



OPEN A misclassification-aware explainable hybrid CNN-vision transformer framework for radiographic weld inspection

Kumar Parmar¹, Rituraj Jain¹, P. T. Anitha², A. Jayanthi³, Ramesh Babu Putchanuthala⁴, Kamal Upreti⁵ & Terefe Bayisa Ayele⁶✉

Proper and valid identification of the defects in welding is also essential in the safety-critical manufacturing systems in order to maintain structural integrity. In general, even though deep learning-based inspection methods have demonstrated high classification accuracy, they frequently fail to distinguish similar-looking types of defects and misclassify them at high costs in practice. The paper presented an intelligent weld inspection framework, hybrid Convolutional Neural Network-Vision Transformer (CNN-ViT) architecture, aimed at improving reliability and interpretability, with potential relevance to Industry 5.0-oriented inspection systems. The suggested model was relatively tested against a tailored lightweight CNN baseline, provided that the same training conditions were implemented with the RIAWELC radiographic weld dataset. It is demonstrated that the hybrid CNN-ViT has the highest accuracy of 98.56%, compared to the CNN baseline (97.90%), at a relative difference of about 31% in the misclassification rate. Misclassification-sensitive analysis with confusion matrices aided by explainability methods with Grad-CAM and transformer self-attention maps proves that the global contextual modeling significantly enhances the discrimination of visually similar defects that include cracks, porosity, and absence of penetration. The trained model was also tested on external data (GDXray dataset) collected under different imaging conditions to test its robustness and it was found to perform well on generalization. Moreover, fivefold cross-validation was performed in order to evaluate how the model performs across different data partitions. These findings validate that the suggested hybrid attention-based framework, in addition to the accuracy improvement, leads to the decrease in the misclassification and the availability of the transparent decision explanations. In general, the suggested solution would be a technically robust and explainable framework that may support future Industry 5.0-aligned intelligent weld inspection systems.

Keywords Weld defect classification, Misclassification analysis, Industry 5.0, Smart manufacturing, Automation, Non-destructive evaluation, Control of welding quality

Welding is a basic process in contemporary manufacturing and building, and the basis of the structural integrity and durability of parts in diverse industrial fields. And this is why it is so important to ensure weld quality, as undetected faults can have a negative impact on safety, performance and reliability¹. Conventionally, Non-Destructive Testing (NDT) which include visual inspection, radiographic testing, ultrasonic testing, magnetic particle testing and liquid penetrant testing techniques have been employed in the assessment of welds^{2–4}. Despite the fact that visual testing is still in use due to its simplicity and low cost, visual testing is highly reliant on the knowledge of the inspector and thus it is subjective as well as prone to inconsistencies³. As industrial production increases, human-driven inspections are struggling to keep up with the demands of speed, accuracy, and consistency, and as a result, fatigue-induced errors and subjective assessments occur⁵.

¹Department of Information Technology, Marwadi University, Rajkot, Gujarat, India. ²Department of MCA, Mohan Babu University, Tirupati, Andhra Pradesh, India. ³BVRIT Hyderabad College of Engineering for Women, Hyderabad, Telangana, India. ⁴Department of Computer Science & Engineering, Narsimha Reddy Engineering College, Secunderabad, Telangana, India. ⁵Department of Computer Science, Christ University, Delhi NCR Campus, Ghaziabad, India. ⁶School of Electrical and Computer Engineering, Wollega University, Nekemte, Ethiopia. ✉email: terefebayisa2018@gmail.com

Recent progress in Artificial Intelligence (AI) and Deep Learning (DL) have revolutionized automated weld inspection systems with improved accuracy, repeatability and scalability⁶. CNNs have been widely used because of their strong capability of combining the extraction of features and their classification into a unified architecture⁵. Such models as AlexNet, ResNet, MobileNet, Faster R-CNN, and variants of YOLO (including lightweight TA-YOLOv8) have been demonstrated to exhibit high performance in terms of detection and classification of the defects in the weld that are pores, cracks, non-penetration, slag inclusion, etc.^{7,8}. These models have demonstrated high levels of accuracy in binary and even multi-class models with some even extending to real-time predictions of weld penetration using architectures like MobileNetV2¹. However, CNN-based approaches are still facing key challenges, especially lack of global feature awareness, and high computational complexity caused by deep parameterized layers⁶.

To overcome all these limitations, Vision Transformers (ViTs) have become a good alternative. Originally designed for natural language processing, ViTs make use of self-attention mechanisms to capture long range dependencies and global contextual relationships between image regions⁹. When applied to computer vision tasks, ViTs have shown excellent performance on the identification of weld penetration states including incomplete fusion, partial penetration, full penetration and excessive penetration^{1,3}, by being able to effectively model both local textures and global semantics. Variants such as the Swin Transformer enhance this process by hierarchical attention mechanisms which enhance the localization of defects and multi-scale representation¹⁰. Such transformer-based approaches have enhanced the robustness, explainability and generalization of the automated quality inspection systems.

Despite these advances, there is an important research gap in this area in which most studies are focused on the accuracy of aggregate measurements whereas they ignore the dynamics of misclassification processes between similar classes of defects. In industrial terms, this control can be expensive, as visually overlapping defects such as cracks, porosity and absence of penetration are quite common, which can cause confusion in automated systems, leading to inconsistent judgments of the quality and increased inspection downtime.

In this aspect, the current research performs a misclassification-conscious comparative study on a custom CNN and a hybrid CNN-ViT network in terms of classifying welding defects. Even though the hybrid CNN-Transformer models are prevalent in current literature, the majority of studies on weld defects primarily investigate the overall accuracy and fail to provide a systematic study of the misclassification behavior and the statistical reliability, or validation criteria based on the explainable nature.

The contribution of this work therefore extends beyond architectural integration. In particular, this paper proposes (i) individual class-based misclassification analysis to lessen the confusion among the visually similar defects, (ii) statistically validating the performance differences by the means of hypothesis testing and confidence intervals, and (iii) generate explainability-based interpretation coordinating transformer attention and error reduction. Besides this, a false negative to a false positive trade-off analysis and external cross-domain validation is carried out to test industrial reliability and robustness. The framework proposed can combine architectural comparison and systematic error analysis with statistical validation to improve the reliability and interpretability of hybrid inspection systems, may support Industry 5.0-oriented inspection frameworks.

Literature review

Deep learning has dramatically changed the way automated weld defect detection is performed-to replace manual radiographic inspection with the use of intelligent vision-based systems. Researchers have experimented with various paradigms of neural networks ranging from the traditional Artificial Neural Networks (ANNs), Convolution Neural Networks (CNNs) to the hybrid meta-learning and transformer-based models for greater accuracy and adaptability.

Kusuma et al.¹¹ used an Artificial Neural Network with 100 hidden layers using ReLU and L-BFGS-B optimizer. Their workflow consisted of heavy preprocessing of welding workpieces images such as resizing, orientation correction, and data augmentation before feeding the ANN with learning on the defect classification¹². Parametric methods for detecting RLA in remote sensing images from space: A hybrid meta-learning+transfer learning framework to transfer pretrained weights from ImageNet to RIAWELC dataset. The proposed model combined few-shot meta-learning for fast adaptation with fine-tuning on a small labeled subset with the support of a data augmentation technique MixUp to accommodate the data scarcity and class imbalance issues. Geng et al.¹³ designed Multi-Channel Fusion CNN (MC-ResNet) architecture that has multiple parallel convolutional branches to extract the features. Extracted features were concatenated and fed into a fully connected layer and SoftMax classifier to differentiate weld defects that are similar in their visual appearance.

Tang et al. improved a ResNet101 backbone by adding a Multi-Head Attention (MHA) mechanism to better extract features in region of interest. The model used attention weights to pay attention to the important areas of the defect regions and adopted the multi-head architecture for contextual learning. Guo et al. YOLOv5 single stage detection framework using augmented data and optimized hyperparameters for end-to-end weld defect localization and classification. Similarly, Riaz et al.¹⁴ used EfficientNet-B0, PANet, and YOLOv8 detection head with multi-head attention to boost feature fusion and scalability across multiple spatial resolutions. Yadav et al.¹⁵ has designed CNN-SVM fusion model, in which the convolutional layers were used for the feature extraction and Support Vector Machine was used for classification. This hybrid configuration made use of CNN based spatial representations and SVM's discriminative capability to provide better decision boundaries. Yang and Jiang¹⁶ proposed using a unified Deep Neural Network that fused multi-level features, including handcrafted geometric features, intensity features, and contrast-based features, in hidden layers. A feature fusion mechanism in the last layer aggregated information from the different hierarchical layers for final comprehensive defect classification.

Liu et al.¹⁷ used transfer learning by adjusting pretrained CNNs, VGG16, InceptionV3, and Inception-ResNetV2, on a specialized dataset of images. The results of this study have focused on the effects of architecture

choice, depth fine-tuning and input resolution on domain adaptation, which provides methodological insight for use in weld defect classification. Ajmi et al.¹⁸ showed the application of the data augmentation and channel substitution technique, where color channels were replaced by the Canny edge and adaptive Gaussian threshold maps to improve the extraction of the structural features in the weld x-ray images. Malarvel and Singh¹⁹ explored a three-stage pipeline process of preprocessing using modified anisotropic diffusion, segmentation using improved Otsu thresholding algorithm, and classification using multi-class SVM with Gaussian radial basis kernel. Hou et al.²⁰ introduced an automatic detection pipeline of the weld seam with image processing to extract the weld seam, DNN classification for patch classification, and the sliding-window algorithm for full-image inspection, which realizes the automatic feature learning of image data with the radiography data. A summary of the related approaches based on deep learning is given in Table 1.

From ANN and CNN architectures, methodologies have been advanced from ANN and CNN architectures to attention and transformer-based integrated methodologies. CNNs (ResNet, VGG16) give good localized feature extraction and attention and transformer models give good understanding of defect regions. Overall, the literature shows consistent progress toward transformer-integrated models, but there is still a lack of focus on misclassification explainability, the focus of this study. The existing research goes further by using a hybrid model CNN-ViT model, with the focus on misclassification analysis and good performance with RIAWELC data set to solve the existing gap on explainable and context-aware weld defect classification.

Even though significant advances have been achieved in the field of weld defect classification using deep learning, the current research has many open limitations that limit their usefulness in safety-critical industrial processes. The bulk of the previous literature focuses on aggregate measures of performance without paying much attention to systematic misclassification phenomena, especially between visually similar defect groups. Moreover, weak incorporation of global contextual reasoning and weak connection between explainability mechanisms and the real error reduction limit the feasibility of the existing models. In order to strictly position the motivation and novelty of the presented research, Table 2 presents the main issues that have been discovered in the literature, the relevant research gaps, and the solutions that the proposed hybrid CNN-ViT framework introduces.

Author & Year	Methodology/model used	Dataset/imaging source	Performance metrics	Key findings/remarks
5	ResNet50-based CNN with augmentation and regularization	RIAWELC, GDXray, Private dataset	RIAWELC = 98.75%, GDXray = 90.26%, Private = 75.83%	Demonstrated dataset quality's impact; strong generalization with stratified CV
11	Artificial Neural Network (ANN) (100 hidden layers, ReLU, L-BFGS-B)	1,288 processed weld images	Accuracy = 97%	Achieved high precision and generalization across multiple defect types
12	Meta-Learning+Transfer Learning (MF algorithm) with MixUp augmentation	RIAWELC dataset	Accuracy = 92%	Combined meta-learning and fine-tuning improved few-shot performance; effective under limited labeled data
13	Multi-Channel Fusion CNN (MC-ResNet)	Radiographic weld images	mAP = 76.37%, Recognition = 97.8%	Parallel feature fusion improved distinction between visually similar defects (porosity vs slag inclusion)
14	Multi-Scale EfficientNet-B0+PANet+YOLOv8+MHA	RIAWELC, GC10-DET, LIB datasets	Accuracy > 98%, Recall/F1 ↑ 8–10%	Multi-scale hybrid improved interpretability and recall; real-time GPU inference suitable for industrial systems
15	CNN-SVM hybrid fusion	4,880 labeled images (5 classes)	Accuracy = 96.92%	Combined CNN feature extraction with SVM classification; robust multi-class recognition
16	Unified Deep Neural Network (DNN) with multi-level feature fusion	Weld radiographic images (11 handcrafted features)	Accuracy = 91.36%	Multi-level feature fusion across hidden layers enhanced classification; pre-training improved small-sample generalization
17	Transfer learning using VGG16, InceptionV3, Inception-ResNetV2	Magnifying Narrow-Band Imaging (M-NBI) medical dataset (methodology relevant for welds)	Accuracy = 97.60	Showed fine-tuning pre-trained CNNs boosts performance on small specialized datasets; findings applicable to weld defect domains
18	CNN with data augmentation and Canny-edge/Binary Gaussian channel substitution	Small weld X-ray dataset	Accuracy = 95%	Demonstrated that channel manipulation and augmentation improve defect recognition on limited data
19	Multi-Class Support Vector Machine (MSVM)+image preprocessing (anisotropic diffusion) and segmentation (improved Otsu)	X-radiography weld images	Accuracy = 96%	Detected porosity, slag, cracks, and lack of penetration; MSVM outperformed ANN/Bayes; limited by handcrafted features
20	Deep Neural Network (DNN) with sliding-window patch classification	X-ray weld images (industrial)	Accuracy = 97.20	Introduced early DNN-based automatic detection framework; eliminated manual feature extraction; achieved reliable classification through patch-based learning
21	YOLOv5 (object detection) with augmentation and hyperparameter optimization	Augmented weld image dataset	Accuracy = 97.92%	Real-time defect localization with strong detection reliability for industrial use
22	ResNet101+Multi-Head Attention (MHA)	JPEGWD dataset	Recall = 0.9634, Precision = 0.9875, mAP = 0.9817	Attention modules enhanced ROI focus; reduced false positives; high defect discrimination
Proposed Study	Hybrid CNN-ViT vs Custom CNN	RIAWELC radiographic dataset	CNN = 97.90%, CNN-ViT = 98.56%	Hybrid model combines CNN locality and ViT global context; reduced confusion between similar defects (cracks vs lack of penetration)

Table 1. Comparative summary of recent deep-learning approaches for weld defect detection and classification.

Observed problem in literature	Identified research gap	How this study addresses the gap
Performance evaluation focuses primarily on overall accuracy and mAP	Lack of systematic analysis of class-wise misclassification behavior	Performs detailed confusion matrix and per-class misclassification analysis
Visually similar weld defects are frequently confused	Insufficient focus on reducing inter-class confusion	Targets misclassification reduction using a hybrid CNN-ViT architecture
CNN models emphasize local textures but lack global context	Limited modeling of long-range spatial dependencies	Integrates transformer self-attention for global contextual reasoning
Explainability is claimed but not empirically validated	Weak linkage between attention mechanisms and error behavior	Demonstrates measurable error reduction attributable to attention modeling
Automation-centric inspection systems dominate	Limited alignment with Industry 5.0 human-centric reliability goals	Frames misclassification reduction as a trust-enhancing decision-support tool

Table 2. Problem–gap–solution mapping highlighting the limitations of existing weld defect classification methods and the contributions of the proposed CNN–ViT framework.

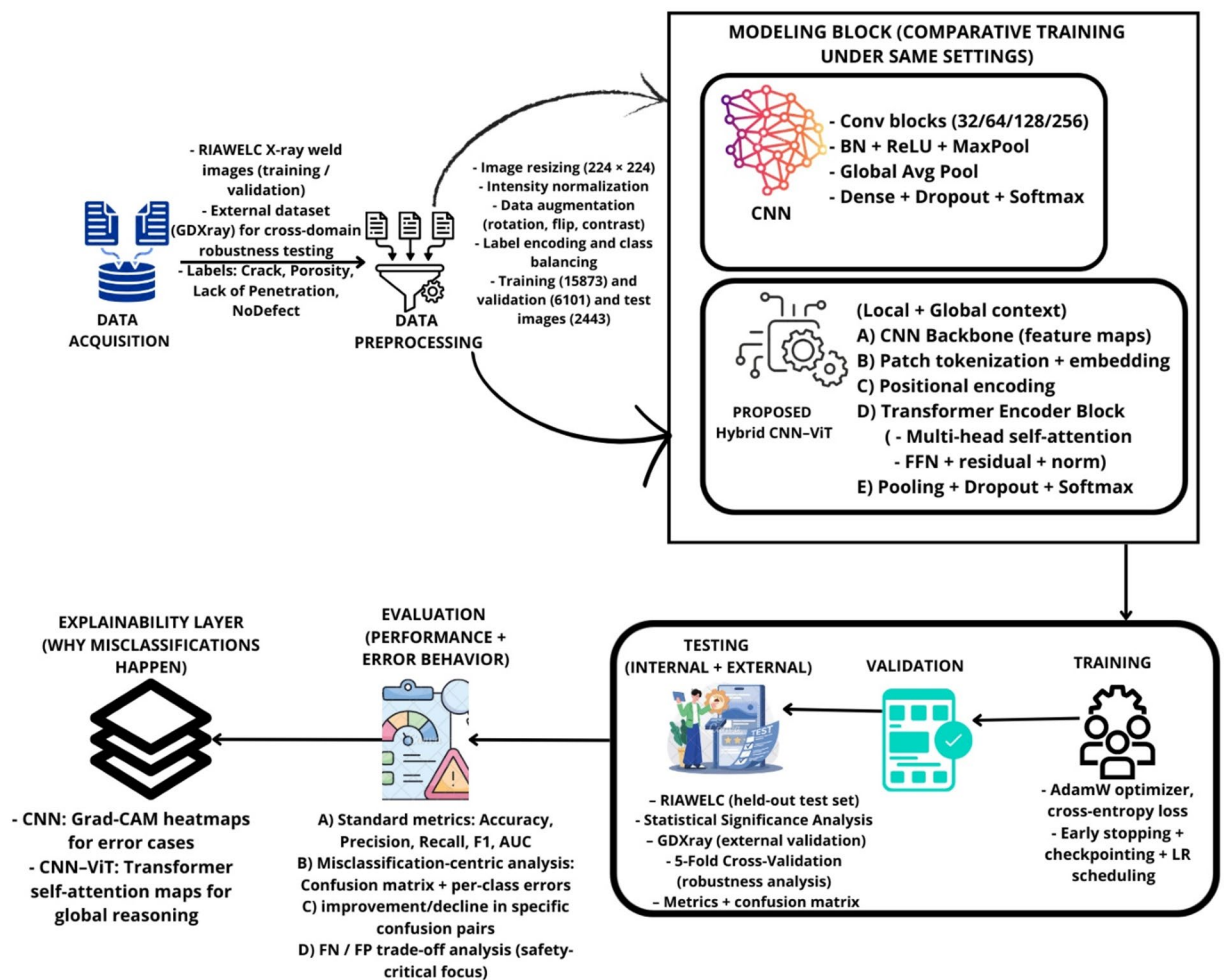


Fig. 1. Overall system architecture of the proposed weld defect classification framework.

Methodology

Figure 1 depicts a general workflow of the proposed system, including data preprocessing, comparative model training, explicit training-validation-testing strategy, evaluation, and explainability analysis.

Dataset

The RIAWELC weld X-ray database^{23,24} was used in this study and contains annotated radiographic images of weld joints that were used in various types of defects. The data set was broken into three, training, validation and testing data sets. This kind of partitioning ensured that each model is optimized, the hyperparameters are optimized and the final performance is assessed using independent subsets of data in order to minimize

additional performance reporting bias. The data set is of high-resolution weld x-ray images with the following properties:

- Image resolution: reduced to 224×224 pixels as an input to the model.
- Training (15,884) and validation (6101) and test images (2432).
- Color space: normalized grayscale radiographs.
- Types of defects: cracks, porosity, penetration absence and no-defect.
- There were predefined training, validation and test subsets to enable reproducible experimentation.

Lack of Penetration and Lack of Fusion in the context of welding metallurgy are different forms of defects. Lack of Penetration is a term used to describe the partial penetration of the weld metal through the joint thickness thereby leaving an inadequate bonding to the root. Lack of Fusion on the other hand refers to poor metallurgical bonding of the weld metal and the base material or between consecutive passes of the weld. Only Lack of Penetration defects are explicitly labeled in the RIAWELC dataset; that is why this research study is limited to the given type of defect to achieve consistency in labels, reproducibility, and consistency with ground-truth definitions in the dataset.

There was observed imbalance in classes, where there were underrepresented classes. To overcome that, data augmentation (rotation, translation, flipping, contrast adjustment) was used in the course of training.

Model architectures

Custom lightweight CNN

The initial model used this research is a specially designed Convolutional Neural Network (CNN) which has been designed to act as a strong baseline to be used in a comparative analysis. This architecture was specifically developed to achieve stepwise hierarchical extraction of the features and effective multi-classification of the radiographic defect images. It takes in RGB input images of dimensions $224 \times 224 \times 3$ in the network. The process of feature learning commences by a starting convolutional block consisting of 32 filters in a 2D convolutional layer with a kernel size of 3×3 , a stride of 1, and same padding and a ReLU activation function. Max-pooling layer (2×2) is then used to decrease spatial dimensions and maintain salient low level visual features like edges and corner patterns.

To allow a more semantically-expressive representation, three consecutive convolutional layers of 64, 128 and 256 filters are then used. The structure of convolution, batch normalization, and max-pooling is always consistent in each stage and enables the propagation of gradients, as well as, the learning of more complex features, to proceed in a stable manner. The layers store both the mid-level features (e.g., textures and contours) and high-level and defect specific features such as irregularities in the shape or characteristic patterns of various defect classes. After convolutional hierarchy, the global average pooling layer used to reduce the spatial information to a small feature vector. This operation has a major effect on minimizing the trainable parameters, improving translation invariance, and reducing overfitting. The classification head is a fully connected dense layer of 512 neurons and ReLU activation and followed by a dropout layer (dropout rate = 0.5) to enhance the generalization performance. The output layer has N neurons, which are the number of categories of defects and the activation function is the SoftMax to produce a normalized distribution of probabilities of all classes.

The function that is used to calculate the class probabilities (the SoftMax) can be defined as follows:

$$\hat{y}_i = \exp(z_i) / \sum_{j=1}^N \exp(z_j) \quad (1)$$

where z_i represents the logit score for class i , $\hat{y}_i$ is the predicted probability for that class, and N denotes the total number of classes.

Hybrid CNN-ViT

The second one was a hybrid convolution neural networks and ViT with the advantage of both local feature extraction and global context modeling. This was allowed by starting with a convolutional backbone so as to acquire low level spatial patterns by 4 successive convolutional blocks with a 32, 64, 128 and 256 filter size. Each block contained Convolution, ReLU activation, batch normalization and maximum pooling, so it was able to identify the successively sophisticated feature, like edges, textures and defect-specific shapes. The extracted feature maps were then divided into patches not overlapping any other and subsequently flattened and projected to an embedding space which is understandable by the transformer. To maintain the positional relationship between the patches, positional encoding layer was implemented in advance of feed the sequence to multiple blocks of transformer.

Each transformer block had a multi-head self-attention mechanism that represented long-range dependencies. It is done by letting every patch perform the maintenance of all the other patches and feedforward layers as well as normalization and residual connections, which stabilize optimization. The self attention operation was coded as:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (2)$$

Q, K, V are the query, key, and value matrices, d_k is the dimension of the key vectors. Once the blocks of the transformer have been traversed, the output sequence is summed with global average pooling layer to form a

small representation. In order to combat the overfitting issue, dropout regularization is employed, followed by a fully connected dense layer with SoftMax activation to get the final class probabilities.

Training strategy

The two models were trained in the same conditions in order to make a fair comparison. AdamW optimizer was applied with a learning rate of 1×10^{-4} and the weight decay of 1×10^{-4} . The categorical cross-entropy loss was used:

$$L = - \sum_{i=1}^N y_i \log(\hat{y}_i) \tag{3}$$

y_i is the actual one-hot encoded label and $\hat{y}_i$ stamp for class i is the probability prediction. It was trained up to 50 epochs with the following protection:

- Model checkpointing—save the lowest loss of validation model.
- Early stopping- termination of training after 10 epochs without any improvement.
- Learning rate scheduling—the learning rate was adaptively decreased as validation performance leveled off.

Evaluation

Both of the models had been trained, and then they were tested on the held-out test set and given special attention on misclassification behavior as opposed to the overall accuracy. Traditional indicators such as precision, recall, F1-score and AUC were however calculated to give baseline performance indicators but the primary analysis was the failure of the models. The confusion matrices were made to illustrate the patterns of errors made among different classes in order to determine the classes that were confused with other classes. Moreover, individual classification report histograms revealed a lack of strength in each of the classes, and we could therefore detect systematic misclassification patterns and not individual errors. This analysis provides deeper insight into systematic misclassification patterns, rather than isolated prediction errors. In addition, false negative and false positive behaviors were analyzed to reflect the asymmetric risk profile of industrial inspection tasks.

Comparative rationale

Performance trade-offs can be organized with the help of the usage of two architectures, which are an optimized CNN and CNN-ViT hybrid model. The CNN is a model that is computationally efficient and lightweight and whose implementation is founded on the concept of hierarchies of convolutional filters to extract features. The hybrid CNN-ViT, on the other hand, takes advantage of the ability of convolutional processing methods to extract local patterns, and the ability of transformer-based self-attention methods to extract long-range dependencies among image regions. When comparing these two models, under the same training and evaluation circumstances, they can be used to identify the strengths and weaknesses of pure convolutional models as compared to hybrid ones. Not only can the effectiveness of the optimized CNN be validated as a baseline using this comparative framework, but also it is possible to demonstrate whether or not the incorporation of transformer modules can be used to realize measurable changes in the classification accuracy, the generalization and robustness. The variations in the structure of the two models are shown in Table 3. The custom CNN with approximately 0.52 million parameters is small and fast, and can be applied to the situation that requires fast inference. Conversely, hybrid CNN and hybrid CNN-ViT are detector architectures constructed using a mixture of various network architectures hybrid CNN-ViT uses transformer encoder blocks following the convolutional feature extraction, which adds 3.3 million parameters. Such complexity allows the hybrid model to simultaneously learn the local spatial features and long-range dependencies that are highly beneficial to the differentiation of visually similar

Component	Proposed CNN (baseline)	Hybrid CNN-vision transformer (ViT)
Input dimensions	224 × 224 × 3	224 × 224 × 3
Feature extractor	4 Hierarchical Blocks: • Conv2D (3 × 3) + ReLU • Batch Normalization • MaxPool (2 × 2) • Filters: 32 → 64 → 128 → 256	CNN Backbone: • Layers 1–13 of Baseline CNN • Output Map: 14 × 14 × 256
Intermediate processing	Local Operations: Spatial downsampling via pooling to reduce dimensionality	Global Operations: • Patch Tokenization (4 patches) • Linear Projection (D = 128) • 8 Transformer Blocks (MSA + MLP)
Feature aggregation	Global Average Pooling 2D Reduces (14,14,256) → (256)	Global Average Pooling 1D Aggregates 4 token embeddings
Classifier	• Dense (512 units) • Dropout (0.5) • Softmax (4 classes)	• Dropout (0.5) • Softmax (4 classes)
Total parameters	~ 523,012 (0.5 Million)	~ 3,317,124 (3.3 Million)
Inference characteristic	High throughput, low latency	Robust handling of complex textures

Table 3. Architectural comparison between the Custom CNN and Hybrid CNN-ViT models.

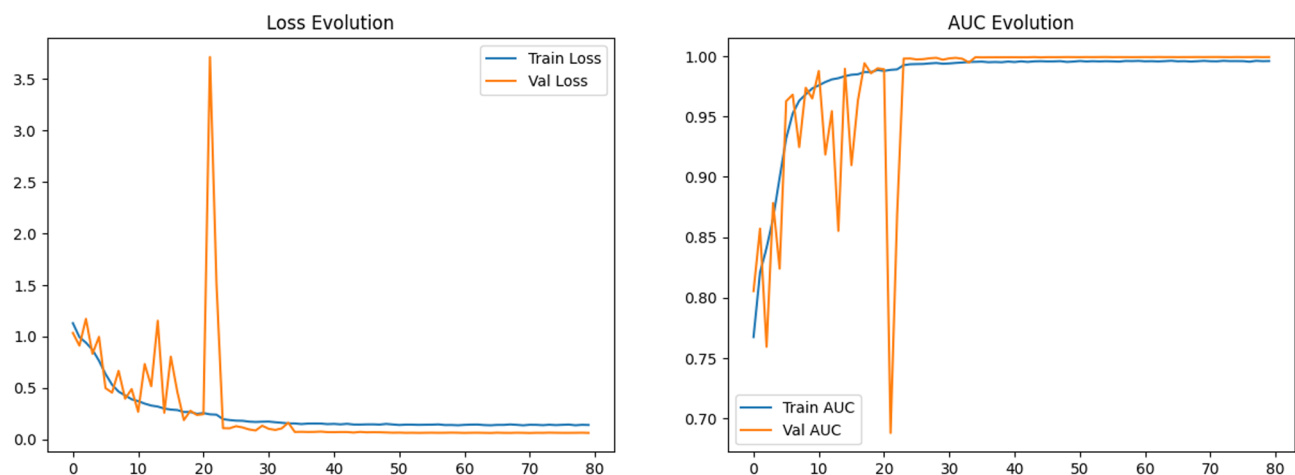


Fig. 2. Training and validation loss and AUC curves for the custom CNN model with smooth converging and stable generalization.

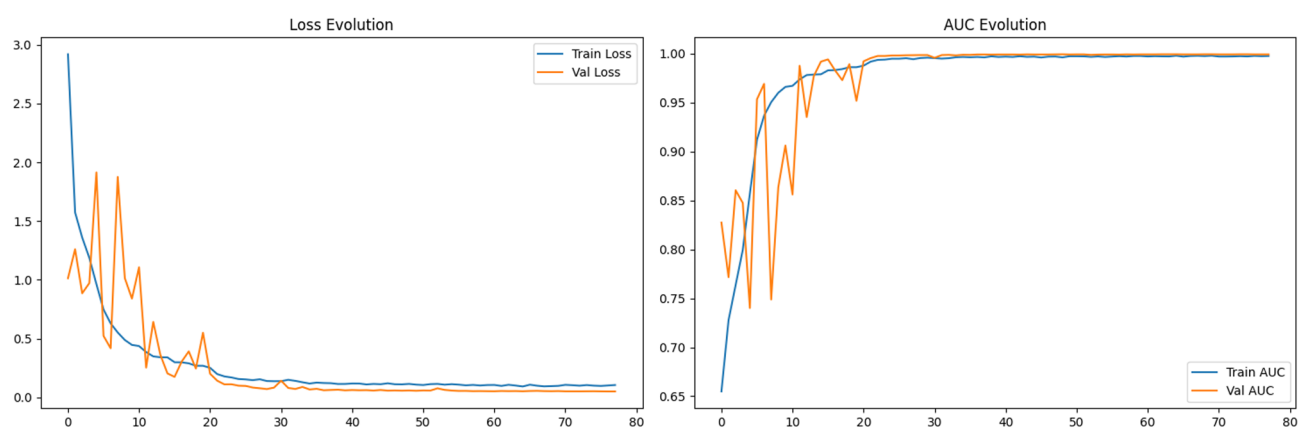


Fig. 3. Training and validation performance of the Hybrid CNN-ViT model, showing faster convergence & improved stability.

categories of defects. The table by displaying side by side both architectures brings out the trade-off between computational efficiency and better representational power.

Results and discussion

Training dynamics and convergence analysis

The findings of the present study provide the evidence of the performance disparities between the Custom CNN and the hybrid CNN-ViT models for the automated classification of the weld defects. To gain a better insight into their functioning, we examined the dynamics of their training, the error patterns and accuracy.

Custom CNN model (Fig. 2) has training dynamics that show a stable convergence behavior. Training and validation loss reduces quickly at the first epochs and slows down in the later epochs at low value and the validation loss changes closely related to the training loss at the early stages of fluctuation. This means that the learning and overall good generalization occurs without continuous overfitting. The AUC curves against the corresponding epochs depict a steep increase in the discriminative performance, and training and validation AUC values exceeded 0.98 in the initial few epochs but remained close to 0.99 afterward and indicates a strong class separability during both training and validation processes.

A similar learning trend is recorded with the hybrid CNN-ViT model (Fig. 3), albeit a slightly more expressive one. Although loss curves show slight oscillations in validation in early epochs, they all tend to smooth to a lower and more stable plateau as the training progresses. Fast feature learning is emphasized by AUC evolution, which is validated with an AUC of about 98–99% in the initial 15–20 epochs and then continues to give high values thereafter. This is the characteristic of the hybrid model that merges CNN and transformer-based global context modeling to extract local features of the object and model the entire object, respectively, leading to quick convergence and consistent correct generalization.

Class	Precision (%)	Recall (%)	F1-score (%)
Crack	99.00	98.00	98.00
Porosities	99.00	97.00	98.00
Lack of penetration	94.00	98.00	96.00
No defect	99.00	100.00	99.00
Overall accuracy (%)			97.90

Table 4. Performance evaluation of the proposed custom CNN model for weld defect classification on the test dataset.

Class	Precision (%)	Recall (%)	F1-score (%)
Crack	99.00	98.00	98.00
Porosities	99.00	99.00	99.00
Lack of Penetration	96.00	98.00	97.00
No defect	100.00	100.00	100.00
Overall accuracy (%)			98.56

Table 5. Performance evaluation of the proposed Hybrid CNN-ViT model for weld defect classification on the test dataset.

Quantitative performance evaluation on the test set

The effectiveness of the proposed models is additionally supported by the performance on the test dataset. The Custom CNN had an overall classification accuracy of 97.90% as indicated in Table 4 and it shows high and valid predictive accuracy in all types of weld defects. Most classes had precision and recall values of over 95% meaning sensitivity and specificity are balanced regardless of the variability inherent in the radiographic weld images. Specifically, the model demonstrated high quality in the detection of Crack and Porosities defects with both F1-score of 98%, and Lack of Penetration has a marginally lower but strong F1-score of 96%. The No Defect category was rated as having close to perfect reliability, indicating the good discriminatory power of the model.

Table 5 also provides the comparative results of the hybrid CNN-ViT model. The overall accuracy of this model was higher at 98.56% and there were steady improvements in terms of accuracy of the model in terms of precision and recall and F1-scores. It is worth mentioning that the Porosities and No Defect categories reached the score of perfection or close to it, whereas the Lack of Penetration class demonstrated better precision (96%) and F1-score (97%). These gains show that the transformer-based global attention mechanism improves feature representations, especially when the defect types share similar visual features and this leads to better and more stable classification.

Confusion-matrix-based misclassification analysis

The confusion matrices as a method of analyzing the error give a better understanding of the capabilities and weaknesses of the two models. The Custom CNN as shown in Fig. 4 depicts a robust diagonal dominance, which proves that the overall classification reliability is high. Nevertheless, the greatest misunderstanding is that between Cracks and Lack of Penetration. In particular, 16 crack samples were incorrectly identified as lack of penetration, and 14 porosity samples were also identified as lack of penetration, which shows that long and non-uniform intensity distributions remain a problem to purely convolutional feature extraction. The porosities and Cracks are also somewhat confused at a very lower frequency. Notably, the No Defect category is reported with close to perfect accuracy (588 corrects and just one false positive), which reduced false alarm occurrences when used in the industrial inspection settings.

The confusion matrix of the hybrid CNN-ViT model (Fig. 5) shows that the number of misclassifications was reduced significantly. There is a rise in correct predictions in all the defect classes with the largest correct prediction of Porosities; the misclassification to Lack of Penetration drops to 14, 6 samples respectively. In the same way, the mixing of Cracks and Lack of Penetration is minimized (16 down to 13 samples). These advances suggest that the global attention mechanism of the transformer boosts the contextual perception and eliminates the non-significant errors. However, the concept of Cracks and Lack of Penetration still contains a certain amount of ambiguity, which highlights the inherent visual similarity of the two types of defects in radiology images. This confirms that even with the huge enhancement of robustness brought by hybrid architectures, there are defect ambiguities that are inherently difficult even with sophisticated deep learning models.

Comparative error reduction and class-level analysis

Figure 6 shows a performance delta heatmap that shows the differences in the prediction performance of the hybrid CNN-ViT model and the baseline CNN by classes. The cells indicate the total change in the count of samples (hybrid CNN-ViT vs CNN) per actual predicted pair of classes. A positive value denotes instances where more samples were rightly classified by the hybrid model as compared to the CNN, but a negative value represents those samples that were misclassified or further misclassified. The heatmap shows that there

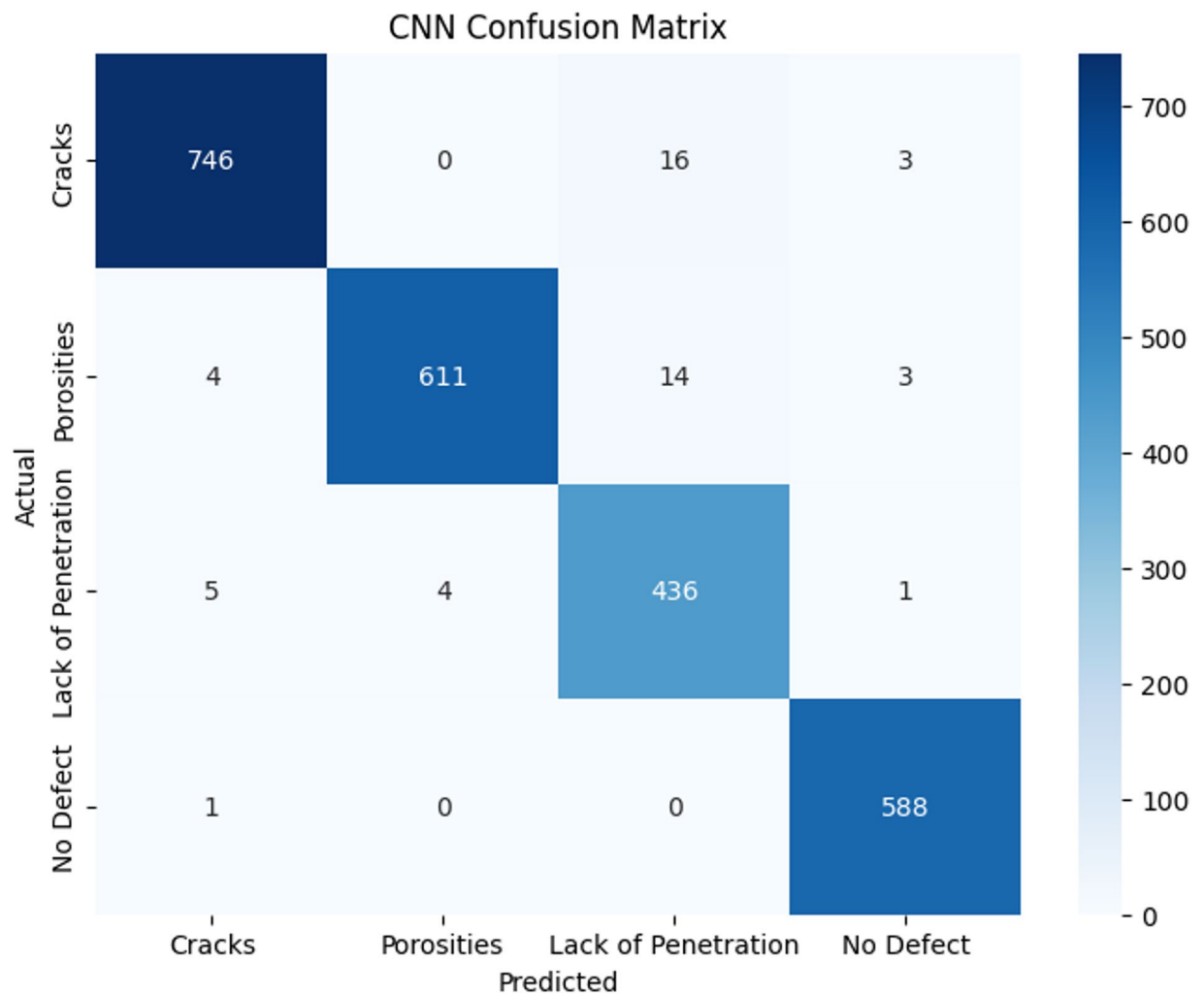


Fig. 4. Confusion matrix illustrating the CNN model's misclassification patterns across four defect classes.

are noticeable improvements in the direction of the diagonal, specifically on the Porosities class of +13 and Cracks of +4, which showed improved true positive recognition in the hybrid architecture. It is also important to note that there are great decreases in acute off-diagonal confusions, in particular, there is a reduction in Porosities misclassified as Lack of Penetration (–8) and Cracks misclassified as Lack of Penetration (–3). These results affirm that global attention mechanism transformer-based method is effective in enhancing contextual differentiation of visual similar defect patterns. The fact that there are no alterations in the No Defect class also suggests that the hybrid model does not compromise high reliability when pursuing defect free welds against false alarms in industrial inspection.

Comparing the two models to each other side by side in Fig. 7 indicates that, despite the similarities in error patterns of the two architectures across defect categories, the overall misclassification patterns of the two models are noticeably different. The hybrid CNN-ViT model yields significantly lower misclassifications of most classes of defects although it is equally robust to non-defective samples. The strongest improvement is noticed in the class of Porosities; misclassifications decrease to 8 (hybrid CNN-ViT) down to 21 (CNN) which is a significant decrease by of about 62%. This enhancement shows the increased capability of the hybrid model to capture the clustered and circular visual properties of porosity defects. The substantial decrease is also noted regarding the Crack category, where errors have been reduced by 19 to 15 which could signify better discrimination of long-shaped crack patterns.

A significant observation in Fig. 7 is that the hybrid CNN-ViT model has slightly higher misclassification of the Lack of Penetration category (11) than the standard CNN (10). Though this difference is little, it is significant and it should be interpreted. Such behavior could be explained by the fact that the attention mechanism of the transformer is more vulnerable to changes and is more concerned with the long-range spatial dependencies and contextual relationships in the radiograph as a whole. The signatures of a defect in Lack of Penetration are often in the form of subtle, low-contrast linear discontinuities, closely desired to what are variations in background intensity or long lines such as crack-like features.

Whereas the CNN is particularly more dependent on localized texture and edge information, and therefore tends to make conservative predictions, the hybrid structure can over-contextualize the relatively small linear

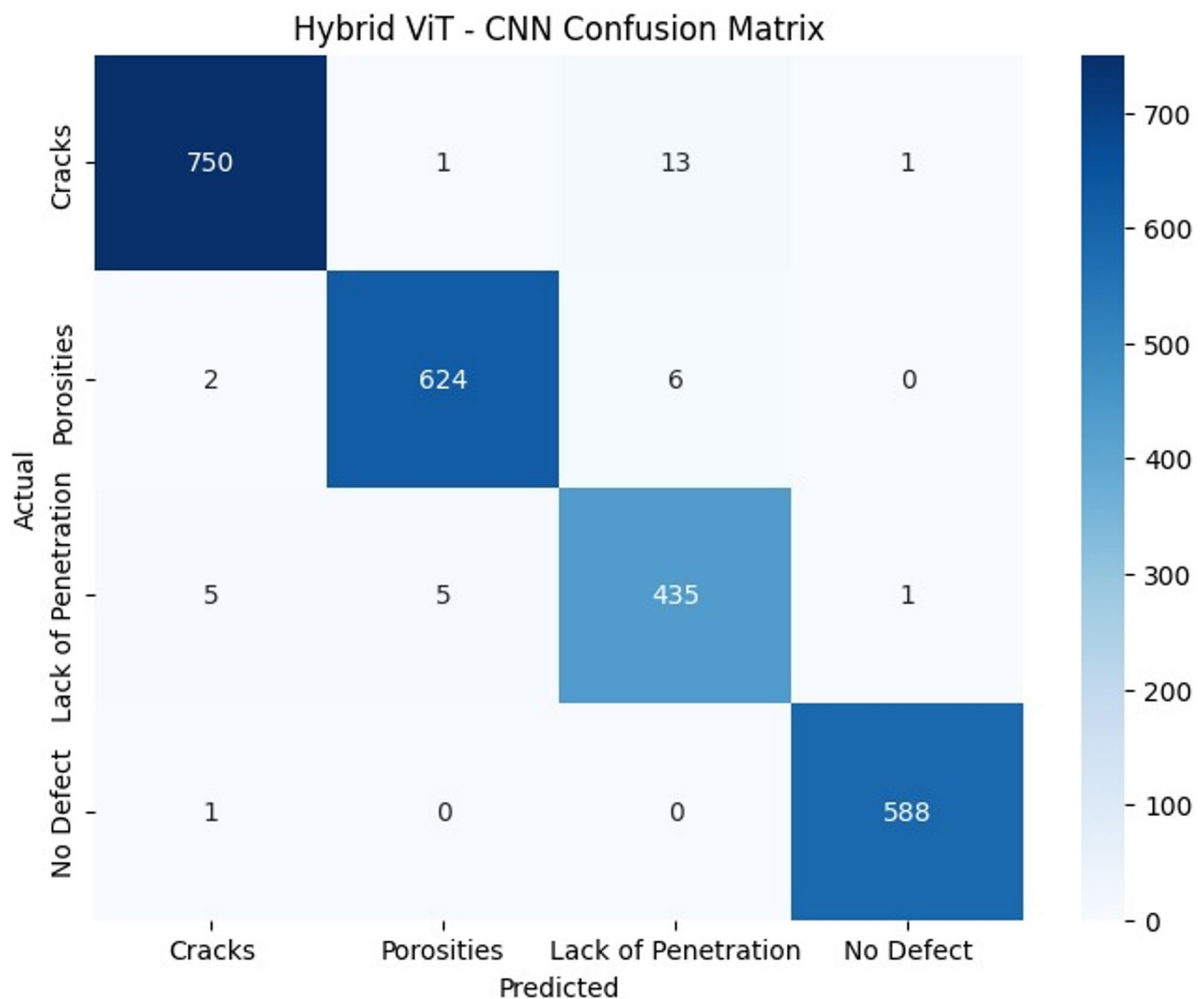


Fig. 5. Confusion matrix for the Hybrid CNN-ViT model, showing reduced cross-class confusion compared to the CNN.

details and make them relate to the similarity of defects to the similar visual defects, which are represented by the similarities. This, by implication, can bring with it a minor confusion to this very fault, although discrimination is enhanced throughout in other categories like Porosities and Cracks. Notably, the seen increase is only a minor one and does not negatively reflect on the overall superiority of the hybrid model, but instead, points to ambiguity inherent in the (Lack of Penetration) class and the trade-off between global contextual thinking and local feature specificity.

Notably, both models have a near one hundred percent reliability in the No Defect category with a single misclassification in each, which proves that both models are well-suited to the reduction of false alarms in industrial inspection processes. In general, the findings in Fig. 6 validate that the hybrid CNN-ViT model is more balanced and robust to error, especially on the visually complex defect classes, through the use of transformer-based global context and convolutional feature extraction. The reason behind such near-perfect performance is critical in an industrial sense, where false positives are reduced, leading to unnecessary and costly rework and/or rejection of otherwise-good welds. This assists in making them more practical in the real world. The drawback in the relative efficiency of the CNN is that it can be used in real-time inspection pipeline applications, but the hybrid CNN-ViT is expected to offer greater confidence in those cases where the chances of misclassification are the most significant.

In Fig. 8, the 5 most common patterns of misclassification of the five most common models are compared between the Custom CNN and the hybrid CNN-ViT on the test data. The bar chart points out the types of errors that are operationally important especially that in which there is confusion between Cracks, Porosities, and Lack of Penetration. Hybrid CNN-ViT model is consistent in the number of misclassifications in the most severe error categories. As an illustration, the misclassification of Porosities to Lack of Penetration is lowered by 14 (CNN) to 6 (hybrid CNN-ViT) samples, whereas the misclassification of Cracks to Lack of Penetration is lowered by 16 to 13 samples. These cuts imply that the hybrid model has a better capacity to retain local texture clues and long-

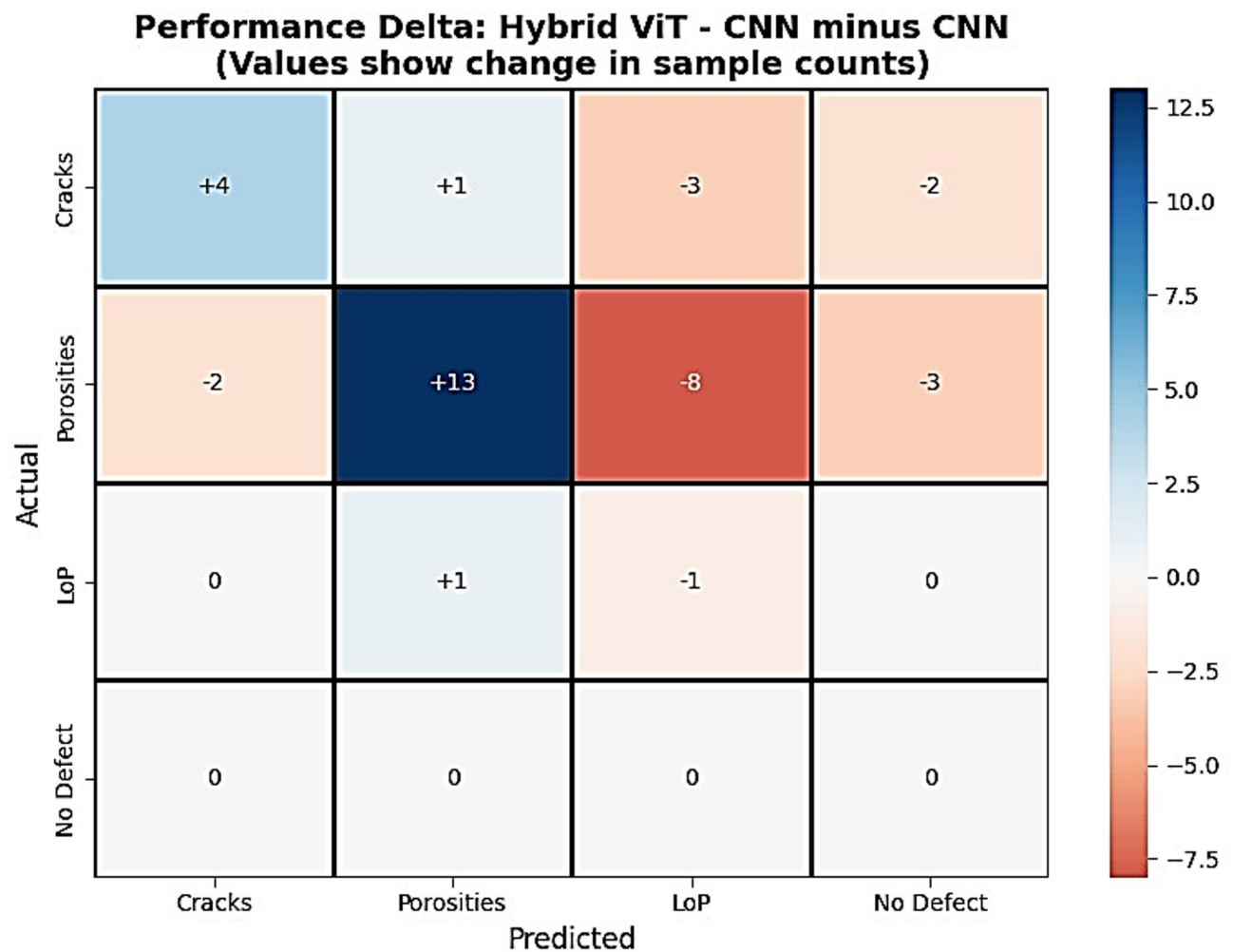


Fig. 6. Performance delta Heatmap between hybrid CNN-ViT and custom CNN.

range structural relations, which is necessary to distinguish defects that have similar radiographic presentations. Even though Lack of Penetration as predictors of Porosities shows a small but positive increase, the net change shows a significant decrease in severe misclassification against the backdrop of a rising trend, which proves the point that the hybrid CNN-ViT network can be more robust and reliable when it comes to safety-critical weld inspection challenges.

False negative-false positive trade-off analysis for industrial deployment

The cost of a classification error is not equally expensive in industrial radiographic weld inspection. Specifically, the risks of false negatives, which are the undetected flaws, are significantly stronger as compared to false positives, which normally lead to further inspection or rework. In that case, measurement of the model performance based on the overall accuracy is not enough to deploy the model safely.

In order to solve this issue, a false negative-false positive trade-off analysis was carried out based on the principle of class-wise recall statistics and confusion-matrix statistics. As presented in Figs. 4 and 5, the proposed Hybrid CNN-ViT model has shown a tendency of better recall of critical defect classes, such as cracks, porosities, and lack of penetration, which show a decreased number of defect cases that have been missed in comparison to CNN baseline. The behavior is additionally measured in Figs. 6, 7 and 8, in which the hybrid architecture shows a distinct reduction of the severe patterns of misclassification, which correlate with defect-to-defect confusion.

Notably, the Hybrid CNN-ViT does not compromise on false positives to make the system more defect sensitive. Both models have close-to-perfect performance in the No Defect category reflected by the high ability of the diagonal superiority of the confusion matrices (Figs. 4 and 5). It means that the hybrid model sustains a low false-alarm value and increases the reliability of defect detection.

Overall, this analysis confirms that the Hybrid CNN-ViT suggested is a good trade-off between false negative and false positive rates, which is crucial when the task of the inspection system is to monitor industrial safety, reliability, and trustworthiness of the operators. The proposed framework ensures that to reduce missed detections, one can also reduce false alarms, which is important in safety-critical manufacturing settings and may support Industry 5.0-oriented inspection frameworks.

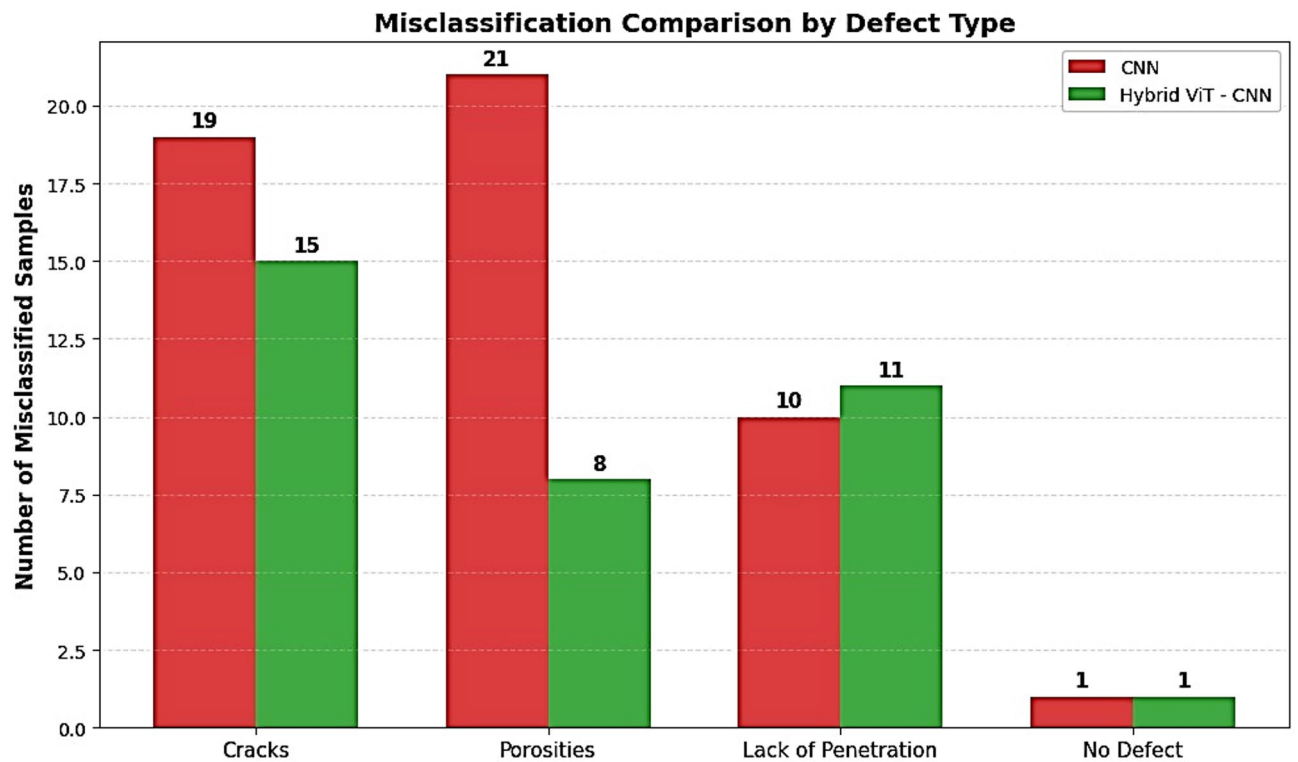


Fig. 7. Per-class misclassification counts comparing the CNN and Hybrid CNN-ViT models.

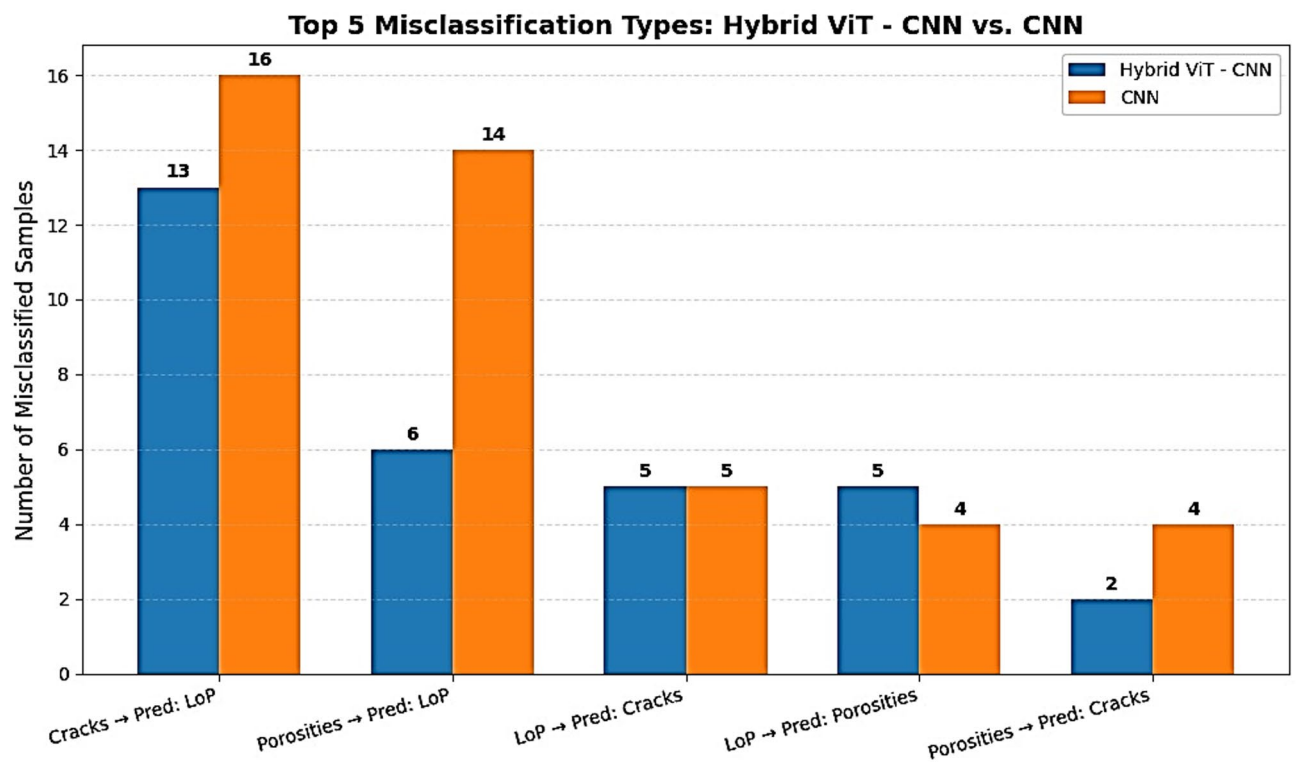


Fig. 8. Comparison of top five misclassification patterns for CNN and hybrid CNN-ViT.

Explainability-driven mechanistic validation

To present the trade-offs in performance, radar chart comparison was drawn (Fig. 9). That may also be regarded as another sign that hybrid CNN-ViT consistently has a small better precision, recall and F1-scores than CNN model that still is competitive in certain significant aspects with the additional advantage of faster inference.

Another factor to be taken into consideration is deployment context. Hybrid CNN-ViT provides better accuracy and consistency. It is more appropriate for safety critical situations which have maximum reliability as a non-negotiable. A hybrid strategy where CNN does the regular scans and hybrid CNN-ViT cross-checks the cases where there is uncertainty may be an efficient way that can be both reliable. By reducing the false alarms and finding patterns of systematic misclassification, the hybrid CNN-ViT model may contribute to Industry 5.0 objectives by supporting more transparent AI-assisted inspection workflows.

As shown in Table 6, the proposed hybrid CNN-ViT framework can be compared with the performance of the selected deep learning-based weld defect classification methods, which are reported in the recent literature. Both the overall classification accuracy and approximate misclassification rate are provided as well to provide an accurate and complete picture of predictive performance and error behavior. The suggested hybrid CNN-ViT model has the highest accuracy of 98.56%, which is equal to the low misclassification rate of ≈ 1.44 , and is better than the custom lightweight CNN (97.90%, ≈ 2.10) and the majority of current models. It performs competitively with the recent state-of-the-art studies and it also shows a stronger decrease in the number of errors in classifications. On the contrary, previous methods (such as by¹⁶ and¹⁹) record significantly larger misclassification (more than 4%), which suggests a relatively poor generalization capability in differentiating visually similar types of weld defects. It is also worth mentioning that, unlike some earlier studies that focus on aggregate accuracy as the only performance measure, this paper also uses misclassification analysis as an additional performance metric, which improves interpretability and offers more insight into the patterns of confusion on the class level. This two-metric assessment scheme highlights the overall strength and utility of the

Final Model Comparison: Metrics Overview

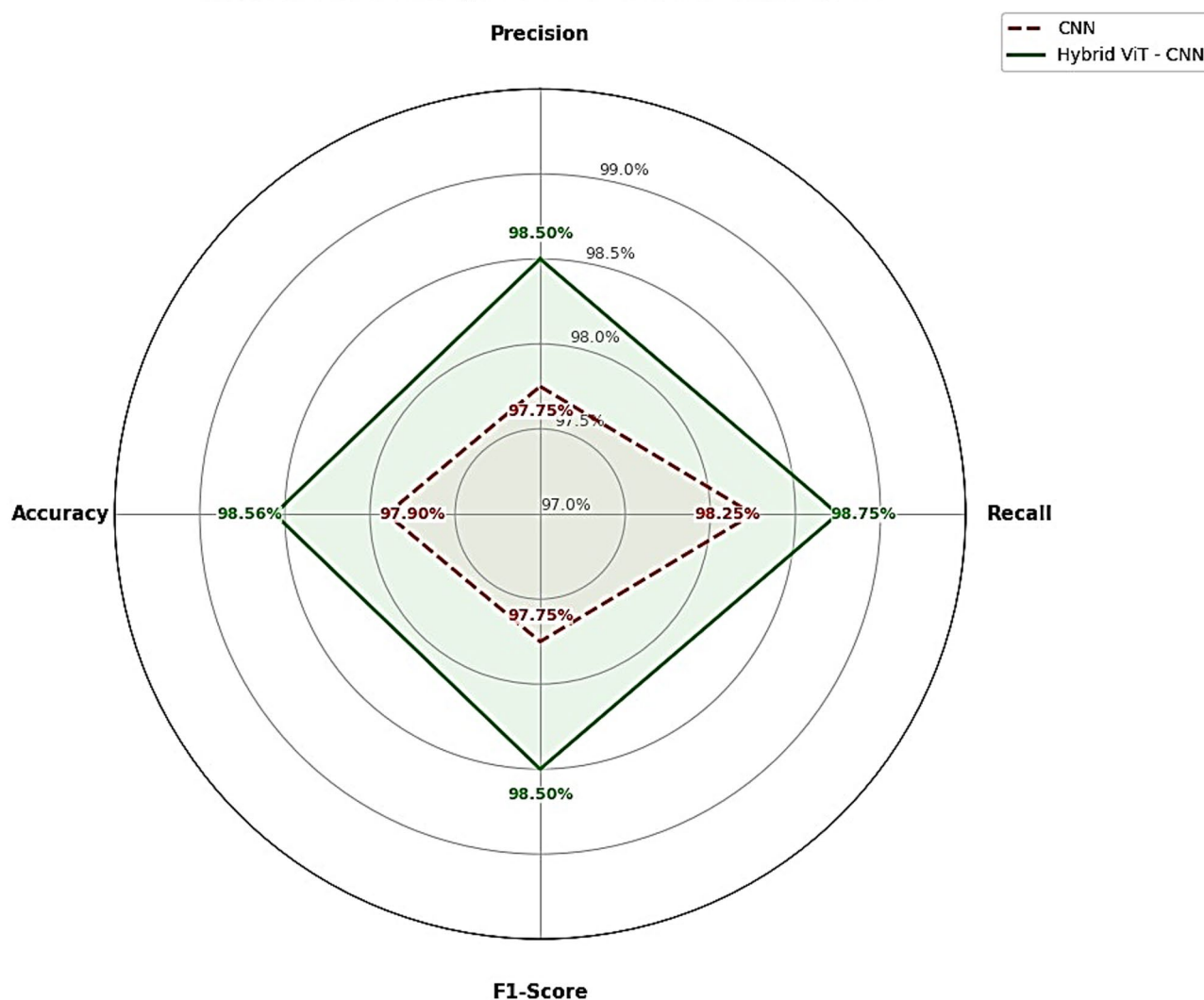


Fig. 9. Radar-chart comparison of CNN and Hybrid CNN-ViT precision, recall, and F1-scores across classes.

Author & Year	Accuracy performance metrics (%)	Approx. misclassification rate (%)
11	97.00	≈ 3.00
12	92.00	≈ 8.00
15	96.92	≈ 3.08
16	91.36	≈ 8.64
17	97.60	≈ 2.40
18	95.00	≈ 5.00
19	96.00	≈ 4.00
20	97.20	≈ 2.80
21	97.92	≈ 2.08
Custom lightweight CNN	97.90	≈ 2.10
Proposed study (hybrid CNN-ViT)	98.56	≈ 1.44

Table 6. Benchmark comparison of misclassification rates across existing deep-learning models and the proposed Hybrid CNN-ViT.

suggested hybrid CNN-ViT system, and supports the notion of its applicability to reliable, open, and human-friendly automated weld inspections, which is one of the key tenets of the Industry 5.0 oriented intelligent quality control.

Ablation study and architectural rationale

To prove the architectural design of the proposed framework hybrid CNN-ViT the structural ablation study was carried out to understand the single and combined contribution of convolutional feature extraction and transformer based global contextual modeling. This analysis aims at justifying the chosen hybrid architecture and showing how local spatial representations and global contextual reasoning can be used to complement one another in classification of weld defects.

There were three model configurations that were tested under the same training conditions:

- (i) An optimized Custom Lightweight CNN without an attention mechanism,
- (ii) A Pure ViT that is trained on unprocessed patches of the image without using a convolution filter, and
- (iii) The suggested hybrid CNN-ViT combination of the two elements.

The Pure ViT model did not converge to meaningful visual representations reaching only an accuracy of 33.0%. This result illustrates the importance of convolutional inductive bias to successfully extract features on medium-scale industrial data set RIAWELC. This finding proves the hypothesis that transformers that are trained directly cannot be data-efficient in this application area and cannot accurately detect low-level spatial structure in radiographic weld images.

In the other two settings, quantitative performance comparison is presented in Table 6. The Custom Lightweight CNN reached a high baseline accuracy of 97.90% which exhibited effective local spatial feature learning through convolutional inductive bias. Nevertheless, although the CNN-only model was highly accurate overall, it had weaknesses in identifying visually similar defect types, especially ones that involved long-range spatial computation along the weld seam, e.g. lack of penetration and porosity.

Conversely, the suggested hybrid CNN- ViT enhanced classification accuracy of up to 98.56%, as indicated in Table 6. The CNN base is effective at deriving discriminative spatial features and the transformer encoder derives global contextual dependencies throughout the weld area. This can be used to interpret defect geometry more coherently and minimize ambiguity in complicated and visually overlapping defect pattern.

This ablation study shows that neither convolutional nor transformer-based modeling can be used to classify defects in the weld using optimal results. Rather, they have a mutually supportive and equilibrium-based architecture, which justifies the design reasoning behind the proposed hybrid CNN ViT structure and justifies its aptness as a tool to be used in the safe and reliable inspection of processes in Industry.

Statistical significance analysis and hypothesis testing

Aggregate performance measures, including accuracy and F1-score, come in handy to give a general idea of the model performance, but not to show whether the performance variations have been statistically significant or may be the result of random error. To overcome this weakness, a formal statistical significance test was applied to prove the hypothesis that the suggested hybrid CNN-ViT model would provide a real enhancement over the optimized CNN baseline.

Since both models were tested using the same validation samples, the test of McNemar was used. The test provided by McNemar is a non-parametric test of statistical hypothesis specifically made to test paired nominal data and find extensive applications in assessing the predictive ability of two classifiers on the same data. The test concentrates on those samples that the two models differ, and therefore, it is suitable especially when comparing models in the supervised classification problems.

Let n_{01} represents the number of samples incorrectly classified by the CNN and correctly classified by the hybrid CNN-ViT and, n_{10} represents the count of samples that were correctly classified by the CNN and wrongfully classified by the hybrid CNN-ViT. The McNemar test statistic is defined as:

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \tag{4}$$

The null hypothesis H_0 is that both models are equal predictive models, i.e. the likelihood of misclassification is the same with both classifiers. Rejection of H_0 indicates a statistically significant difference in performance. The test was performed on the validation set comprising $N = 6101$ samples. The contingency values that resulted were:

- Hybrid CNN-ViT correct, CNN incorrect (n_{01}): 82 samples
- CNN correct, Hybrid CNN-ViT incorrect (n_{10}): 49 samples

With these values, the McNemar test gave a chi-square value of $\chi^2 = 7.817$ and a p-value 0.00518. The null hypothesis is rejected since the p-value is way below the standard significance level of 0.05. This finding validates that not only is the observed gain of the hybrid CNN-ViT model as compared to the CNN baseline statistically significant, but it is not due to chance. Notably, the hybrid CNN-ViT was able to correct 82 misclassified cases which the CNN could not solve, but with relatively fewer new errors. This statistically confirmed decrease of misclassification is a direct indication that the incorporation of transformer-based global contextual reasoning into convolutional feature extraction is effective. Combined with the independent test set results and the explainability-oriented misclassification analysis, the statistical results support the fact that hybrid CNN-ViT is a strong and capable enhancement of the conventional convolutional architectures in the safety-critical weld defect classification.

Computational efficiency and deployment feasibility

Industrial online inspection systems demand low latency inferences, as well as small model footprints. In order to assess the feasibility of deployment of the proposed hybrid CNN-ViT framework, we compared the computational efficiency of the proposed framework with the optimized Custom CNN regarding the number of its parameters, model size, as well as inference latency and throughput. The hybrid CNN-ViT has approximately 3.32 million parameters (12.66 MB), which is much more than 0.52 million parameters (2.00 MB) of the CNN baseline as summarized in Table 7. The hybrid CNN-ViT has a smaller inference latency (59.12 ms vs. 70.52 ms), and a larger throughput of 16.9 FPS, than the CNN (14.2 FPS). It means that the global attention mechanism based on transformers does not impose a prohibitive computational cost and can be effectively implemented in practice. The findings indicate that the suggested hybrid scheme can be implemented in practice in real-time or near-real-time industrial inspection, especially on the current industrial PCs or edge AI solutions. Moreover, moderate model size allows further optimization by standard deployment methods (e.g. quantization, pruning, and hardware-accelerated inference) so that the framework can be applied to embedded and resource constrained inspection systems.

Explainability-driven misclassification analysis

In order to clarify the complexity of the performance measures and the confusion-matrix statistics, the explainability-driven analysis was performed to learn about the inner decision-making process of the existing CNN and the proposed hybrid CNN-ViT model. Two visualization strategies complementary to each other were used: Grad-CAM heatmaps of CNN model and self-attention maps of transformer component. Collectively, such visualizations give an understanding of where the models focus on during prediction and the cause of some misclassification patterns. The cases of representative misclassifications generated by the Custom CNN model with the help of Grad-CAM are depicted in Fig. 10. In every instance of defect, the first row shows the original radiographic picture with the true label (T), predicted label (P), and the confidence score. The second and the third row show manually identified regions of interest (ROIs) of the image that are visually salient defects, and the third row superimposes the Grad-CAM heatmap on the original image.

Grad-CAM visualizations show that the CNN mostly depends on the localized texture activations and the intensity gradient, and frequently focuses on the attention on the small and high-contrast areas instead of the entire morphology of the defect. When the cracks are false identified as lack of penetration, the CNN only detect the cracks on short and discrete parts of the crack, not in its long and continuous structure. On the same note, porosity flaws are often misinterpreted with lack of penetration when clustered voids are observed under non-uniform background intensity because the CNN reacts to scattered point-wise activations and not the overall spatial structure. On a few samples that were misclassified, Grad-CAM makes false alarms on areas that are not related to the actual defect, suggesting that it is sensitive to the variations in the background and not the geometry of the defect. This piecemeal and regionally prejudiced focus is what causes the incessant mix-up that is evident between cracks, porosities, and absence of penetration in the CNN confusion map. Although useful in defects

Model	Parameters	Model size (MB)	Latency (ms)	FPS
Optimized CNN	~ 523,012	2.00	70.52	14.2
Hybrid CNN-ViT	~ 3,317,124	12.66	59.12	16.9

Table 7. Computational efficiency comparison.

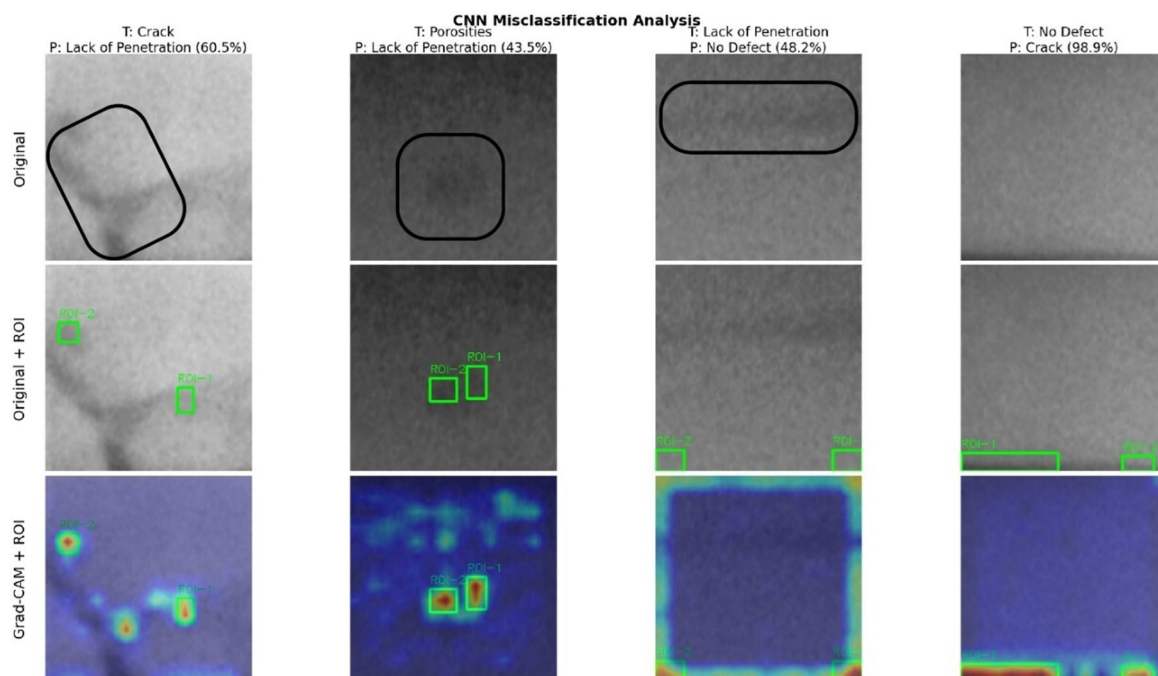


Fig. 10. CNN Grad-CAM-based misclassification analysis for radiographic weld defect classification.

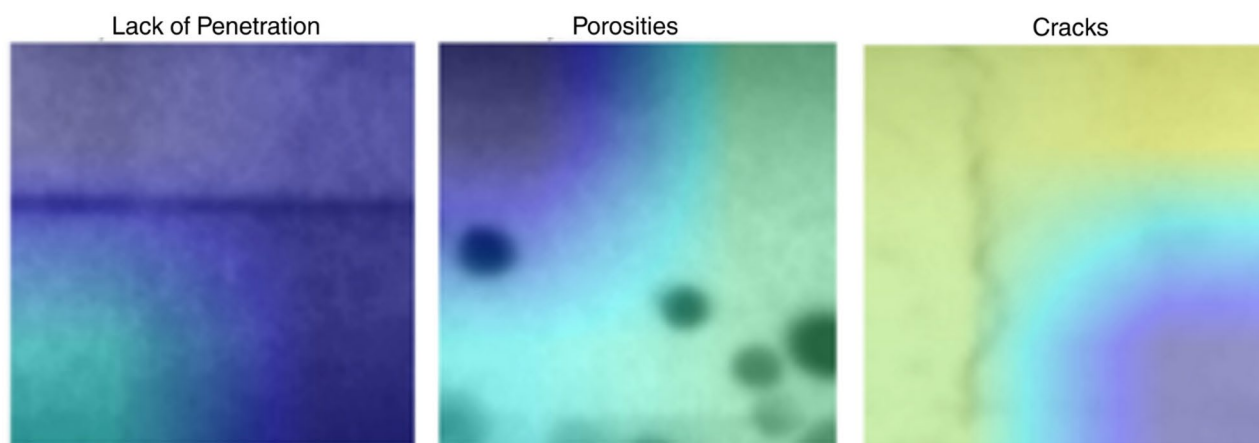


Fig. 11. Class-wise transformer self-attention maps of the Hybrid CNN-ViT model for weld defect interpretation.

with a strong local contrast, the CNN lacks the capability of bringing out bigger contextualization to aid in robust discrimination in more complex radiographic conditions.

In Fig. 11, the attention maps of the different classes were obtained with the help of the transformer part of the hybrid CNN-ViT model on correctly classified lack of penetration, porosity, and crack samples. In contrast to Grad-CAM, which shows gradient based importance in the convolutional layers, transformer attention maps are maps that visualize world-wide self-attention correlation among image regions and this shows how spatial interdependencies are modeled across the whole image.

To loss of penetration defects, activations of the attention maps are broadly distributed horizontally with respect to the weld seam, suggesting that the model represents low-contrast discontinuities over large spatial scales. The focus in the case of porosity defects is centered around several spatially clustered circular patterns, which indicates an appreciation on clustered void patterns as opposed to isolated texture points. Crack defects show long and continuous tracks of attention following along the entire crack path which represent structural continuity not previously acknowledged by the CNN in a complete sense.

It can be seen in these patterns of attention that the transformer component effectively combines long-range contextual information, allowing the whole interpretation of defect geometry and spatial relations. Subsequently,

the hybrid CNN-ViT model addresses the weaknesses of CNN Grad-CAM visualizations, and the degree of inter-class confusion is decreased, as well as the robustness to visually overlapping defect types is enhanced.

When combined, Figs. 10 and 11 provide supplementary information about the decision-making behavior of the tested models. Grad-CAM analysis shows that attention patterns of the CNN are very local and spatial and diffuse and this is one of the factors that have enhanced its misclassification behavior. Contrarily, the transformer attention maps exhibit the ability of self-attention systems to acquire global contextual relations within the weld area. This empirical evidence justifies the use of a Vision Transformer on convolutional backbone.

The hybrid CNN-ViT frame can be used to obtain a better and more consistent representation of weld defects with a lowered amount of localization and global context modeling, including both. The mode of failure detected in the baseline CNN is alleviated directly in this integration and improves the system robustness in identifying visual similar defect classes. The findings obtained through the explainability support the interpretability, as well as the reliability of the proposed model, which can provide a technical foundation for deployment in safety-critical manufacturing environments, including Industry 5.0-oriented systems.

The explainability-based reasoning in this section offers mechanistic support of the found misclassification decrease of the hybrid CNN-ViT model. The results of Grad-CAM indicate that the only convolutional feature learning methodology has certain constraints in detecting long defect morphology, whereas the transformer self-attention maps show that global contextual modeling can make the interpretation of defect geometry of weld seams more consistent. This architectural difference is directly related to the statistical and empirical reduction in confusion between visually related defects like cracks, porosities and absence of penetration.

Defect-specific causes of misclassification

Even though both models have high total accuracy, residual errors of misclassification occur because of the inherent visual properties of weld defects in radiographic images. These are mainly due to morphological confusion of types of defects and how spatial features are acquired by various architectures.

The cracks are usually observed as elongated, low-contrast discontinuities, which are usually narrow. Instead of emphasizing the linear structure, the CNN puts a lot of emphasis on short cracks that do not necessarily represent the linear structure as a whole, confounding with lack of penetration defects that also exhibit linear structures. This ambiguity can be reduced by the hybrid CNN-ViT which models long-range spatial continuity using global self-attention, though in a few cases of very low-contrast there is some minor confusion.

Porosities are typified by circular cavities with different sizes and densities. With fragmented activations on each individual pore in the CNN, this is likely to blur out the overall spatial pattern, especially in the non-uniform background intensity case, leading to false classification as lack of penetration. The CNN-ViT enhances the discrimination process through the combination of space-based relationships among the several pores with residual errors arising primarily in abnormal or sparsity of pore patterns.

No penetration is the most challenging one as it appears low-contrast and it is located along the weld seam. These features can be confused with the CNN-based models as background-variation, or crack-like features. Although the CNN-ViT has a drawback of seam-level contextual reasoning, a low degree of confusion is still the result of the intrinsic ambiguity of this form of defect.

The defect samples of any kind do not have uniform intensity and are always detected by the two models meaning that there are no defect-free welds with a high probability of rejection and few false alarms.

This defect-based analysis shows that this visual overlap between defect morphologies is the major cause of misclassification, and not model limitations. The hybrid CNN-ViT architecture is an efficient way to remove such limitations and ensures better robustness in terms of safety-related weld inspection conditions by hybridizing local feature extraction with global contextual modeling. In order to further determine the extent to which the defect-specific behavior is generalized outside of the training distribution, the proposed model was also tested on an external dataset under different imaging conditions.

External dataset validation and cross-domain robustness analysis

External validation with GDXray dataset²⁵, a collection of radiographic images obtained under varying imaging conditions, acquisition setups and processes, was used to assess the robustness and generalization ability of the proposed hybrid CNN-ViT beyond the RIAWELC dataset. This dataset offers a strict standard of evaluation of cross-domain generalization.

It was assessed with 883 samples (227 Crack, 310 Porosity and 346 Lack of Penetration). The proposed model had a overall accuracy of 92.98 in the classification, and the F1-scores are 95.22% (Crack), 91.00% (Porosity) and 93.20% (Lack of Penetration). The associated confusion matrix (Fig. 12) reveals that the diagonal is highly dominant, which means that it remains consistent in recognizing defects in a domain where there is a shift.

Though a moderate decrease in performance is noticed against in-dataset testing, this is normal in cross-dataset testing and this fact validates that the proposed hybrid architecture is learning to transferable representation and not dataset-related features. These findings indicate that the hybrid CNN-ViT is capable of preserving a high level of discrimination even in a wide range of imaging factors that favor its usage in real-life industrial inspection systems based on heterogeneous data sources.

5-fold cross-validation robustness analysis

A fivefold cross-validation experiment on the RIAWELC data was performed to deal with the possible issues of split-dependency and to give a strict evaluation of the generalization ability of the framework. The data was divided into five mutually exclusive folds and in each run, four folds were used to train and the other fold was used as the hold-out validation folds. Table 8 gives the fold-wise classification accuracies. The model got classification accuracy of 95.07%, 95.35%, 94.94%, 95.19%, and 96.42% across the five folds respectively, with the mean cross-validation accuracy standing at $95.39\% \pm 0.53\%$ (SD). The t-distribution and 95% confidence

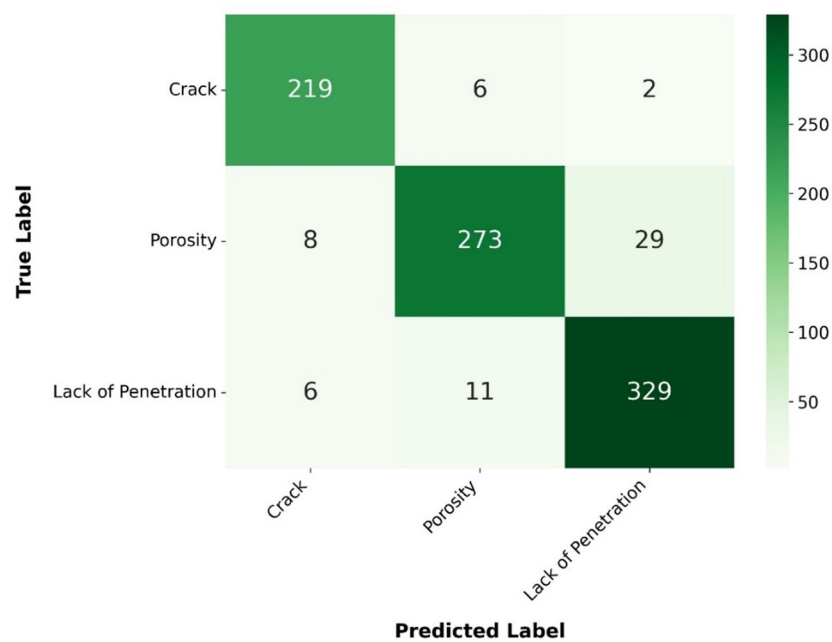


Fig. 12. Confusion matrix obtained from external cross-dataset validation on the GDXray benchmark.

Fold	Accuracy (%)	Macro F1	Weighted F1
Fold 1	95.07	0.9495	0.9506
Fold 2	95.35	0.9520	0.9535
Fold 3	94.94	0.9492	0.9493
Fold 4	95.19	0.9508	0.9518
Fold 5	96.42	0.9626	0.9641
Mean ± SD	95.39 ± 0.53	0.953 ± 0.005	0.954 ± 0.006
95% CI (accuracy)	94.73–96.05	–	–

Table 8. 5-fold cross-validation performance summary.

levels provided a confidence interval estimate of $95.39\% \pm 0.66\%$, which is a range in performance of 94.73% to 96.05%. The small value of the confidence interval shows that the predictive behavior is statistically constant over a range of data partitions. The overall misclassification rate consisted of only 4.61% (1125 samples) of the total number of samples that were assessed in all folds (24,417). The low fold-wise variability, constant class-wise F1-scores and small confidence interval all confirm the strength and the strength of generalization of the proposed framework with various training-validation combinations supporting its capability to be used to reliably inspect industrial weld defects.

Conclusion and future work

The paper involved a strict comparative analysis of a bespoke CNN and a hybrid CNN-ViT model in automated weld defect classification of radiographic images. The custom CNN also formed a solid foundation by learning low- and mid-level spatial representations in a manner that it reached a 97.90% overall classification accuracy with a misclassification rate of around 2.10%. Based on this base, this hybrid CNN-ViT model further enhanced predictive performance achieved with accuracy of 98.56% and misclassification rate of 1.44%, which represents a relative improvement rate of nearly 31% in classification errors compared to the baseline CNN.

An extended analysis of misclassification by class has shown that the hybrid architecture greatly alleviates confusion of visually similar defect classes particularly cracks, porosity, and lack of penetration. This is enhanced by the self-attention of the transformer that allows us to model longer-range spatial dependencies and global contextual relationships that cannot be adequately modeled by convolutional operations on their own. Notably, the two models were found to be near-perfect in detecting defect-free welds thus lowering the false positives and decreasing false rejection or rework of good components in the inspections processes in the industries.

Besides the higher classification accuracy, the hybrid CNN-ViT framework offers a better model interpretability due to the attention mechanisms offered by transformers, which give the framework a more consistent understanding of space and localization of defects in the background. The explainability-driven

analysis promotes building of transparent and reliable decision-support systems which is critically important to be adopted in safety-critical and human-centric manufacturing settings in line with Industry 5.0 principles.

Although these are good results, there are still some shortcomings. The dataset used is skewed in terms of classes and has a very small amount of diversity, which can be translated into residual ambiguity in particular defect classes. The next steps of work will be the expansion and diversification of radiographic datasets, ensemble and uncertainty-aware approaches to learning, and the creation of lightweight versions of transformers that can be deployed in the field of real-time. In addition, more extensive incorporation of explainable artificial intelligence systems, such as state-of-the-art attention visualization and human-in-the-loop validation will also be considered to increase further transparency, strength, and operator confidence. Moreover, fivefold cross-validation was performed to verify the strength of the proposed framework, which provided an average value of 95.39% with low standard deviation and a tight 95% confidence interval, which means that the predictive behavior of the framework remains stable across different training-validation divides.

Overall, the presented misclassification rate decrease, along with the strong inference accuracy and better interpretability, suggest the suggested hybrid CNN-ViT framework to be a stable and scalable solution to the automated inspection of the quality of welds. The enhanced misclassification control and explainability solutions as shown in this paper indicate possible applicability to the sustainable and humanistic manufacturing paradigms. Nonetheless, specific studies on human-in-the-loop validation and industrial implementation are necessary to justify complete compliance with non-destructive evaluation systems based on Industry 5.0.

Data availability

The dataset used in this study is publicly available from Totino B, Spagnolo F, Perri S (2023) RIAWELC: A Novel Dataset of Radiographic Images for Automatic Weld Defects Classification. *International Journal of Electrical and Computer Engineering Research* 3:13–17. <https://doi.org/10.53375/ijecer.2023.320>. Mery D, Rizzo V, Zscherpel U, et al. (2015) GDxray: The Database of X-ray Images for Nondestructive Testing. *J Nondestruct Eval* 34:42. <https://doi.org/10.1007/s10921-015-0315-7>

Received: 1 January 2026; Accepted: 16 March 2026

Published online: 22 March 2026

References

- Chandrasekhar, N., Muthukumaran, V. & Das, C. R. Real-time determination of weld penetration status during A-TIG welding of stainless steel employing deep learning approach. *Weld. World* **69**, 2727–2744. <https://doi.org/10.1007/s40194-025-02021-6> (2025).
- Avilés-Cruz, C., Aguilar-Sánchez, M., Vargas-Arista, B. & Escudero-Celis, G. A deep learning model to identify the initial parameters of dissimilar welded joints under impact fracture. *Weld. World* <https://doi.org/10.1007/s40194-025-01944-4> (2025).
- Biasuz Block, S., Dutra da Silva, R., Eugenio Lazzaretti, A. & Minetto, R. LoHi-WELD: A novel industrial dataset for weld defect detection and classification, a deep learning study, and future perspectives. *IEEE Access* **12**, 77442–77453. <https://doi.org/10.1109/ACCESS.2024.3407019> (2024).
- Rabe, P., Schiebahn, A. & Reisgen, U. Development and validation of a generalized, AI-based inline void defect detection solution for FSW based on force feedback. *Weld. World* **69**, 499–514. <https://doi.org/10.1007/s40194-024-01895-2> (2025).
- Palma-Ramírez, D. et al. Deep convolutional neural network for weld defect classification in radiographic images. *Heliyon* **10**, e30590. <https://doi.org/10.1016/j.heliyon.2024.e30590> (2024).
- Zhang, L. et al. Lightweight DCGAN and MobileNet based model for detecting X-ray welding defects under unbalanced samples. *Sci. Rep.* **15**, 6221. <https://doi.org/10.1038/s41598-025-89558-0> (2025).
- Liu, W. et al. A lightweight TA-YOLOv8 method for the spot weld surface anomaly detection of body in white. *Appl. Sci.* **15**, 2931. <https://doi.org/10.3390/app15062931> (2025).
- Díaz-Cano, I. et al. Automated fillet weld inspection based on deep learning from 2D images. *Appl. Sci.* **15**, 899. <https://doi.org/10.3390/app15020899> (2025).
- Contreras Ortiz, A., Santiago, R. R., Hernandez, D. E. & Lopez-Montiel, M. Multiclass evaluation of vision transformers for industrial welding defect detection. *Math. Comput. Appl.* **30**, 24. <https://doi.org/10.3390/mca30020024> (2025).
- Guo, E. & Chen, F. Heterogeneous attention multi-scale network for efficient weld seam classification. *Sci. Rep.* **15**, 12270. <https://doi.org/10.1038/s41598-025-91186-7> (2025).
- Kusuma, F. A., Fatchan, M. & Zy, A. T. Advanced ANN techniques for precise detection and classification of welding defects. *Int. J. Integr. Sci. Technol.* **2**, 516–525. <https://doi.org/10.59890/ijist.v2i5.1907> (2024).
- Xia, W. et al. Automatic identification of complex defects in pressurized vessel welds based on meta-learning-fine-tuning. *SSRN Electron. J.* <https://doi.org/10.2139/ssrn.4848587> (2024).
- Geng, H., Li, Z. & Zhou, Y. Research on weld defect detection and evaluation technology based on deep learning. *Front. Eng. Sci. Res.* **1**, 1. <https://doi.org/10.56028/fesr.1.1.1.2023> (2023).
- Riaz, W., Qi, X., Ji, J. & Ullah, A. Multi-scale feature convolutional modeling for industrial weld defects detection in battery manufacturing. *Fractal Fract.* **9**, 611. <https://doi.org/10.3390/fractalfract9090611> (2025).
- Yadav V, Banerjee D, Sharma N, Singh V (2024) Automated welding defect recognition through deep learning fusion: CNN and SVM integration. In: 2023 4th International Conference on Intelligent Technologies (CONIT). IEEE, pp 1–6
- Yang, L. & Jiang, H. Weld defect classification in radiographic images using unified deep neural network with multi-level features. *J. Intell. Manuf.* **32**, 459–469. <https://doi.org/10.1007/s10845-020-01581-2> (2021).
- Liu X, Wang C, Hu Y, et al (2018) Transfer learning with convolutional neural network for early gastric cancer classification on magnifying narrow-band imaging images. In: 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, pp. 1388–1392
- Ajmi, C. et al. Using deep learning for defect classification on a small weld X-ray image dataset. *J. Nondestruct. Eval.* **39**, 68. <https://doi.org/10.1007/s10921-020-00719-9> (2020).
- Malarvel, M. & Singh, H. An autonomous technique for weld defects detection and classification using multi-class support vector machine in X-radiography image. *Optik* **231**, 166342. <https://doi.org/10.1016/j.ijleo.2021.166342> (2021).
- Hou, W. et al. Automatic detection of welding defects using deep neural network. *J. Phys. Conf. Ser.* **933**, 012006. <https://doi.org/10.1088/1742-6596/933/1/012006> (2018).
- Guo J, Dong L, Xu H (2024) Welding image defect detection based on Yolov5. In: 2024 4th International Conference on Neural Networks, Information and Communication (NNICE). IEEE, pp 469–472

22. Tang, W. et al. Research on welding defect classification methods based on convolutional neural networks. *Int. J. Comput. Sci. Inform. Technol.* **4**, 150–161. <https://doi.org/10.62051/ijcsit.v4n1.19> (2024).
23. Perri, S., Spagnolo, F., Frustaci, F. & Corsonello, P. Welding defects classification through a convolutional neural network. *Manuf. Lett.* **35**, 29–32. <https://doi.org/10.1016/j.mfglet.2022.11.006> (2023).
24. Totino, B., Spagnolo, F. & Perri, S. RIAWELC: A novel dataset of radiographic images for automatic weld defects classification. *Int. J. Electric. Comput. Eng. Res.* **3**, 13–17. <https://doi.org/10.53375/ijecer.2023.320> (2023).
25. Mery, D. et al. GDxray: The database of X-ray images for nondestructive testing. *J. Nondestr. Eval.* **34**, 42. <https://doi.org/10.1007/s10921-015-0315-7> (2015).

Author contributions

K.P.: Conceptualization, Formal analysis, Methodology, Validation, Visualization, Writing—original draft, Writing—review & editing. R.J.: Conceptualization, Formal analysis, Methodology, Validation, Visualization, Writing—original draft, Writing—review & editing. P.A.: Methodology, Validation, Visualization, Writing—original draft, Writing—review & editing. A.J.: Methodology, Validation, Visualization, Writing—original draft, Writing—review & editing. R.P.: Methodology, Validation, Visualization, Writing—original draft, Writing—review & editing. K.U.: Methodology, Validation, Visualization, Writing—original draft, Writing—review & editing. T.A.: Conceptualization, Formal analysis, Methodology, Validation, Visualization, Writing—original draft, Writing—review & editing.

Funding

There is no funding assistance offered for the publication of this paper.

Declarations

Competing interests

The authors declare no competing interests.

Consent to publications

All authors are exhibiting their consent to publish this paper.

Ethics approval

No Participation of human occurs in this implementation process.

Additional information

Correspondence and requests for materials should be addressed to T.B.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026