accuracy	Precisoin	Recall	F1-score
ResNet				0.788	0.804	0.790	0.787
D-ResNet		√	√	0.816	0.824	0.818	0.818
F-ResNet	√			0.792	0.797	0.796	0.794
DF-ResNet (without weight sharing)	√	√		0.821	0.842	0.822	0.818
DF-ResNet	√	√	√	0.850	0.851	0.854	0.852

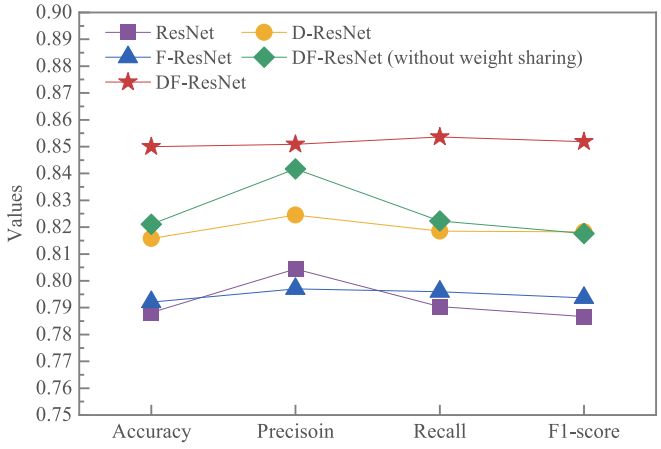


Fig. 17. Comparison of evaluation metrics in ablation experiments.

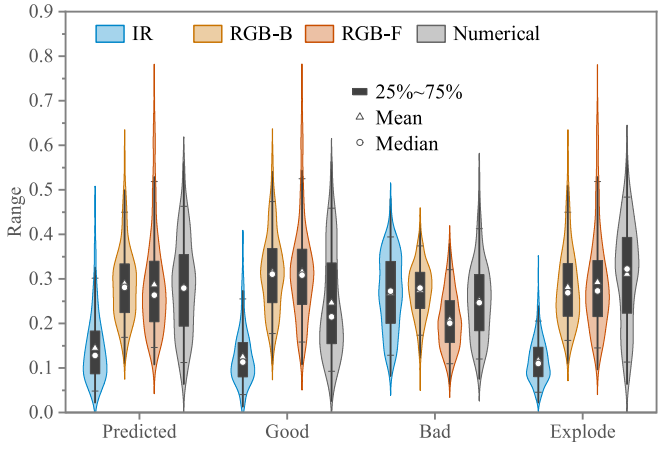


Fig. 18. Contribution proportions of different modal information in model predictions.

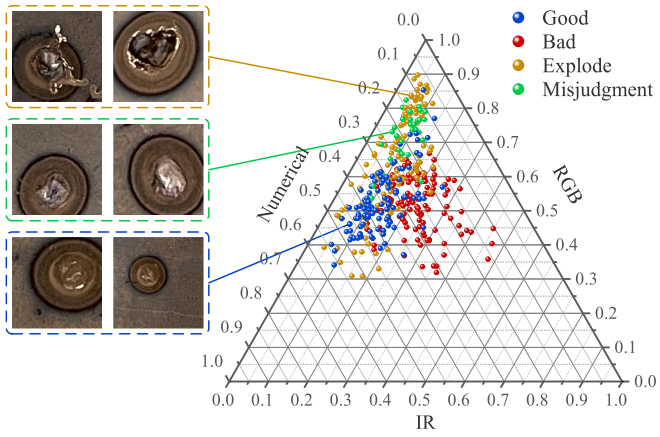


Fig. 19. Modal contribution distribution of test data.

geometry of the weld nugget, correlating the energy field patterns around the weld with welding quality. The DF-ResNet extracts morphological structures and texture features of the weld surface based on digital images of the weld's front and back surface post-welding, and the ANN captures the latent influence of RSW's input and output parameters on welding quality. The ensemble model fuses the predictions of the three base models, comprehensively considering the infrared images, digital images, and welding input-output parameters to accurately discriminate the quality of resistance spot welding.

(3) The targeted network architecture design effectively enhances the model's feature extraction ability for surface images. To fully demonstrate the individual contributions of the FPN mechanism, dual-image input structure, and weight-sharing strategy to feature extraction of surface images, we constructed four variants: a baseline ResNet, a dual-input weight-sharing network based on ResNet (D-ResNet), an FPN-enhanced ResNet (F-ResNet), and an FPN-enhanced dual-input network without weight sharing. All models are trained on the preprocessed dataset using the same hyperparameters listed in Table 5. For single-input networks, the RGB-B and RGB-F datasets are merged to augment the sample size. The evaluation metrics on the test set are presented in Table 7, and Fig. 17 visualizes their comparison via line plots. It can be observed that the conventional ResNet achieves only 0.788 accuracy and an F1-score of 0.787. Incorporating the FPN module raises accuracy and F1-score to 0.792 and 0.794, respectively, which is a smaller gain than that achieved by D-ResNet, where accuracy and F1-score climb to 0.816 and 0.818, representing improvements of 3.55 % and 3.94 %. The dual-input structure provides the model with richer visual information from resistance spot welds, enabling the extraction of additional features for classification, especially when defects exist only on one surface. Additionally, when feature extraction networks did not share weights (using two structurally identical but independently weighted F-ResNet backbones), Accuracy reaches 0.821, while Precision attains 0.842. In this scenario, each backbone is trained exclusively on either RGB-B or RGB-F images. Given that RGB-B and RGB-F images share similar characteristics despite differing capture positions, the weight sharing strategy unquestionably enables more effective use of limited data to train the feature extractor. Consequently, DF-ResNet attains an accuracy of 0.850 and an F1-score of 0.852, representing gains of 7.87 % and 8.26 % over the base ResNet. The targeted network design enables full and rational utilization of data, enabling rapid and precise assessment of welding quality.

(4) The interpretability analysis reveals that the ensemble model makes comprehensive use of information from diverse modalities, while its misclassifications are associated with the model's referencing biases toward different modal features. MM-Shap calculates the contribution ratios of different data modalities to predictions by tracking feature activation levels in each input when predicting specific classes. As shown in Fig. 18, when

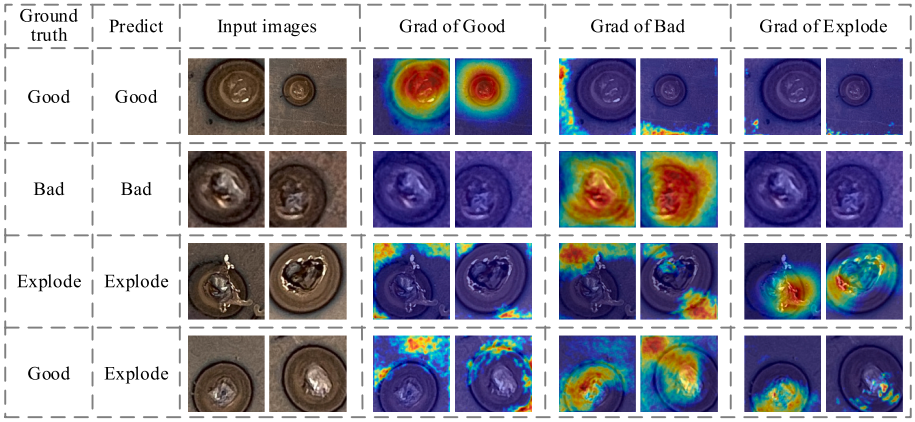


Fig. 20. Grad-Cam analysis of different target categories.

predicting the category with the highest probability score, attributes average contribution scores of 0.1446, 0.2887, 0.2868, and 0.2799 to the IR, RGB-B, RGB-F, and Numerical inputs, respectively, indicating that multiple data sources inform its predictions. Notably, the model relies more heavily on infrared image data when computing the prediction probability for "Bad" than for "Good" or "Explode", suggesting that "Bad"-relevant information may primarily reside in thermal distribution around solder joints. This implies that improving the quality and resolution of infrared images in subsequent data collection may help enhance the recognition of the "Bad" samples. Fig. 19 more clearly displays the contribution distribution across predicted classes, with RGB-B and RGB-F contributions aggregated as the comprehensive RGB modality contribution. It can be observed that contribution clusters vary with predicted class: "Good" predictions preferentially draw on numerical inputs, "Bad" predictions favor infrared images, and "Explode" predictions prioritize surface images. Additionally, for samples misclassified from "Good" to "Explode", the model shows increased preference for RGB modality information. According to Fig. 20, accurate predictions rely on focusing on solder joint morphology. For example, the model attends to the spatter region surrounding the weld when correctly identifying "Explode" samples. Misclassification may be influenced by the misleading textures surrounding the solder joints, where white regions exhibit color characteristics similar to those of splash areas, potentially leading the model to assign an "Explode" prediction. Therefore, introducing attention-based constraint mechanisms to direct the model toward informative regions rather than distracting features, or augmenting training with negative samples that exhibit similar distracting features, is able to strengthen the model's discriminative ability. The combined use of MM-Shap and Grad-Cam provides a comprehensive insight into the basis for the model to make classifications.

5. Discussion

This paper proposes an ensemble multimodal deep learning approach for the identification of resistance spot welding quality. A DF-ResNet is designed to simultaneously process two input images using a weight-shared backbone network for feature extraction on images captured from both front and back surfaces of weld spots post-welding. Experimental results show that the dual-input weight-sharing network performs well for surface image feature extraction, improving the recognition effect by 7.87 %. Multimodal fusion further improves it by 7.76 % on this basis. Although the current model achieves promising performance, further comparison with alternative methods is required.

CNN architectures have been widely used in image processing due to their feature extraction and end-to-end training capabilities. Notable architectures include AlexNet, VggNet, and GoogleNet, which excelled in competitions like the ImageNet Challenge [69]. Additionally, lightweight models like MobileNetv2 and SqueezeNet, designed for efficiency, use techniques like inverted residual blocks and Fire Modules to reduce computational cost [70]. With the advent of Transformer architectures, Vision Transformer captures global dependencies by dividing images into patches and leveraging a multi-head self-attention mechanism [71]. These models are selected for comparison against the proposed method.

Selection of fusion strategies critically influences the performance of multimodal classifiers. Multiple approaches have been used for information fusion beyond Dempster-Shafer theory. Bayesian Model Averaging (BMA) [72] weights predictions by each model's posterior probability and averages them to produce more stable and dependable inferences. Voting is a simple and intuitive fusion strategy [73]. Specifically, hard voting aggregates model outputs by majority rule on class labels, while soft voting typically averages class prediction probabilities as the final decision criterion. Furthermore, fuzzy logic [74] handles uncertainty and fuzziness in the classification process by defining membership functions. Additionally, feature-level concatenation [75] is a fundamental fusion method that directly joins the feature vectors from different modalities into a unified vector for input to a classifier.

Several comparative experiments are conducted to further validate the model's performance and advantages. By replacing the backbone for feature extraction in DF-ResNet, dual-input weight-sharing networks corresponding to various base models are established. Training and evaluation are conducted to compare these models with D-ResNet and DF-ResNet. These models are trained on the preprocessed dataset described earlier using the same hyperparameters as in Table 5. Performance is evaluated using identical metrics: Accuracy, Precision, Recall, and F1-score, with results presented in Table 8. Detailed classification results for each model are plotted in confusion matrices shown in Fig. 21. Building on this, Fig. 22 demonstrates the impact of sample size reduction on model performance, comparing well-performing SqueezeNet and GoogleNet from Table 8 against the backbone F-

Table 8
Comparison of evaluation metrics for various models.

Backbone	Accuracy	Precision	Recall	F1-score
AlexNet	0.524	0.388	0.537	0.432
VggNet	0.611	0.645	0.610	0.602
GoogleNet	0.761	0.819	0.771	0.743
MobileNetv2	0.734	0.740	0.739	0.738
SqueezeNet	0.789	0.802	0.795	0.792
Vision Transformer	0.703	0.727	0.703	0.702
ResNet	0.816	0.824	0.818	0.818
F-ResNet	0.850	0.851	0.854	0.852

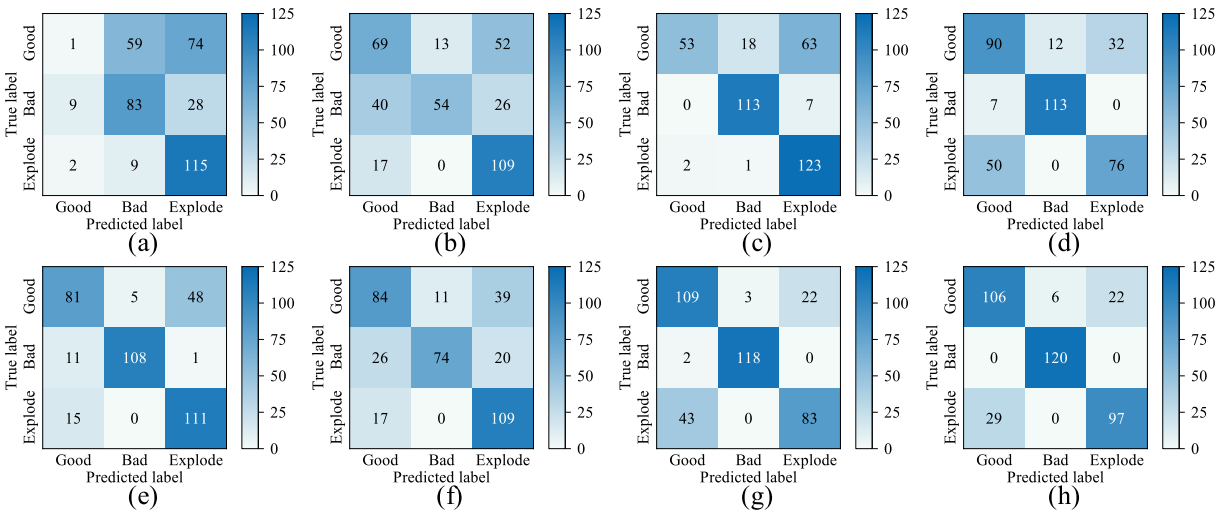


Fig. 21. Confusion matrix for model classification with backbone (a) AlexNet, (b) VggNet, (c) GoogleNet, (d) MobileNetv2, (e) SqueezeNet, (f) Vision Transformer, (g) ResNet, and (h) F-ResNet.

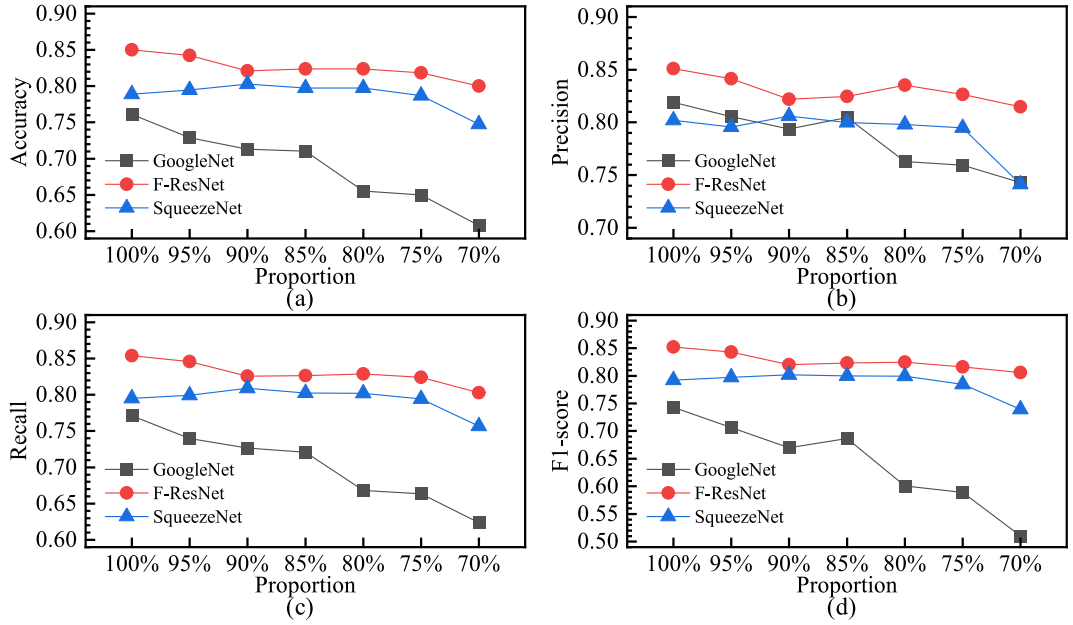


Fig. 22. Values of evaluation indicators when reducing the number of samples in the training set: (a) Accuracy; (b) Precision; (c) Recall, and (d) F1-score.

ResNet used in this study. Further comparative analysis assessed the performance of different fusion strategies, as shown in Fig. 23. These results are analyzed as follows.

(1) The ResNet-based backbone network demonstrates optimal performance in comprehensive feature extraction from resistance spot welding surface images. According to evaluation results in Table 8, all other models achieve an Accuracy below 0.8 on the test set. AlexNet attains an Accuracy of only 0.524 due to its simple architecture and limited feature extraction capacity, ranking it lowest among all models. As shown in Fig. 21 (a), AlexNet failed to learn discriminative features for the "Good" category, with nearly all such samples misclassified as "Bad" or "Explode". SqueezeNet achieves an Accuracy of 0.789, a Precision of 0.802, a Recall of 0.795, and an F1-score of 0.792, outperforming other comparative models yet remaining inferior to both the F-ResNet and ResNet backbones employed in this study. Fig. 21 reveals that performance differentials primarily concern the identification of the "Good" and "Explode" samples. For instance, GoogleNet attains high recall for Bad and Explode samples but misclassifies over half of the Good instances. While MobileNetv2 and SqueezeNet effectively distinguished "Bad" and "Explode" categories, they exhibited significant misclassification rates when processing "Good" samples. In contrast, the DF-ResNet with F-ResNet backbone utilized in this study also exhibits misclassifications, but they are mainly misjudgments between "Good" and "Explode" categories, and the number is also lower than that of other models. In summary, the FPN-enhanced ResNet backbone emerges as the most suitable feature extractor for processing surface images of weld joints.

(2) Testing on data subsets demonstrates that the model exhibits strong robustness to reductions in training sample size. In the experiments, training samples are randomly removed to create subsets comprising 95 %, 90 %, 85 %, 80 %, 75 %, and 70 % of the original training set, while the test set remains unchanged. As

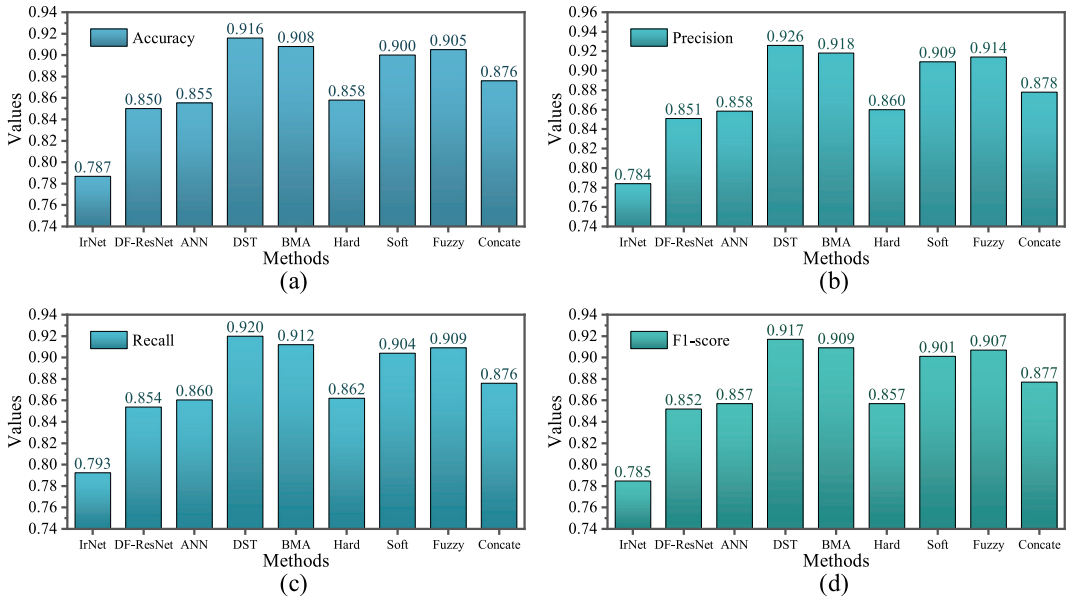


Fig. 23. Comparison under different ensemble methods of (a) Accuracy; (b) Precision; (c) Recall; (d) F1-Score.

depicted in Fig. 22, GoogleNet exhibited significant performance degradation with reduced training data. At 70 % training data, its Accuracy, Precision, Recall, and F1-score drop to 0.608, 0.743, 0.624, and 0.510, respectively, with average decreases of 0.67 % in accuracy, 0.31 % in precision, 0.64 % in recall, and 1.04 % in F1-score per 1 % data reduction. By contrast, when using an F-ResNet backbone on the 70 % subset, the evaluated Accuracy, Precision, Recall, and F1-score are 0.800, 0.815, 0.803, and 0.806, respectively. Compared with using the complete training set, the performance of the model only decreases by 5.88 %, 4.25 %, 6.01 % and 5.43 % respectively. On average, each 1 % reduction in training data results in drops of 0.20 % in accuracy, 0.14 % in precision, 0.20 % in recall, and 0.18 % in F1-score. Additionally, SqueezeNet maintains stable performance initially, with no marked declines across the four metrics. However, once the training subset falls below 75 %, its metrics also begin to decline rapidly. In summary, the F-ResNet-based feature extractor accurately identifies weld surface images even under small-sample conditions.

(3) Dempster-Shafer evidence theory demonstrates prominent advantages as a fusion strategy for multimodal welding defect detection. As shown in Fig. 23, the result of DS theory performs most prominently, achieving an accuracy of 91.6 % and an F1-score of 91.7 %. This represents an improvement of 6.07 % in accuracy and 6.01 % in F1-score over the best base model (ANN). The simultaneous optimization of precision (92.6 %) and recall (92.0 %) validates the significant advantage of DS theory in reconciling conflicts between models and fusing uncertain information. By contrast, other fusion methods show graded differences while outperforming the base models: BMA and fuzzy logic achieved F1-scores of 90.9 % and 90.7 %, respectively. Soft voting reached 90.1 % F1, demonstrating the effectiveness of probability-weighted fusion. Conversely, hard voting and feature concatenation yield lower F1-scores of 85.7 % and 87.7 %, indicating the limitations of discrete label voting and simple feature-level fusion. Overall, DS theory systematically integrates multi-model discriminative information via basic probability assignment functions, achieving a superior balance among precision, recall, and overall generalization.

Compared to prior works that often rely on single-modality data or simplistic fusion strategies, our approach demonstrates how effective integration of heterogeneous data sources can substantially improve detection reliability and operational adaptability. By achieving high accuracy and real-time processing capabilities, the proposed method offers a practical and scalable solution for automated quality inspection in civil engineering applications, such as the construction of steel structures and bridges. It provides a foundation for more intelligent and autonomous welding monitoring systems, thereby promoting the advancement of automated and robotic technologies in infrastructure construction projects.

6. Conclusions

This paper presented an ensemble multimodal deep learning approach for the recognition and classification of resistance spot welding quality. The framework integrates a dual-input weight-sharing network, Dempster-Shafer evidence theory, and two model interpretation methods. Dual-input weight-sharing networks and diverse model interpretation methods effectively improve the full utilization of training data and model understanding for subsequent optimization. Unlike conventional multimodal approaches, this method enables more parameter-efficient joint representation learning, conflict-aware probabilistic fusion, and explicit model behavior interpretation. This combination offers a systematically explainable and highly accurate solution for automated welding quality inspection, advancing the intelligence and reliability of non-destructive evaluation in construction projects. Its effectiveness is fully verified through training, evaluation and comparative experiments a publicly available dataset.

Testing and evaluation based on real experimental data indicate that: (1) The proposed approach achieves strong multimodal classification performance on RSW data, with macro averaged Precision, Recall, and F1-score of 0.926, 0.920, and 0.917, respectively, and an overall Accuracy of 0.916. (2) The integration of models enables the decision to comprehensively reference information contained within different modalities, effectively enhancing accuracy. (3) The tailored backbone network design boosts surface image classification performance, with Accuracy and F1-score improving by 7.87 % and 8.26 %, respectively. (4) MM-SHAP analysis reveals that the ensemble model exhibits modality preferences when making predictions for different categories. (5) The ResNet-based backbone more effectively extracts weld surface features compared to classic methods such as AlexNet, VGGNet, and

GoogleNet. This research demonstrates that integrating multimodal data with DS theory-based fusion and interpretability analysis provides a robust, interpretable, and practical solution for industrial welding defect detection. This method can be effectively integrated with the sensor system into the automated production process [76]. By processing the multi-modal data collected during the welding process, it can achieve efficient and timely identification of welding defects.

While the proposed method achieves excellent results, there are still some limitations requiring further research in the future. For one thing, the information fusion method used in this study is fusion at the decision level, neglecting the cross-influence between features in different modalities. Future efforts could develop more complex models to facilitate deep feature fusion, exploring Transformer-based cross-modal fusion or joint feature space architectures. For another, this study identifies welding quality after welding completion. Future efforts could explore real-time detection and prediction during the welding process to promptly adjust parameters and avoid defect formation. Furthermore, testing the model in a wider range of welding process scenarios, such as arc welding and laser welding, will help validate its robustness. If future research can be conducted with training on more representative and balanced large-scale datasets, the generalization ability of the model and the practical significance of the conclusions will be further enhanced. Furthermore, testing in more welding process scenarios such as arc welding and laser welding, as well as with more defect samples and in real factory environments, will help verify the robustness of the model.

CRediT authorship contribution statement

Shiqiang Tang: Writing – original draft, Data curation, Visualization, Validation, Methodology, Writing – review & editing. **Feilong Fei:** Resources. **Limao Zhang:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization. **Jinfeng Yu:** Methodology.

Declaration of competing interest

We declare that the manuscript has not been submitted to, nor is under review at, another journal or other publishing venue. All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version. The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China (Nos. 72571110, 72271101) and the Hubei Provincial Key Research and Development Program (No. 2024DJC007).

Data availability

Data will be made available on request.

References

[1] L. Zhang, J. Guo, X. Fu, RLK. Tiong, P. Zhang, Digital twin enabled real-time advanced control of TBM operation using deep learning methods, Autom Constr (2024) 158, https://doi.org/10.1016/j.autcon.2023.105240.
[2] H. Wang, Y. Rong, J. Xu, Y. Huang, G. Zhang, Application and trends of point cloud in intelligent welding: state of the art review, J. Manuf. Syst. 79 (2025) 48–72, https://doi.org/10.1016/j.jmsy.2025.01.001.
[3] T. Wang, P. Upadhyay, S. Whalen, A review of technologies for welding magnesium alloys to steels, Int. J. Precis. Eng. Manuf.-Green Technol. 8 (3) (2021) 1027–1042, https://doi.org/10.1007/s40684-020-00247-x.
[4] V. Vasan, N.V. Sridharan, R.J. Balasundaram, S. Vaithiyanathan, Ensemble-based deep learning model for welding defect detection and classification, Eng. Appl. Artif. Intell. 136 (2024) 108961, https://doi.org/10.1016/j.engappai.2024.108961.
[5] A. Cardellicchio, M. Nitti, C. Patruno, N. Mosca, M. di Summa, E. Stella, V. Renò, Automatic quality control of aluminium parts welds based on 3D data and artificial intelligence, J. Intell. Manuf. 35 (4) (2024) 1629–1648, https://doi.org/10.1007/s10845-023-02124-1.
[6] Z. Zhang, Z. Yang, W. Ren, G. Wen, Random forest-based real-time defect detection of Al alloy in robotic arc welding using optical spectrum, J. Manuf. Process. 42 (2019) 51–59, https://doi.org/10.1016/j.jmapro.2019.04.023.
[7] L. Xu, S. Dong, H. Wei, Q. Ren, J. Huang, J. Liu, Defect signal intelligent recognition of weld radiographs based on YOLO V5-IMPROVEMENT, J. Manuf. Process. 99 (2023) 373–381, https://doi.org/10.1016/j.jmapro.2023.05.058.
[8] W. Cui, K. Song, Y. Wang, G. Lv, Y. Yan, H. Yu, X. Li, A rapid screening method for suspected defects in steel pipe welds by combining correspondence mechanism and normalizing flow, IEEE Trans. Industr. Inform. 20 (9) (2024) 11171–11180, https://doi.org/10.1109/TII.2024.3399934.
[9] X. Yu, P. Zuo, J. Xiao, Z. Fan, Detection of damage in welded joints using high order feature guided ultrasonic waves, Mech. Syst. Signal Process. 126 (2019) 176–192, https://doi.org/10.1016/j.ymssp.2019.02.026.
[10] A. Zolfaghari, A. Zolfaghari, F. Kolahan, Reliability and sensitivity of magnetic particle nondestructive testing in detecting the surface cracks of welded components, Nondestruct. Test. Eval. 33 (3) (2018) 290–300, https://doi.org/10.1080/10589759.2018.1428322.
[11] W. Sun, D.R. Symes, C.M. Brenner, M. Böhnel, S. Brown, M.N. Mavrogordato, I. Sinclair, M. Salamon, Review of high energy x-ray computed tomography for non-destructive dimensional metrology of large metallic advanced manufactured components, Rep. Prog. Phys. 85 (1) (2022) 016102, https://doi.org/10.1088/1361-6633/ac43f6.
[12] G. Liu, D. Yang, J. Ye, H. Lu, Z. Wang, Y. Zhao, A real-time welding defect detection framework based on RT-DETR deep neural network, Adv. Eng. Inform. 65 (2025) 103318, https://doi.org/10.1016/j.aei.2025.103318.
[13] A. Raj, U. Chadha, A. Chadha, R.R. Mahadevan, B.R. Sai, D. Chaudhary, S. K. Selvaraj, R. Lokeshkumar, S. Das, B. Karthikeyan, R. Nagalakshmi, V. Chandramohan, H. Hadidi, Weld quality monitoring via machine learning-enabled approaches, Int. J. Interact. Des. Manuf. (IJIDeM) (2023), https://doi.org/10.1007/s12008-022-01165-9.
[14] A.A. Melakhsou, M. Batton-Hubert, Welding monitoring and defect detection using probability density distribution and functional nonparametric kernel classifier, J. Intell. Manuf. 34 (3) (2023) 1469–1481, https://doi.org/10.1007/s10845-021-01871-3.
[15] J. Sun, C. Li, X.J. Wu, V. Palade, W. Fang, An effective method of weld defect detection and classification based on machine vision, IEEE Trans. Industr. Inform. 15 (12) (2019) 6322–6333, https://doi.org/10.1109/TII.2019.2896357.
[16] D. Amirkhani, M.S. Allili, L. Hebbache, N. Hammouche, J.F. Lapointe, Visual concrete bridge defect classification and detection using deep learning: a systematic review, IEEE Trans. Intell. Transp. Syst. 25 (9) (2024) 10483–10505, https://doi.org/10.1109/TITS.2024.3365296.
[17] X. Wang, U. Zscherpel, P. Tripicchio, S. D'Avella, B. Zhang, J. Wu, Z. Liang, S. Zhou, X. Yu, A comprehensive review of welding defect recognition from X-ray images, J. Manuf. Process. 140 (2025) 161–180, https://doi.org/10.1016/j.jmapro.2025.02.039.
[18] U. Sreedhar, C.V. Krishnamurthy, K. Balasubramaniam, V.D. Raghupathy, S. Ravisankar, Automatic defect identification using thermal image analysis for online weld quality monitoring, J. Mater. Process. Technol. 212 (7) (2012) 1557–1566, https://doi.org/10.1016/j.jmatprotec.2012.03.002.
[19] F. Xu, Y. Xu, H. Zhang, S. Chen, Application of sensing technology in intelligent robotic arc welding: a review, J. Manuf. Process. 79 (2022) 854–880, https://doi.org/10.1016/j.jmapro.2022.05.029.
[20] W. Huang, R. Kovacevic, A laser-based vision system for weld quality inspection, Sensors 11 (1) (2011) 506–521, https://doi.org/10.3390/s110100506.
[21] Y. Pan, L. Zhang, Roles of artificial intelligence in construction engineering and management: a critical review and future trends, Autom. Constr. 122 (2021) 103517, https://doi.org/10.1016/j.autcon.2020.103517.
[22] Z. Chen, X. Niu, J. Liu, K. Khan, Y. Liu, Seismic study on an innovative fully-bolted beam-column joint in prefabricated modular steel buildings, Eng. Struct. 234 (2021) 111875, https://doi.org/10.1016/j.engstruct.2021.111875.
[23] X.-L. Gu, B. Zhang, X.-K. Huang, X.-L. Wang, Experimental investigation on progressive collapse resistance of precast concrete frame structures with different beam-column connections, J. Build. Eng. 56 (2022) 104803, https://doi.org/10.1016/j.jobe.2022.104803.
[24] Z. Zhao, X. Cheng, Y. Li, M. Diao, H. Guan, Y. An, Progressive collapse analysis of precast reinforced concrete beam-column assemblies with different dry connections, Eng. Struct. 287 (2023) 116174, https://doi.org/10.1016/j.engstruct.2023.116174.
[25] X.C. Liu, A.X. Xu, A.L. Zhang, Z. Ni, H.X. Wang, L. Wu, Static and seismic experiment for welded joints in modularized prefabricated steel structure, J. Constr. Steel Res. 112 (2015) 183–195, https://doi.org/10.1016/j.jcsr.2015.05.003.
[26] T. Qiu, J. Zhang, X. Chen, Z. Xu, D. Su, R. Song, T. Cui, Experimental investigation and mechanical model for assembled joints of prefabricated two-wall-in-one diaphragm walls, Eng. Struct. 275 (2023) 115285, https://doi.org/10.1016/j.engstruct.2022.115285.
[27] Q. Liu, P. Song, X. Wang, H. Guo, X. Gao, R. Liu, Y. Long, Full-scale evaluation of key performance of novel welded connectors in precast airport pavement under fatigue loading, Structures 81 (2025) 110104, https://doi.org/10.1016/j.istruc.2025.110104.

[28] D. Chen, W. Xu, H. Qian, J. Sun, J. Li, Effects of non-uniform temperature on closure construction of spatial truss structure, J. Build. Eng. 32 (2020) 101532, https://doi.org/10.1016/j.jobe.2020.101532.
[29] E. Şahin, N.N. Arslan, D. Özdemir, Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning, Neural Comput. & Applic. 37 (2) (2025) 859–965, https://doi.org/10.1007/s00521-024-10437-2.
[30] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626, https://doi.org/10.1109/ICCV.2017.74.
[31] S. Frank, E. Bugliarello, D. Elliott, Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 9847–9857, https://doi.org/10.18653/v1/2021.emnlp-main.775.
[32] L. Parcalabescu, A. Frank, MM-SHAP: A Performance-agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 4032–4059, https://doi.org/10.18653/v1/2023.acl-long.223.
[33] J. Wang, K. Chen, H. Yang, L. Zhang, Ensemble deep learning enabled multi-condition generative design of aerial building machine considering uncertainties, Autom. Constr. 157 (2024) 105134, https://doi.org/10.1016/j.autcon.2023.105134.
[34] Q. Li, Q. Yao, L. Sun, H. Ma, C. Zhang, N. Wang, Effect of micro-galvanic corrosion on corrosion fatigue cracking of the weld joint of high strength bridge steel, Int. J. Fatigue 170 (2023) 107568, https://doi.org/10.1016/j.ijfatigue.2023.107568.
[35] J. Huo, N. Hou, W. Sun, L. Wang, J. Dong, Analyses of dynamic characteristics and structure optimization of tunnel boring machine cutter system with multi-joint surface, Nonlinear Dynam. 87 (1) (2017) 237–254, https://doi.org/10.1007/s11071-016-3038-0.
[36] Y. Guo, X. Li, D. Jin, H. Liu, Y. Fang, Assessment of slurry chamber clogging alleviation during ultra-large-diameter slurry tunnel boring machine tunneling in hard-rock using computational fluid dynamics-discrete element method: a case study, J. Rock Mech. Geotech. Eng. 17 (8) (2025) 4715–4734, https://doi.org/10.1016/j.jrmge.2024.09.059.
[37] M. Hasan, M. Lu, Bridging AI and explainability in civil engineering: the Yin-Yang of predictive power and interpretability, AI Civil Eng. 4 (1) (2025) 21, https://doi.org/10.1007/s43503-025-00066-6.
[38] Y. Hong, M. Yang, B. Chang, D. Du, Filter-PCA-based process monitoring and defect identification during climbing helium arc welding process using DE-SVM, IEEE Trans. Ind. Electron. 70 (7) (2023) 7353–7362, https://doi.org/10.1109/TIE.2022.3201304.
[39] Y.-J. Xia, Z.-W. Su, Y.-B. Li, L. Zhou, Y. Shen, Online quantitative evaluation of expulsion in resistance spot welding, J. Manuf. Process. 46 (2019) 34–43, https://doi.org/10.1016/j.jmapro.2019.08.004.
[40] Y. Li, Y.F. Li, Q.L. Wang, D. Xu, M. Tan, Measurement and defect detection of the weld bead based on online vision inspection, IEEE Trans. Instrum. Meas. 59 (7) (2010) 1841–1849, https://doi.org/10.1109/TIM.2009.2028222.
[41] M. Amarnath, N. Sudharshan, P. Srinivas, Automatic detection of defects in welding using deep learning - a systematic review, Mater. Today Proc. (2023), https://doi.org/10.1016/j.matpr.2023.03.268.
[42] S.B. Jha, R.F. Babiceanu, Deep CNN-based visual defect detection: survey of current literature, Comput. Ind. 148 (2023) 103911, https://doi.org/10.1016/j.compind.2023.103911.
[43] X. Fu, M. Wu, R.L.K. Tiong, L. Zhang, Data-driven real-time advanced geological prediction in tunnel construction using a hybrid deep learning approach, Autom. Constr. 146 (2023) 104672, https://doi.org/10.1016/j.autcon.2022.104672.
[44] W. Liu, J. Hu, J. Qi, Coarse-to-fine vision-based welding spot anomaly detection in production lines of body-in-white, J. Manuf. Syst. 81 (2025) 144–154, https://doi.org/10.1016/j.jmsy.2025.05.003.
[45] D. Dhruva Kumar, C. Fang, Y. Zheng, Y. Gao, Semi-supervised transfer learning-based automatic weld defect detection and visual inspection, Eng. Struct. 292 (2023) 116580, https://doi.org/10.1016/j.engstruct.2023.116580.
[46] W. Dai, D. Li, D. Tang, H. Wang, Y. Peng, Deep learning approach for defective spot welds classification using small and class-imbalanced datasets, Neurocomputing 477 (2022) 46–60, https://doi.org/10.1016/j.neucom.2022.01.004.
[47] X. Yang, X. Liu, Q. Wu, G. Wen, S. Mei, G. Liao, T. Shi, VMMAO-YOLO: an ultra-lightweight and scale-aware detector for real-time defect detection of avionics thermistor wire solder joints, Front. Mech. Eng. 19 (3) (2024) 21, https://doi.org/10.1007/s11465-024-0793-3.
[48] W. Dai, D. Li, Y. Zheng, D. Wang, D. Tang, H. Wang, Y. Peng, Online quality inspection of resistance spot welding for automotive production lines, J. Manuf. Syst. 63 (2022) 354–369, https://doi.org/10.1016/j.jmsy.2022.04.008.
[49] R. Miao, Z. Shan, Q. Zhou, Y. Wu, L. Ge, J. Zhang, H. Hu, Real-time defect identification of narrow overlap welds and application based on convolutional neural networks, J. Manuf. Syst. 62 (2022) 800–810, https://doi.org/10.1016/j.jmsy.2021.01.012.
[50] S. Wang, E. Zhang, L. Zhou, Y. Han, W. Liu, J. Hong, 3DWDC-net: An improved 3DCNN with separable structure and global attention for weld internal defect classification based on phased array ultrasonic tomography images, Mech. Syst. Signal Process. 229 (2025) 112564, https://doi.org/10.1016/j.ymssp.2025.112564.
[51] R. Zhang, D. Liu, Q. Bai, L. Fu, J. Hu, J. Song, Research on X-ray weld seam defect detection and size measurement method based on neural network self-optimization, Eng. Appl. Artif. Intell. 133 (2024) 108045, https://doi.org/10.1016/j.engappai.2024.108045.
[52] J. Zhou, D. Wang, J. Chen, Z. Feng, B. Clarson, A. Baselhuhn, Autonomous nondestructive evaluation of resistance spot welded joints, Robot. Comput. Integr. Manuf. 72 (2021) 102183, https://doi.org/10.1016/j.rcim.2021.102183.
[53] J. Zhou, Z. Xi, S. Wang, B. Yang, Y. Zhang, Y. Zhang, A real spatial–temporal attention denoising network for nugget quality detection in resistance spot weld, J. Intell. Manuf. 35 (6) (2024) 2743–2764, https://doi.org/10.1007/s10845-023-02160-x.
[54] J. Liu, F. Jiang, S. Tashiro, S. Chen, M. Tanaka, A physics-informed and data-driven framework for robotic welding in manufacturing, Nat. Commun. 16 (1) (2025) 4807, https://doi.org/10.1038/s41467-025-60164-y.
[55] C. Xu, L. Xu, S. Zhao, L. Yu, C. Zhang, Complementary knowledge augmented multimodal learning method for yarn quality soft sensing, Eng. Appl. Artif. Intell. 133 (2024) 108057, https://doi.org/10.1016/j.engappai.2024.108057.
[56] J. Wang, Z. Zhang, Z. Bai, S. Zhang, R. Qin, J. Huang, G. Wen, On-line defect recognition of MIG lap welding for stainless steel sheet based on weld image and CMT voltage: feature fusion and attention weights visualization, J. Manuf. Process. 108 (2023) 430–444, https://doi.org/10.1016/j.jmapro.2023.10.081.
[57] G. He, X. Gao, H. Yang, AETMC-FCVT: An end-to-end welding defect detection and classification method based on magneto-optical infrared bi-imaging system, Mech. Syst. Signal Process. 224 (2025) 112058, https://doi.org/10.1016/j.ymssp.2024.112058.
[58] K. Zhou, P. Yao, Overview of recent advances of process analysis and quality control in resistance spot welding, Mech. Syst. Signal Process. 124 (2019) 170–198, https://doi.org/10.1016/j.ymssp.2019.01.041.
[59] Z. Wang, P. Wang, K. Liu, P. Wang, Y. Fu, C.-T. Lu, C.C. Aggarwal, J. Pei, Y. Zhou, A comprehensive survey on data augmentation, arXiv Preprint (2024), https://doi.org/10.48550/arXiv.2405.09591 arXiv:2405.09591.
[60] J. Zhang, X. Xi, J. Du, X. He, M. Wu, Y. Wei, Normalized intrinsic deep features based zero-watermarking scheme for remote sensing images using U-net and K-means, in: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2025, pp. 1–15, https://doi.org/10.1109/JSTARS.2025.3586280.
[61] J. Cui, C. Lv, X. Qu, J. Du, H. Wang, Development of an intelligent CNN-LSTM-attention model for acoustic emission-based fracture detection and structural health monitoring in marine steel structures, Ocean Eng. 339 (2025) 122002, https://doi.org/10.1016/j.oceaneng.2025.122002.
[62] M. Momeny, A.A. Neshat, M.A. Hussain, S. Kia, M. Marhamati, A. Jahanbakhshi, G. Hamarneh, Learning-to-augment strategy using noisy and denoised data: improving generalizability of deep CNN for the detection of COVID-19 in X-ray images, Comput. Biol. Med. 136 (2021) 104704, https://doi.org/10.1016/j.compbiomed.2021.104704.
[63] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.
[64] T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936–944, https://doi.org/10.1109/CVPR.2017.106.
[65] T. Gudiyangada Nachappa, S. Tavakkoli Piralilou, K. Gholamnia, O. Ghorbanzadeh, O. Rahmati, T. Blaschke, Flood susceptibility mapping with machine learning, multi-criteria decision analysis and ensemble using Dempster Shafer theory, J. Hydrol. 590 (2020) 125275, https://doi.org/10.1016/j.jhydrol.2020.125275.
[66] Y. Kessentini, T. Burger, T. Paquet, A Dempster–Shafer Theory based combination of handwriting recognition systems with multiple rejection strategies, Pattern Recogn. 48 (2) (2015) 534–544, https://doi.org/10.1016/j.patcog.2014.08.010.
[67] Y.-L. Tao, X.-C. Liu, X. Chen, W.-B. Cui, L.-X. Guo, Analysis and design of column splicing flange joints with different weld types, J. Constr. Steel Res. 212 (2024) 108324, https://doi.org/10.1016/j.jcsr.2023.108324.
[68] L. Domínguez, E.A. Rivas-Araiza, J.C. Jáuregui-Correa, J.L. González-Córdoba, J. C. Pedraza-Ortega, A. Takács, Resistance spot welding insights: a dataset integrating process parameters, infrared, and surface imaging, Data Brief 59 (2025) 111373, https://doi.org/10.1016/j.dib.2025.111373.
[69] L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, J. Big Data 8 (1) (2021) 53, https://doi.org/10.1186/s40537-021-00444-8.
[70] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, Y. Miao, Review of image classification algorithms based on convolutional neural networks, Remote Sens. 13 (2021), https://doi.org/10.3390/rs13224712.
[71] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, D. Tao, A survey on vision transformer, IEEE Trans. Pattern Anal. Mach. Intell. 45 (1) (2023) 87–110, https://doi.org/10.1109/TPAMI.2022.3152247.

[72] L. Wei, S. Jiang, J. Dong, L. Ren, Y. Liu, L. Zhang, M. Wang, Z. Duan, Fusion of gauge-based, reanalysis, and satellite precipitation products using Bayesian model averaging approach: determination of the influence of different input sources, J. Hydrol. 618 (2023) 129234, https://doi.org/10.1016/j.jhydrol.2023.129234.
[73] H. Zahid, O. Elmansoury, R. Yaagoubi, Dynamic predicted mean vote: An IoT-BIM integrated approach for indoor thermal comfort optimization, Autom. Constr. 129 (2021) 103805, https://doi.org/10.1016/j.autcon.2021.103805.
[74] X. Han, F. Chen, J. Ban, FMFN: a fuzzy multimodal fusion network for emotion recognition in ensemble conducting, IEEE Trans. Fuzzy Syst. 33 (1) (2025) 168–179, https://doi.org/10.1109/TFUZZ.2024.3373125.
[75] Y. Li, M. El Habib Daho, P.-H. Conze, R. Zeghlache, H. Le Boité, R. Tadayoni, B. Cochener, M. Lamard, G. Quellec, A review of deep learning-based information fusion techniques for multimodal medical image classification, Comput. Biol. Med. 177 (2024) 108635, https://doi.org/10.1016/j.compbiomed.2024.108635.
[76] P. Sassi, P. Tripicchio, C.A. Avizzano, A smart monitoring system for automatic welding defect detection, IEEE Trans. Ind. Electron. 66 (12) (2019) 9641–9650, https://doi.org/10.1109/TIE.2019.2896165.