Grenze International Journal of Engineering and Technology, June Issue

Comprehensive Visualization of Model Predictions using XAI Techniques for Metal Surface Defects in Grayscale Images

Sharanya R Shetty[1] and Guruprasad[2]
[1]Student, Nitte (Deemed to be University) NMAM Institute of Technology (NMAMIT) Nitte Karkala, India
Email: shettysharanya2001@gmail.com
[2]Assistant Professor, Nitte (Deemed to be University) NMAM Institute of Technology (NMAMIT) Nitte Karkala, India
Email: pguru88@gmail.com

***Abstract*— Explainable Artificial Intelligence (XAI) techniques help in visually representing defects during industrial inspection to show model's transparency and reliability. The paper introduces a metal surface defect classification system utilizing a ResNet-50 backbone fused with four Explainable AI (XAI) methods to enhance interpretability level in industrial inspection. A transfer-learning approach with ImageNet-pretrained ResNet-50 is first kept frozen and then selectively fine-tuned to get the deeper layers engaged with defect-specific features. Efficient defect recognition is made possible by a Global Average Pooling, dropout layers, and dense units of a lightweight classification head. The model achieves 93.4% validation accuracy on the NEU-DET dataset for scratches and inclusion. To understand model's predictions XAI techniques like Grad-CAM, LRP, Occlusion Sensitivity, and LIME are used to identify visual explanation of defect-relevant regions. A qualitative comparison helps to understand region of defect on surface metals defects, also identifies major drawback like noise. Methods used point to non-defect areas such as printed text and logos and consider them as most significant regions. This provides valuable insights into robustness and reliability of XAI techniques in real industrial environment.**

***Index Terms*— Explainable Artificial Intelligence (XAI), Gradient-weighted Class Activation Mapping (Grad-CAM), Layer-wise Relevance Propagation (LRP), Local Interpretable Model-Agnostic Explanations (LIME), Occlusion Sensitivity, ResNet-50 and Transfer Learning.**

I. INTRODUCTION

In recent years, computer vision models based on deep learning of Artificial Intelligence (AI) such as CNNs and Vision Transformers (ViT) have in many cases been able to reach high accuracy in the field of industrial quality inspection [1]. These models operate as black boxes that do not offer any kind of explanation and they do not reveal anything about their rationale or decision, making processes [2]. In industrial scenarios such as surface defects on metals detection reliability, safety and responsibility are important. Explainable Artificial Intelligence (XAI) is way for models to be open, where AI models become understandable by highlighting most contributing image region or features to decision and prediction. XAI deliver visual maps, relevance scores and concept-based explanations, which enable user to see if model is really identifying defect areas or if it is just recognizing textures or logos as noise.

Grenze ID: 01.GIJET.12.2.87_23

Metal surfaces usually have small and textured defects, which becomes difficult to detect them. In grayscale pictures, small changes that may be scratches, inclusions or oily stains are usually hardly distinguishable. Explainable Artificial Intelligence (XAI) techniques are assistance by providing visualization and thus locating defect in detection models [3]. The technicians can verify model's highlighted region and by checking if it really focuses on defect area, which in turn leads to increase in trust and accuracy of automated systems. XAI provides manufacturing process analysis by exposing features that have impact on defect classification and same time it assists in model debugging and performance optimization, increasing overall reliability and transparency in industrial material quality inspection.

Explainable Artificial Intelligence (XAI) is a way to bring together AI and humans by focusing on making AI more understandable and hence its predictions transparent [4]. Some of XAI techniques like Gradient weighted Class Activation Mapping (Grad-CAM), Layer wise Relevance Propagation (LRP), Local Interpretable Model agnostic Explanation (LIME) and Occlusion Sensitivity Analysis provide not only visual but also quantitative evidence of which parts of the image were considered by the model. In domain of metal surface inspection, NEU Surface Defect Database that contains six common defect types and GC10 DET dataset with ten industrial defect categories have been used as standard benchmarks for development and evaluation of explain able defect detection models.

Combining XAI techniques with these datasets not only helps human understand model predictions but also make it possible for automated inspection systems to be reliable, trustworthy and auditable, thereby making them suitable for industrial applications.

II. Related Works

Existing literature presents a wide range of defect-detection approaches that increasingly integrate XAI for interpretability and reliability.

Bento et al. [5] presents a novel method of using XAI for model interpretation and to enhance deep learning performance. The authors trained a MobileNet model on a metal surface defect dataset containing images with unwanted overlays such as text and logos, which caused model to learn irrelevant features. By using XAI technique Layer-wise Relevance Propagation (LRP), they have visualized which region of image is influenced for predictions and observed that model focused on overlays like text and logos instead of actual defects. Using this they applied preprocessing methods like Gaussian blur, image cropping, censor bars and GAN-based generative inpainting to remove overlays and retrained model. The improved dataset led to 20% increase in F1-score.

Kotsiopoulos et al. [6] integrates deep, learning defect detection, post-hoc explainability and human-expert feedback loops to improve both performance and trust. On dataset of images from hard metal production and multi, sensor data, system applies a convolutional neural network to figure out if metal sample is defective or not. Later, it applies explainable AI techniques such as SHAP to give reasons behind predictions to human operators. AI model in this paper had performance score of over 90%. Human reviewers are able to inspect, correct and justify results. Received feedback is applied in model refinement.

Aboulhosn et al. [7] presents deep CNN as main architecture for feature extraction and classification in steel surface defect detection is focus of paper to guide high-level hierarchical features and also reduce training time and amount of data required, transfer learning with pretrained backbones ResNet, VGG, or EfficientNet used. To create texture irregularities and local defect patterns of steel images, authors use a series of convolution batch normalization and max-pooling layers. System employs Explainable AI techniques Grad-CAM to make model more comprehendible and produces class, discriminative heatmaps to visualize most important defect areas recognized by CNN and interpretability methods such as LRP or SHAP are used to provide pixel-level relevance mapping and feature attribution. These methods form robust, interpretable and high-performance pipeline for automated industrial defect detection.

Almodawar et al. [8] details a deep CNN architecture like ResNet50 or VGG19 trained on texture datasets to understand multi-scale spatial representations which are vital for texture prediction. Proposed Multi-Layer Grad-CAM (MLGC) method goes beyond single activation layer by retrieving gradient-weighted class activation maps from several intermediate convolutional layers. Fusion method uses layer-wise attention weighting that allows model to pinpoint texture-discriminative regions more precisely than standard Grad-CAM. MLGC module is integrated post-hoc method which is model-agnostic and can be used with different CNN backbones. Fine-tuned MobileNetV2 model with accuracy of 98.3%. Multi-layer explanation method increases level of model's transparency and at same time it raises transparency of model in texture classification tasks.

Yue et al. [9] presents Metal-YOLOX builds on anchor-free YOLOX architecture while incorporating specialized modules to enhance detection irregular metal surface defects under complex textures and low-contrast conditions. To locate tiniest defects, model upgrades backbone based on CSPDarknet by strengthening PANet feature pyramid and improving multi-scale feature extraction of network. By introducing channel attention and dynamic convolution, Metal Feature Enhancement Module (MFEM) is designed to excite discriminative texture cues, thus ensuring fine defect patterns present in input are represented robustly. Metal-YOLOX has 79%, 69%, and 81% on NEU DET, GC-10 and Aluminium datasets accuracy respectively. Paper uses XAI methods Grad-CAM and feature activation visualization, that are combined with intermediate convolutional layers to get regions which have most influence on decisions.

Ullmann et al. [10] idea is to use layer deep CNN or transformer-based network trained using millimetre-wave imaging data, in order to spatial-frequency patterns concerning both surface and subsurface defects. Author uses multi-scale convolutional encoders featured in model extract detailed structural signatures from mmWave reflections. Explainable AI methods Grad-CAM, Integrated Gradients and Layer-wise Relevance Propagation are used by system. These XAI maps provides tool to check if network is concentrating on actual defect regions instead of noise, reflections or imaging artifacts.

Patel et al. [11] introduce a Federated Learning setup where various semiconductor fabs collectively train a CNN-based defect detection model. They do this by sharing gradients instead of raw wafer images. FedAvg aggregation is performed by a central server that maintains privacy and high classification accuracy across heterogeneous fab data. To represent wafer defects at detailed level, lightweight CNNs are used with test accuracy of 98%. Grad-CAM, SHAP and Integrated Gradients are explainable AI methods employed.

Demircioglu et al. [12] provides extract texture from metal surface images by applying Gabor Filter Bank comprising multiple orientations and scales. Handcrafted feature vector is successively reduced by PCA or LDA and then classified by SVM, Random Forest and shallow neural networks. XAI methods like SHAP, LIME, and Gabor-activation visualization identify most significant filter responses, thereby revealing defect related texture patterns. Method is further improved by normalization and cross validation to obtain reliable metal texture classification.

Qian et al. [13] suggest a Mixup-enhanced training strategy for performing defect cause analysis at strip steel manufacturing stage by interpretable linear classifiers Linear Regression, Logistic Regression and Linear SVM are used. SHAP and LIME are applied to identify most influential process parameters and improve feature interpretability and Hanchate et al. [14] carry out vibration signal analysis through FFT, STFT and wavelet packet transforms followed by defect prediction using Random Forest, Gradient Boosting or simple neural networks. SHAP and LIME facilitate explainable analyses by pinpointing main features or frequency bands enabling development of noise, robust and real time XAI driven industrial quality monitoring systems.

Song et al. [15] propose a method for localizing defects that focusses on multi-constraint saliency and combines global contrast, texture deviation and structural irregularity through Gabor filters, LBP and gradient, based operators to produce XAI, like saliency maps and Pang et al. [16] introduce a MAML based meta, learning technique allows fast adaptation to new classes of industrial defects even with very few labelled samples, outperforming standard supervised models in few-shot settings.

TABLE I: RESEARCH GAP FOR RELATED WORKS

Paper	Dataset	Methodology	Research Gaps
Bento et al. [5] (2021)	GC10-DET [18]	Used LRP to identify model bias caused by text/logos	Did not compare multiple XAI methods and study of behaviour across different techniques.
Kotsiopoulos et al. [6] (2024)	BAUMER and HIKVISION camera	DL + XAI SHAP + human feedback for defect classification	Uses only basic XAI techniques and no evaluation of XAI sensitivity to elements like text/logos.
Aboulhosn et al. [7] (2024)	Severstal [20]	CNN-based steel defect detection with Grad-CAM	Only uses one or two XAI techniques and no robustness evaluation under noise/artifacts.
Almodawar et al. [8] (2025)	NEU-DET [19]	Multi-Layer Grad-CAM for texture classification	Focus on Grad-CAM variations only and no study of false positives.
Yue et al. [9] (2023)	NEU-DET GC-10	Metal-YOLOX with Grad-CAM for interpretability	XAI limited to Grad-CAM and no test on images with logos/text.
Ullmann et al. [10] (2025)	mmWave Image Dataset	CNN/Transformer with Grad-CAM, IG, LRP for mmWave defects	No qualitative heatmap comparison across industrial image noise and no study of non-defect region activation.
Patel et al. [11] (2024)	MixedWM38 dataset.	Federated defect detection with Grad-CAM, SHAP, IG	No visual comparison across multiple XAI maps and no qualitative-only interpretability analysis.
Song et al. [15] (2020)	SD-saliency-900 dataset [21]	Saliency-based defect localization	No comparison with modern XAI techniques and no analysis under real-world artifacts.

Table I shows explainable defect detection with XAI methods has been extensively studied. Bulk of such research has been limited to one or two methods (usually Grad-CAM or Integrated Gradients) and has not dealt with a comprehensive qualitative comparison of multiple interpretability techniques. Besides that, examination of models in prior works is done using clean and controlled datasets, so there is no consideration of XAI performance in presence of real-world non-defect factors such as logos, printed text or labels in image. Consequently, problem of false-positive visual explanations that is of great importance has surfaced and still remains without a solution.

The issue that has been largely ignored is visual evaluation of most realistic and practical XAI methods for numerous industrial applications. This presents qualitative evaluation framework aimed at comparing four key XAI techniques Grad-CAM, LRP, Occlusion Sensitivity, and LIME on identical dataset without relying on any saliency ground-truth. Paper illustrates the behaviour of XAI methods towards the images with non-defect elements, and therefore, it reveals that usually they interpret logos or text as defects.

III. PROPOSED METHOD

This work proposes an end-to-end explainable deep learning framework for multi-class image classification using a fine-tuned ResNet-50 backbone integrated with multiple post-hoc XAI techniques. The Fig. 1 Shows that the proposed method includes (i) structured dataset preparation and preprocessing, (ii) model architecture with fine-tuned ResNet50 model (iii) Training procedure with two-stage model training (transfer learning + fine-tuning) and performance evaluation using confusion matrices and classification reports and (iv) generation of comprehensive visual explanations using Grad-CAM, LRP, LIME and Occlusion Sensitivity. The complete workflow is implemented in Google Colab [17] and is fully reproducible.

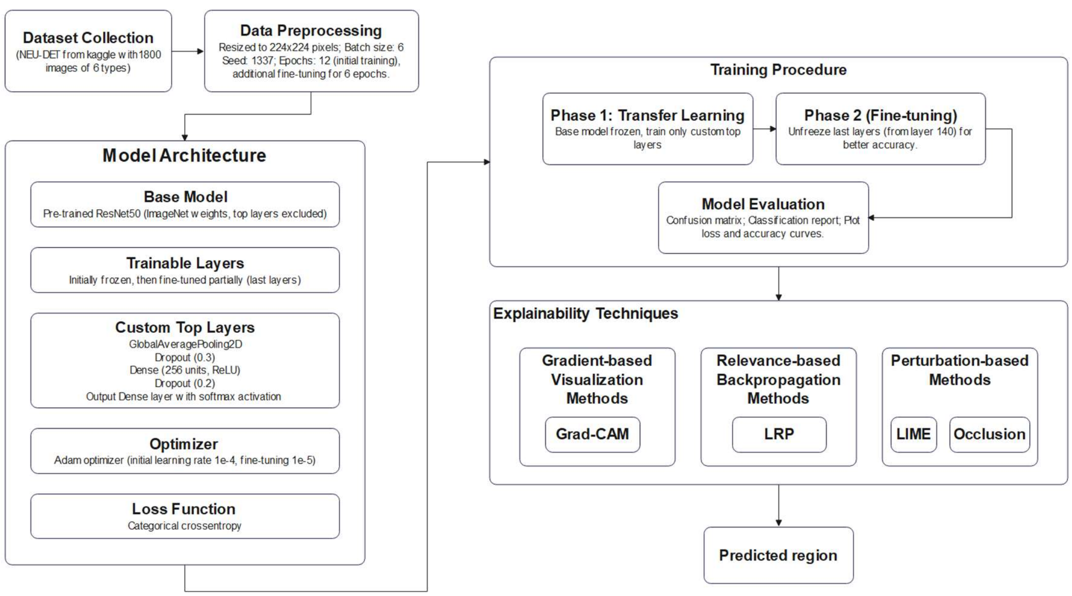


Figure 1. Flow diagram of Proposed Method for ResNet50 model and XAI Techniques

A. Dataset Preparation

The dataset is organized into TRAIN and VALIDATION directories, each containing class-specific subfolders. Images are automatically loaded using TensorFlow image_dataset_from_directory() API with categorical labels. To ensure consistency and reduce data leakage, framework verifies class distribution in both sets. All images are resized to 224 × 224 pixels to match with ResNet-50 input format. Data augmentations such as random flips, rotations, zoom, and contrast changes are used.

B. Model Architecture

ResNet-50 pretrained on ImageNet is used as convolutional back bone with a transfer-learning based architecture. The base network is loaded with frozen weights during first phase of training.

Feature extractor which is on top layer, lightweight classification head is added:

- Global Average Pooling
- Dropout (0.3)
- Dense layer with 256 ReLU units
- Dropout (0.2)
- Final SoftMax layer with N outputs

This hybrid architecture allows efficient learning of domain-specific features while benefiting from pretrained hierarchical representations.

C. Training Procedure

Phase-1: Frozen Backbone Training: First stage all convolutional layers of ResNet-50 remain frozen and only classification head is trained using Adam optimizer (learning rate 1e-4) with categorical cross-entropy loss. This step rapidly adapts top layers to target domain.

Phase-2: Selective Fine-Tuning: In second stage, deeper layers from index 140 onward are unfrozen. This selective unfreezing preserves low-level feature extraction while allowing high-level task-specific refinement. Fine-tuning is performed with a smaller learning rate (1e-5) to prevent catastrophic forgetting.

Model Evaluation: It evaluates confusion matrix, classification report like precision, recall and F1-score with plot of training and validation loss and accuracy curves. The final model is saved in model.keras format after training completion.

D. Explainable AI Techniques Used

Grad-CAM creates class, specific heatmaps by calculating gradients of predicted class score relative to final convolutional layer (conv5_block3_out of ResNet-50), thus a coarse saliency map that visually represents model decision influencing areas is produced. LRP propagates relevance scores back through network using gradient input approximation that eventually results in pixel-level relevance maps. Occlusion Sensitivity determines the regions of image that are more important for the prediction of model by analysing how prediction score decreases when local regions of image are hidden. LIME highlights image regions that contribute most to the predicted class using super pixel perturbations and a local linear surrogate model.

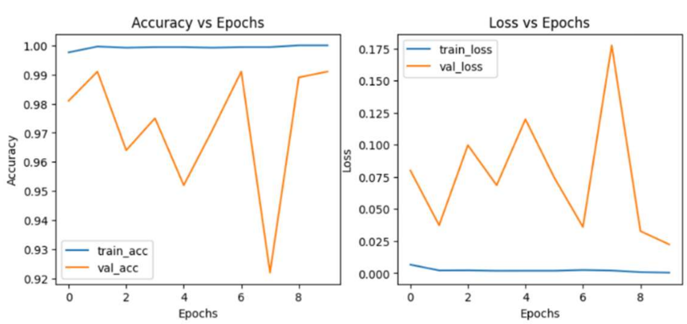
IV. RESULT AND DISCUSSION

This chapter presents the experimental results of proposed metal surface defect classification system based on ResNet50 deep learning model and Explainable AI (XAI) framework. The evaluation includes model performance metrics, training behaviour analysis, and comparative visual interpretation using multiple XAI techniques such as Grad-CAM, LRP, Occlusion Sensitivity, and LIME.

A. Model Performance

Fig. 2 shows ResNet50 model provides strong classification on NEU-DET validation dataset and obtains validation accuracy of 93.4%

Confusion Matrix as shown in Fig. 3 for total of 5000 images consisting of inclusion and scratches plot of confusion matrix is created. Each type consists of 2500 images. 2000 for training and 500 for validation of each class. For class inclusion validation out of 500, 434 is correctly predicted as class inclusion and 66 are detected as class scratches. For class scratches validation out of 500, 500 are correctly predicted as class scratches and 0 are detected as class inclusion.



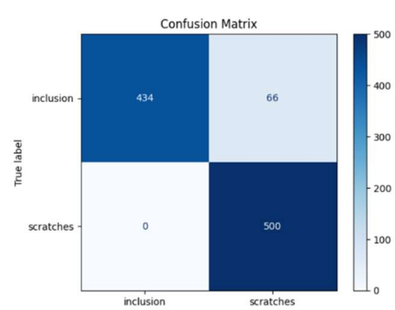


Figure 2. Accuracy and Loss of Validation for ResNet50 model. Figure 3. Confusion Matrix for ResNet50 Model

B. XAI Visualization

The XAI pipeline produced interpretability maps for Grad-CAM, LRP, Occlusion Sensitivity, and LIME. In Fig. 4, Grad-CAM generated high-intensity activation around defect region, confirming network's ability to focus on the spatially relevant area. The LRP map provided a high-resolution relevance visualization. The method physically isolated most relevant pixel areas. These areas highlight edges and structural aspects of defect that were irregular. When main defect regions were covered, occlusion map exhibited largest drops in system confidence thus it was a way of regions discriminative importance being validated. By localizing significant segments, LIME singled out super pixels which were highly correlated with defect. Highlighted super pixels were parts of defective texture which confirmed classifier's dependence on defect-specific regions rather than background noise.

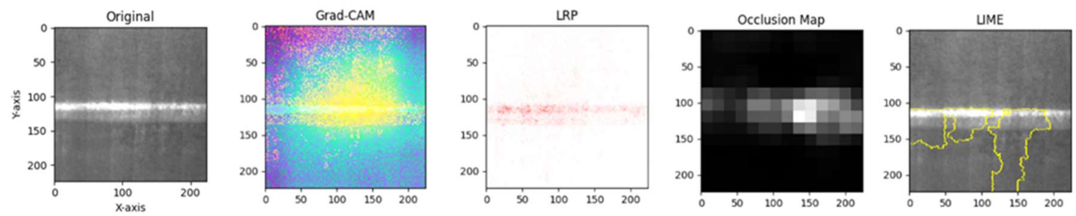


Figure 4. Scratches image for Original, GradCAM, LRP, Occlusion Map and LIME visualization

In Fig. 5, Original image sample featured additional non-defect elements, namely alphanumeric text (upper-left corner) and a logo (lower-right corner). All applied XAI methods consistently indicated these areas as most relevant ones for model's prediction. The response of model was such because newly added structures contained high-contrast edges and localized patterns that are very different from uniform metal surface texture. Consequently, network wrongly identified these artifacts as most discriminative features, which made Grad-CAM, LRP, Integrated Gradients, Occlusion Sensitivity, and LIME to locate these areas as ones to which model was most sensitive. Such a result highlights that a classifier is picking up on spurious features and thus, is highly likely to misinterpret foreign markings as defect features. XAI outputs are very helpful in identifying such bias, but impose requirement of having clean, standardized and annotation free input images during inference phase.

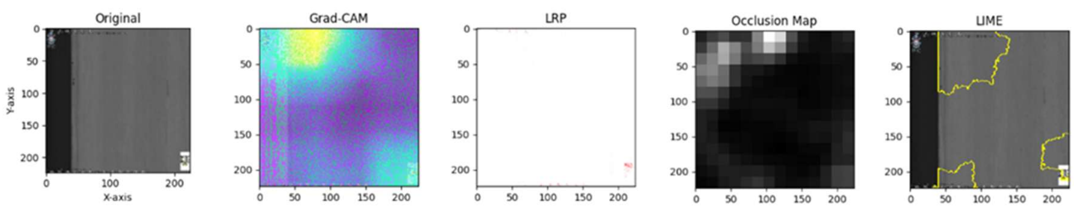


Figure 5. Original image with noise as input and output as Grad-CAM, LRP, Occlusion Map and LIME visualization

If input image sample has extra non-defect elements like an alphanumeric text (upper-left corner) and logo (lower-right corner), XAI techniques that were used in all instances unanimously indicated these areas as most relevant for model's prediction.

IV. CONCLUSION

This paper uses four Explainable AI (XAI) methods such as Grad-CAM, LRP, Occlusion Sensitivity and LIME to visually represent defect detection on metal surfaces area. Results demonstrate that these methods are pointing to actual defect spots, but at same time, they also locate non-defect areas, e.g., printed texts and logos, because these have been having sharp edges and a more regular structure similar to defect. This reveals a substantial limitation of model. Model gets confused and thinks that high contrast artifacts are defects, which exposes its biases and cases in which it fails. Instead of relying on quantitative metrics that need pixel level ground truth, work relies on qualitative visualization as more visual representable interpretability method for real industrial settings.

Further research work should mainly focus on ways enhancing model robustness can be achieved by removing or masking irrelevant artifacts detected during pre-processing and by augmenting datasets with synthetic text and logo variations that model fails to recognize. Attention mechanism implementation, region masking methods together with domain knowledge about metal surface texture can help bring the number of false positive explanations down to almost zero. By employing quantitative interpretability metrics and advanced or noise robust XAI methods, trust component will be even stronger. Greater diversity and number of real-world samples in datasets and testing models in live production scenarios will not only benefit generalization side but also speed up creation of fully operational XAI based defect inspection systems.

REFERENCES

[1] Chefer, H., Gur, S., & Wolf, L. (2021). "Transformer interpretability beyond attention visualization". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 782-791).

[2] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.

[3] Oliveira, R. M. A., Sant'Anna, Â. M. O., & da Silva, P. H. F. (2024). Explainable machine learning models for defects detection in industrial processes. Computers & Industrial Engineering, 192, 110214.

[4] Moosavi, S., Farajzadeh-Zanjani, M., Razavi-Far, R., Palade, V., & Saif, M. (2024). Explainable AI in manufacturing and industrial cyber–physical systems: A survey. Electronics, 13(17), 3497.

[5] V. Bento, M. Kohler, P. Diaz, L. Mendoza, and M. A. Pacheco, "Improving deep learning performance by using explainable artificial intelligence (XAI) approaches," Discover Artificial Intelligence, vol. 1, no. 1, p. 9, 2021.

[6] T. Kotsiopoulos, G. Papakostas, T. Vafeiadis, V. Dimitriadis, A. Nizamis, A. Bolzoni, D. Bellinati, D. Ioannidis, K. Votis, D. Tzovaras et al., "Revolutionizing defect recognition in hard metal industry through AI explainability, human-in-the-loop approaches and cognitive mechanisms," Expert Systems with Applications, vol. 255, p. 124839, 2024.

[7] Z. Aboulhosn, A. Musamih, K. Salah, R. Jayaraman, M. Omar, and Z. Aung, "Detection of manufacturing defects in steel using deep learning with explainable artificial intelligence," IEEE Access, 2024.

[8] A. Almodawar, A. Ahmad, G. Hristov, and Y. Tashtoush, "Multi-layer grad-cam (MLGC) Explainable AI model for texture prediction," 09 2025, pp. 175–181.

[9] X. Yue, J. Chen, and G. Zhong, "Metal surface defect detection based on Metal-YOLOX," International Journal of Network Dynamics and Intelligence, pp. 100020–100020, 2023.

[10] I. Ullmann and J. Sch¨ur, "Explainable AI for reliable defect recognition and localization in non-destructive testing with millimeter-wave imaging," IEEE Access, 2025.

[11] T. Patel, R. Murugan, G. Yenduri, R. H. Jhaveri, H. Snoussi, and T. Gaber, "Demystifying defects: Federated learning and Explainable AI for semiconductor fault detection," IEEE Access, vol. 12, pp. 116987 117007, 2024.

[12] P. Demircioglu, A. Seckin, J. Torgersen, I. Bogrekci, and N. Durakbasa, "Metal surface texture classification with gabor filter banks and XAI," pp. 033–050, 2023.

[13] J. Qian, X. Zhang, S. Pi, Z. Song, and X. Zhang, "Mix up enhanced linear robust explainable model for identifying key factors of surface defects in strip steel manufacturing," in 2024 6th International Conference on Industrial Artificial Intelligence (IAI). IEEE, 2024, pp. 1–6.

[14] A. Hanchate, S. T. Bukkapatnam, K. H. Lee, A. Srivastava, and S. Kumara, "Explainable AI (XAI)-driven vibration sensing scheme for surface quality monitoring in a smart surface grinding process," Journal of Manufacturing Processes, vol. 99, pp. 184–194, 2023.

[15] G. Song, K. Song, and Y. Yan, "Saliency detection for strip steel surface defects using multiple constraints and improved texture features," Optics and Lasers in Engineering, vol. 128, p. 106000, 2020.

[16] S. Pang, L. Zhang, Y. Yuan, W. Zhao, S. Wang, and S. Wang, "Adaptive MAML: Few-shot metal surface defects diagnosis based on model agnostic meta-learning," Measurement, vol. 223, p. 113612, 2023.

[17] Y. Name, "Metal surface defect classification with XAI– google Colab notebook," Google Colab Notebook, 2025, accessed: October 2025. [Online]. Available: https://colab.research.google.com

[18] A. Kim, "GC10-DET: Surface defect dataset," Kaggle Dataset, 2020, accessed: 04 December 2025. [Online]. Available: https://www.kaggle.com/datasets/alex000kim/gc10det

[19] K. Dikshit, "Neu surface defect database," Kaggle Dataset, 2020, accessed: 04 December 2025. [Online]. Available: https://www.kaggle.com/datasets/kaustubhdikshit/neu-surface-defect-database

[20] Kaggle, "Severstal Steel Defect Detection," Kaggle Competition Dataset, 2019, accessed: 12 December 2025. [Online]. Available: https://www.kaggle.com/c/severstal-steel-defect-detection

[21] A. Kim, "SDSaliency900," Kaggle Dataset, 2020, pixel-wise binary maps (ground truth) accessed: 12 December 2025. [Online]. Available: https://www.kaggle.com/datasets/alex000kim/sdsaliency900