



# Visual coating inspection framework via self-labeling and multi-stage deep learning strategies

Changheon Han<sup>1</sup> · Jiho Lee<sup>1</sup> · Martin B. G. Jun<sup>1,2</sup> · Sang Won Lee<sup>3</sup> · Huitaek Yun<sup>4</sup>

Received: 12 September 2023 / Accepted: 13 March 2024  
© The Author(s) 2024

## Abstract

An instantaneous and precise coating inspection method is imperative to mitigate the risk of flaws, defects, and discrepancies on coated surfaces. While many studies have demonstrated the effectiveness of automated visual inspection (AVI) approaches enhanced by computer vision and deep learning, critical challenges exist for practical applications in the manufacturing domain. Computer vision has proven to be inflexible, demanding sophisticated algorithms for diverse feature extraction. In deep learning, supervised approaches are constrained by the need for annotated datasets, whereas unsupervised methods often result in lower performance. Addressing these challenges, this paper proposes a novel deep learning-based automated visual inspection (AVI) framework designed to minimize the necessity for extensive feature engineering, programming, and manual data annotation in classifying fuel injection nozzles and discerning their coating interfaces from scratch. This proposed framework comprises six integral components: It begins by distinguishing between coated and uncoated nozzles through gray level co-occurrence matrix (GLCM)-based texture analysis and autoencoder (AE)-based classification. This is followed by cropping surface images from uncoated nozzles, and then building an AE model to estimate the coating interface locations on coated nozzles. The next step involves generating autonomously annotated datasets derived from these estimated coating interface locations. Subsequently, a convolutional neural network (CNN)-based detection model is trained to accurately localize the coating interface locations. The final component focuses on enhancing model performance and trustworthiness. This framework demonstrated over 95% accuracy in pinpointing the coating interfaces within the error range of  $\pm 6$  pixels and processed at a rate of 7.18 images per second. Additionally, explainable artificial intelligence (XAI) techniques such as t-distributed stochastic neighbor embedding (t-SNE) and the integrated gradient substantiated the reliability of the models.

**Keywords** Automated visual inspection · Autonomous data labeling · Explainable artificial intelligence · Coating boundary pinpointing

## Introduction

Product inspection constitutes a vital phase in manufacturing to ensure the quality of products. Manufacturers endeavor to execute meticulous inspections, resulting in the inspection process frequently constituting the most substantial portion of production expenses (Chin & Harlow, 1982). Visual inspection, which scrutinizes functional and cosmetic imperfections, is known as a versatile, simple, cost-efficient, and contactless approach (Babic et al., 2021; Chin & Harlow, 1982). However, it primarily relies on human inspectors' capabilities (Harris, 1969), and the performance of visual inspection is susceptible to human factors that may be influenced by elements such as the inspector's expertise, the intricacy of the inspection task, product defect rates, and

---

✉ Huitaek Yun  
htyun@kaist.ac.kr

<sup>1</sup> School of Mechanical Engineering, Purdue University, West Lafayette, IN 47906, USA

<sup>2</sup> Indiana Manufacturing Competitiveness Center (IN-MaC), Purdue University, West Lafayette, IN 47906, USA

<sup>3</sup> School of Mechanical Engineering, Sungkyunkwan University, Suwon, Gyeonggi-do 16419, Republic of Korea

<sup>4</sup> Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea

repeatability (Chin & Harlow, 1982; Harris, 1969; Megaw, 1979). In other words, human factors might cause unreliable and costly outcomes despite the merits of visual inspection.

It is particularly evident that the issues are connected to production loss in the field of coating inspection, where both precision and speed are important due to the heightened vulnerability of coated surfaces to flaws, defects, or discrepancies (Doering et al., 2004). Rapid and accurate inspection is paramount, alongside the need for mitigating reliance on human factors to avoid those risks. As one of the potential solutions, machine vision reduced human error through the automation of data acquisition, analysis, and image assessment (Alonso et al., 2019). Nonetheless, its application is often restricted to specific situations, with limited scope for adaptability (Psarommatis et al., 2019), necessitating extensive feature engineering and sophisticated algorithms to function effectively across diverse scenarios.

Deep learning has enhanced visual inspection by automatically abstracting features from data even in intricate and high-dimensional scenarios (LeCun et al., 2015, p. 2; Ren et al., 2022; Rusk, 2016). Deep learning can be classified into supervised and unsupervised approaches, each with inherent challenges. Supervised learning necessitates extensive manually annotated datasets for model training, which can be labor-intensive and impede the efficiency of deep learning applications (Shinde et al., 2022; Wang & Shang, 2014). On the other hand, unsupervised learning, while not requiring annotated data, relies on manually set thresholds, often resulting in unacceptable performance in product inspection tasks (Chow et al., 2020; Kozamernik & Bračun, 2020). Addressing the challenge of minimizing human intervention in data labeling is thus important to enhance the efficacy of deep learning-based inspection methodologies.

This study focuses on advancing the capabilities of automated visual inspection (AVI) by developing a novel framework that autonomously annotates and inspects products with high accuracy. Given the cost-effectiveness and high precision the manufacturing sector requires, this research targeted pinpointing coating interfaces on fuel injection nozzles, which demands pixel-level accuracy. The novelty of this research lies in the application of self-supervised learning and autonomous data annotation algorithms, eliminating the need for human labeling and precisely locating the coating interfaces. Thus, a robust AVI framework is capable of pinpointing coating interfaces from scratch via autonomous data annotation, which is expected to improve the efficiency and reliability of the inspection process in real-world practical applications.

Moreover, most deep learning applications do not provide an explanation of their autonomous decision-making processes, and it hinders understanding the process and exacerbates the reliability of models (Gunning et al., 2019). Explainable artificial intelligence (XAI) offers transparency

in decision-making and aids the robustness and applicability of deep learning models (Cooper et al., 2023; Lee et al., 2022; Liu et al., 2022; Wang et al., 2019, 2023). Consequently, this study investigated the interpretability and explainability to validate the reliability of the deep learning models based on t-distributed stochastic neighbor embedding (t-SNE) and integrated gradient techniques.

The major contributions are outlined as follows:

- The autonomous deep learning-based coating inspection framework is proposed. It identifies coated nozzles, generates self-labeled datasets, trains models, and pinpoints coating interfaces without human intervention.
- The interpretation and explanation of the autoencoder (AE) and convolutional neural network (CNN)-based detection models are provided based on the concept of XAI to validate the reliability of the developed models.

This paper consists of the following. “**Related work**” Section describes an overview of the related literature in machine vision, computer vision, deep learning, and visual inspection applications. “**Methodology**” Section explains the proposed framework. Then, the detailed analysis and discussion follow in “**Results and discussions**” Section. Lastly, “**Conclusion**” Section describes the concluding remarks of the research.

## Related work

Industry 4.0 catalyzes the application of data-driven approaches in quality management, which plays a critical role in productivity and reliability in the manufacturing domain (Psarommatis et al., 2022). The surge in customer demand for diverse products increases the complexity of products and production systems and necessitates flexible and versatile automated inspection solutions for effective quality control (Jacob et al., 2018; Psarommatis et al., 2019). The emergence of contemporary technologies such as deep learning, the internet of things, sensing, and computer vision initiates a paradigm shift in product inspection techniques (Oztemel & Gursev, 2020).

Katirci et al. (2021) introduced a novel automated inspection technique based on electrical and thermal conductivities. Their approach, however, necessitates immersing an object in a solution or scattering powders over it, potentially causing damage and deformation in some cases. To prevent damage to products, nondestructive and contactless inspection solutions have been investigated. A laser displacement sensor is one of the nondestructive and contactless automated inspection methods and has a high resolution and sampling rate, which fits for precisely measuring a surface thickness in real-time without contact (Gryzagoridis, 2012; Nasimi & Moreu,

2021). However, it is not applicable to identifying diverse textures of fuel injection nozzle interfaces. Additionally, such approaches demand complicated and finely calibrated measurement systems compared to machine vision-based AVI using Charge-Coupled Device (CCD) or Complementary Metal–Oxide–Semiconductor (CMOS) cameras. They are implemented to offer an autonomous pipeline that captures and processes images and detects geometric or texture features for decision-making based on defined criteria in machine vision systems due to their versatility and cost-effectiveness (Golnabi & Asadpour, 2007; Noble, 1995; Ren et al., 2022).

As captured images often encounter issues like noise and uneven contrast and brightness, the Fourier transform (Brigham & Morrow, 1967) and the wavelet transform (Graps, 1995) aid to adjust images, eliminate unnecessary data, and emphasize useful information. The two-dimensional discrete Fourier transform (Gonzalez & Faisal, 2019) is widely utilized for image processing, and the wavelet transform is a useful tool to denoise and find a sparse representation in an image (Khatami et al., 2017). The Stationary Wavelet Transform (SWT) is known for offering an approximation to investigate singularities at the interface and maintain the consistency of image size (Nason & Silverman, 1995; Wang et al., 2010). Despite the benefits of denoising and highlighting features (Bai & Feng, 2007), Ren et al. (2022) stated that transform techniques require high computing costs. It also introduced detailed wavelet transform research including denoising (Jain & Tyagi, 2015; Luisier et al., 2007), image fusion (Daniel, 2018; Xu et al., 2016), and image enhancement (Jung et al., 2017; Yang et al., 2010) and categorized visual inspection approaches into classification, localization, and segmentation problems.

Traditionally, visual inspection methods such as the histogram technique, histogram of oriented gradients, scale-invariant feature transform, and speeded-up robust features have been used (Ren et al., 2022). However, there are drawbacks like computational load, high requirements for image quality, and irrelevance of spatial features. The support vector machine (SVM) is a prevalent machine learning approach and proves the advantage of classification problems. Nonetheless, the SVM is not suitable for complex multi-classification problems and requires manual adjustment for hyperparameters. The k-means clustering algorithm is an unsupervised machine learning approach and can be applied to multi-classification problems. For instance, Park et al. (2008) demonstrated an AVI framework implementing computer vision and the k-means clustering algorithm to find a defect in cigarette packing by segmenting regions in a package. The k-means clustering algorithm, however, is also improper to use for large and complex datasets (Guan et al., 2003).

Recent advances in deep learning circumvent the shortcomings of traditional approaches. Deep learning techniques

automatically extract features from datasets and aid to develop improved AVI frameworks. CNN is the most popular supervised learning architecture in various fields including the realm of computer vision including self-driving, facial identification, robotic manufacturing, and remote space exploration, which yields substantial advancements (Gu et al., 2018; LeCun et al., 1998; Lee et al., 2022; Li et al., 2022; Park et al., 2023; Yun et al., 2023a, 2023b), and have been broadly employed in AVI applications (Alonso et al., 2019; Park et al., 2016; Singh & Desai, 2022; Wang et al., 2019; Yun et al., 2020).

In terms of the coating inspection, Ficzero et al. (2022) presented an AVI method using RGB images from a CCD/CMOS camera that could be readily applied in the industrial field and showed the feasibility of accurately inspecting tablet coating quality in real-time using machine vision and deep learning techniques. Despite its advantage, their study focused on detection and classification rather than pinpointing defects, requiring manually labeled datasets. As there are not enough references to pinpoint the dimension of a defect within a few pixels of error, this paper aimed to develop a framework that precisely locates the interface of a fuel injection nozzle while minimizing manual data labeling and feature engineering processes with the simple configuration of data acquisition. In short, the shortcoming of supervised learning architectures is the necessity of labeled data (Wang & Shang, 2014) implying that human intervention in labeling is indispensable for assembling training datasets.

AE is an unsupervised learning or semi-supervised learning model comprised of encoder and decoder layers bridged by a latent space (LS) that compresses data and extracts distinct features (Yun, 2023a, 2023b). It enables unsupervised feature learning and assesses data similarity based on a reconstruction error from the loss function and would be a useful tool to improve the performance of deep learning at the initial stage of developing another model (Bengio et al., 2014). Erhan et al. (2010) introduced the concept of unsupervised pre-training to help deep learning performance and Feng et al. (2020) proposed a self-taught learning technique using unlabeled data to enhance the detection performance of target samples. Both studies emphasized that unsupervised feature learning can effectively substitute manual data annotation.

In the visual inspection domain, Chow et al. (2020) implemented a convolutional AE model to detect defects in concrete structures based on an anomaly detection method comparing reconstruction errors of surface images. Their model was trained by defect-free images, which used a non-labeling process. Nevertheless, a threshold should be defined to distinguish defects, and additional analysis was required to get their exact location. Moreover, the recall and precision were from 30.1% to 91.3%, and it has a limitation to be implemented in real-world inspection applications in the manufacturing domain. Kozamernik and Bračun (2020)

introduced a method to detect defects on the surface of an automotive part automatically. The model, however, did not provide the position of the defects in a few pixels of the error range and the specific value of accuracy was not indicated. Studies on pinpointing a particular surface within a few pixels of the error based on the difference in texture for a practical industrial application have still been limited.

In short, although deep learning resolves many issues in the visual inspection domain, CNN approaches, supervised learning, still require human effort to annotate data, and AE methods, unsupervised learning, also necessitate defining a threshold and do not attain enough accuracy for practical applications. This prompted the development of an advanced AVI framework in the present study, which combines the benefits of automated data labeling via AE's unsupervised feature learning and the precise pinpointing capabilities of CNN's supervised feature learning. Meanwhile, many deep learning models are like a black box and do not illustrate their decision-making processes clearly. It prevents understanding and aggravates the credibility of models. To address this, XAI is getting significant attention for its ability to enhance the reliability of deep learning models. The XAI approach is advantageous for visualizing data structure and interpreting prediction basis to provide intuitive insights into a model's decision-making. Recent studies have leveraged the advantages of XAI to present cases where the explainability of AVI models was substantially improved. Al Hasan et al. (2023) introduced the interpretable and explainable AVI system to detect hardware Trojans and defects in semiconductor chips. Gunraj et al. (2023), developed the SolderNet, an explainable deep learning system aimed at improving the inspection of solder joints in electronics manufacturing, by providing a more transparent approach.

As the prominent method in XAI for data visualization, t-SNE (Maaten & Hinton, 2008) is regarded as an unsupervised algorithm that visualizes high-dimensional data by reducing its dimensionality to levels, typically two or three dimensions, that can be visually perceptible by humans. The t-SNE is particularly effective in preserving probabilistic similarities among samples when translating high-dimensional data into a lower-dimensional space, thereby effectively revealing complex data structures and patterns. In contrast, principal component analysis (PCA) (Wold et al., 1987), while a standard approach for data reduction and visualization, is limited by its linear transformation approach focused on maximizing data variance, thus struggling with non-linear data relationships. While uniform manifold approximation and projection (UMAP) (McInnes et al., 2020) is efficient and adaptable for large datasets, it occasionally merges distinct clusters, which potentially obscures the finer local structures that are more effectively preserved by t-SNE.

Additionally, integrated gradients, which represent the intensity of the network, are computed by accumulating the

gradients along the path line (Lundstrom et al., 2022). The heatmap technique, a popular visualization tool in the computer vision domain, aids in identifying the most important parts of an input image through neural networks of a deep learning model for classification problems using calculated integrated gradients (Qi et al., 2020; Selvaraju et al., 2020). The deletion metric quantitatively measures the impact of removing these significant parts, which if correct, would hinder the accurate detection by a deep learning model (Petsiuk et al., 2018).

According to the surveyed literature, this study implemented AE and CNN architectures to conjugate their advantages and built a feasible AVI framework for coating inspection based on self-supervised and self-annotation learning with AE and CNN architectures to mitigate the required resources for establishing the models. The explanation of the developed models was investigated to ensure their credibility. The following section describes the detailed methodology.

## Methodology

This study proposes an AVI framework pinpointing coating interfaces with the autonomous data annotation technique to mitigate the necessity of human intervention. As Fig. 1 illustrates, the proposed framework comprises six parts: (1) Classifying nozzle types, (2) Cropping uncoated surface images, (3) Training autoencoder model 1 (AE1), (4) Generating datasets automatically annotated as coated and uncoated surface classes, (5) Training a CNN-based detection model, and (6) Improving performance and validating the trustworthiness of models. First, the framework identifies two uncoated nozzles with the highest GLCM energy values to train autoencoder model 1 (AE1). AE1 abstracts features from uncoated nozzles and classifies mixed nozzle images into coated and uncoated classes. Subsequently, autoencoder model 2 (AE2) is trained with cropped surface images from the classified uncoated nozzles. Coated and uncoated partial surface images were autonomously extracted and annotated according to the region exhibiting the highest reconstruction error on each nozzle. Thirdly, an initial CNN-based detection model is constructed utilizing the automatically collected and labeled images. This model is refined through an iterative training strategy with transfer learning, using datasets generated by a previously trained CNN model to improve accuracy. Lastly, the YOLOv8 (Jocher et al., 2023) algorithm accelerates localization speed by narrowing down the size of a detection area.

This framework addressed not only the manual data annotation issue but also the imbalance of the initial dataset. As an unsupervised learning approach, AE is capable of training a model with unclassified data. With just the two uncoated nozzle images identified by the GLCM method, AE1 was

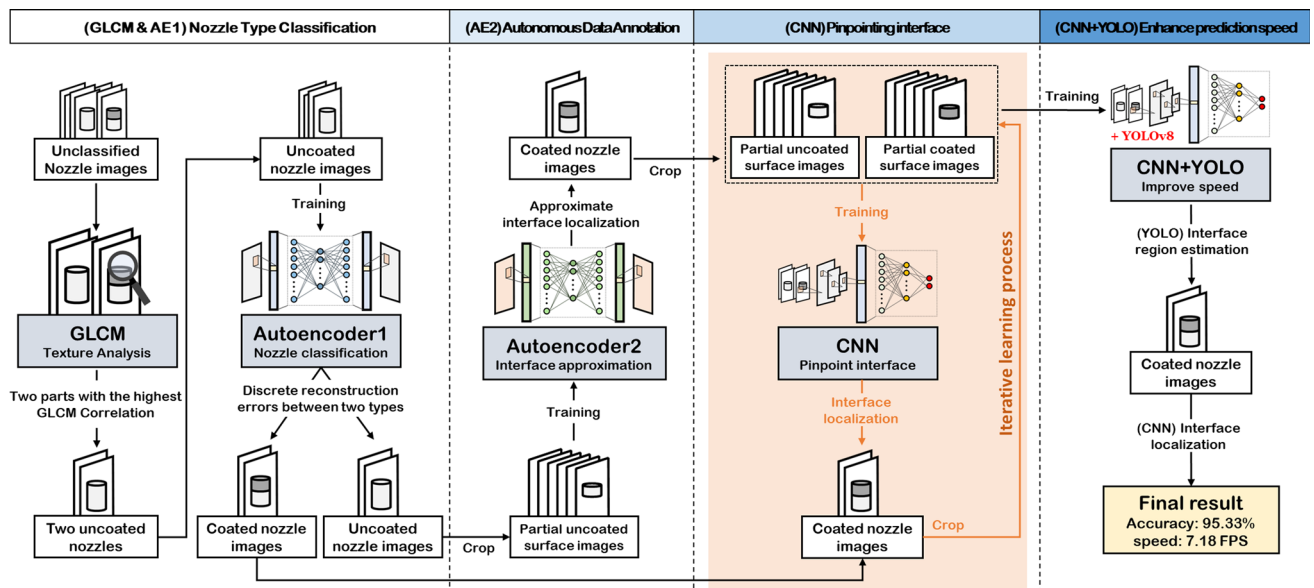


Fig. 1 Pipeline of the proposed framework

effectively trained, categorizing the imbalanced dataset into distinct coated and uncoated nozzle classes. On the other hand, AE2 was developed using exclusively uncoated surface images, thereby resolving the dependency for dataset balancing. For the construction of CNN models, cropped surface images were utilized. Having both coated and uncoated surfaces, it was possible to generate balanced datasets from the coated nozzle images. While the exact number of images in each class varied during the iterative learning process, approximately 120,000 coated and 100,000 uncoated surface images were collected.

In this section, the proposed framework is explained in detail. First, the image acquisition process is demonstrated. The second and third sections explain the texture analysis process and the two AE models that classified the types of nozzles and surfaces and generated an initial training dataset. Subsequently, how CNN-based detection models were built is illustrated. Lastly, the final deep learning model fabricated by integrating the YOLOv8 algorithm and the best CNN-based detection model is described.

### Procedure of data collection

Figure 2 depicts the setup to collect raw fuel injection nozzle images. The image acquisition system was comprised of an industrial monochrome camera (Cognex In-Sight 9000) connected to a laptop of which the specification was AMD Ryzen 5 5500U, 8 GB RAM, 256 GB SSD, and Windows 11. Every side of a part was captured by placing a nozzle on a turntable rotating with a speed of 1.6 rpm. Moreover, direct lighting was applied by an LED desk lamp to provide constant brightness. Eventually, 15 images per nozzle were

acquired at  $4096 \times 3000$  pixel resolution, yielding a dataset of 75 images from 5 uncoated and 525 images from 35 coated nozzles. Since the collected images included unnecessary background, the region of interest (ROI) of each image was cropped to  $512 \times 1024$  pixels based on detecting significant differences between adjacent pixels in both vertical and horizontal directions as boundaries by the developed computer vision algorithm. The image was cut further inward horizontally to exclude the curved interface region. The resultant images were mixed to begin this investigation from scratch. After the classification of the mixed image set into uncoated and coated nozzle categories by gray level co-occurrence matrix (GLCM) and AE methods, the set was further divided into training and test sets as per the requirements of the deep learning model. Table 1 details the configuration of the dataset.

### GLCM & autoencoder model 1—classification between coated and uncoated parts

AE1 model was developed under the hypothesis that it could produce variable reconstruction errors for coated and uncoated parts from their distinct features, and the reconstruction errors would serve as classification metrics. However, training an AE model requires a pre-categorized set of images, which was not initially available in the early stage of this investigation. To address this issue, a GLCM method was employed for obtaining an initial categorized image set for AE training. Proposed by Haralick (1979), GLCM is a statistical texture analysis technique that quantifies a two-dimensional histogram of paired pixels within a specific spatial distance.

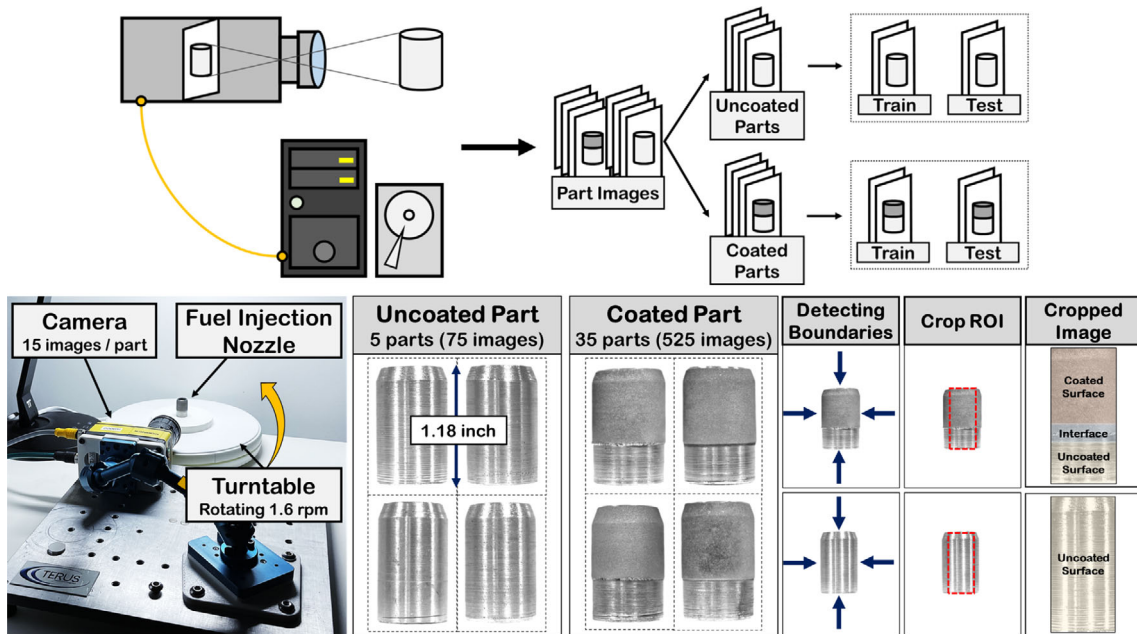


Fig. 2 Experimental setup for image acquisition and processing steps

Table 1 Configuration of dataset

Total		AE1		AE2		CNN	
Category	Dataset	Nozzle parts	Part images	Nozzle parts	Part images	Nozzle parts	Part images
Uncoated part	Train	2	30	5	75	0	0
	Test	3	45	0	0	0	0
Coated part	Train	0	0	15	225	= AE2	= AE2
	Test	35	525	20	300	= AE2	= AE2

GLCM of an image presents the frequency of pairs of pixels with a specific intensity based on parameters (Gadkari, 2004) as shown in Fig. 3.  $x$  and  $y$  is the horizontal and vertical coordinates of a point in an image. The size of an image is represented by  $w$  and  $h$ . An angle  $\theta$  is evaluated by the degree of rotation from the horizontal line passing through a reference point to a line penetrating another point and a reference point. Typically, angle  $\theta$  is quantized by 0, 45, 90, and 135 degrees. Furthermore,  $i$  and  $j$  are indices corresponding to the axis of a co-occurrence matrix and indicate pixel intensities.  $d$  is the distance between a reference point to another one. In consequence, the  $(i, j)$ th entry of GLCM  $g$  is defined as Eq. (1).

$$g_{ij, \theta=0^\circ} = \sum_{x=0}^w \sum_{y=0}^h \begin{cases} 1, & \text{if } I(x, y) = i \text{ and } I(x + d, y) = j \\ 0, & \text{otherwise} \end{cases}$$

$$g_{ij, \theta=45^\circ} = \sum_{x=0}^w \sum_{y=0}^h \begin{cases} 1, & \text{if } I(x, y) = i \text{ and } I(x + d, y - d) = j \\ 0, & \text{otherwise} \end{cases}$$

$$g_{ij, \theta=95^\circ} = \sum_{x=0}^w \sum_{y=0}^h \begin{cases} 1, & \text{if } I(x, y) = i \text{ and } I(x, y - d) = j \\ 0, & \text{otherwise} \end{cases}$$

$$g_{ij, \theta=135^\circ} = \sum_{x=0}^w \sum_{y=0}^h \begin{cases} 1, & \text{if } I(x, y) = i \text{ and } I(x - d, y - d) = j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

With a GLCM parameter distance of 1 and angles at 0, 45, 90, and 135 degrees, GLCM energy values were calculated for each whole surface of coated and uncoated nozzle parts by Eq. (2). Two nozzle parts with the highest GLCM energy values were selected for training AE1 since the higher GLCM energy value demonstrated more consistent textures in uncoated parts.

$$Energy = \sum_i \sum_j g_{ij}^2 \tag{2}$$

Subsequently, autoencoder model 1 (AE1) was developed to separate mixed images into coated and uncoated nozzle images thoroughly. AE1 comprised two CNN layers with

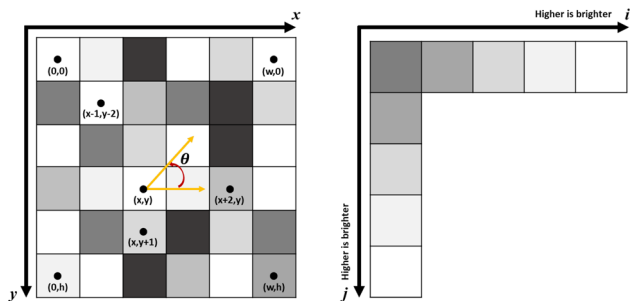


Fig. 3 Parameters of GLCM

a two-pixel stride and 3-pixel zero padding connected with fully connected layers and the latent space with 64 neurons as depicted in Fig. 4. This architecture was determined through a grid search. Among the gathered training sets, one uncoated part was used to train the model, and another one was utilized to get a maximum reconstruction error. The Adam optimizer was employed with a learning rate of 0.001, and the mean squared error (MSE) loss function was applied. The number of epochs for the training was 100. This training was done by PyTorch 1.13.1 with GPU acceleration (Paszke et al., 2019) operating on Ubuntu 20.04.5 LTS with an Intel i7-11700K CPU, 64GB RAM, and Nvidia RTX A5000 GPU. The same system was also utilized to develop the other deep learning models of this investigation.

Evaluating the performance of AE1 was anchored on maximum reconstruction errors, which were automatically established at 1.529 during the prediction by AE1 on the validation set. Notably, the minimum reconstruction error recorded for the coated nozzle parts was 1.565, consistently exceeding the maximum reconstruction error of 0.812 for the uncoated nozzle parts. This led to a flawless 100% classification accuracy by AE1 in distinguishing between the coated and uncoated nozzle parts. Additionally, the findings were supported by Fig. 5, which presented a validation of the GLCM energy analysis. While GLCM alone was not fully capable of differentiating between coated and uncoated parts, it did reveal a pattern: uncoated parts predominantly registered higher GLCM energy values compared to their coated counterparts.

Fig. 4 AE Model 1 configuration

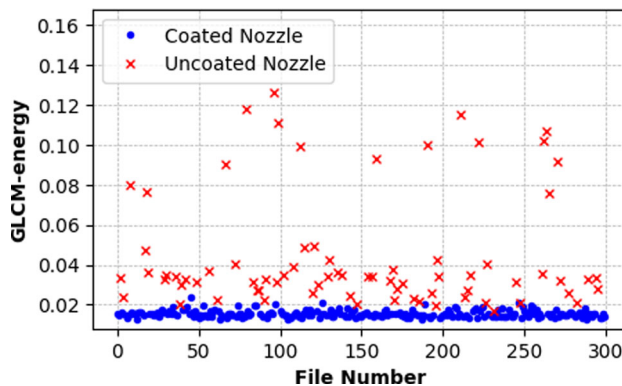
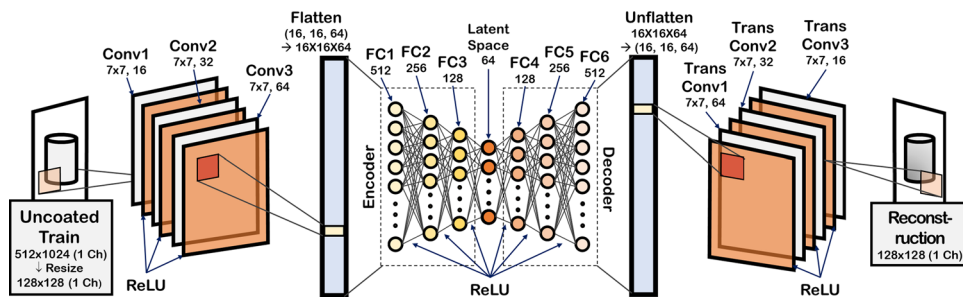


Fig. 5 GLCM energy values of the nozzle images

### Autoencoder model 2—generating training dataset without manual data labeling

Figure 6 illustrates GLCM analysis on cropped coated and uncoated surfaces at a size of 512 × 8 pixels. In this analysis, GLCM correlation values were computed by Eq. (3), where  $\mu$  and  $\sigma$  are the mean and standard deviation of  $g$ . It was also used to quantify linear dependencies across each type of surface (Gadkari, 2004).

$$\text{Correlation} = \frac{\sum_i \sum_j (ij) g_{ij} - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{3}$$

While uncoated surfaces generally yielded higher GLCM values, it was insufficient for precisely distinguishing between the two types of surfaces. Specifically, utilizing a threshold for GLCM correlation values, which was established through k-means clustering, resulted in a classification accuracy of less than 7% within a ± 6-pixel error range. Given these limitations, the study transitioned to implementing deep learning techniques for more precise identification of coating interfaces.

AE2 was designed to estimate interface locations and autonomously generate datasets for training CNN-based detection models, eliminating the need for manual labeling. The uncoated nozzle parts were already classified by AE1, enabling the immediate extraction of segmented images of uncoated surfaces. The input height of segmented images was

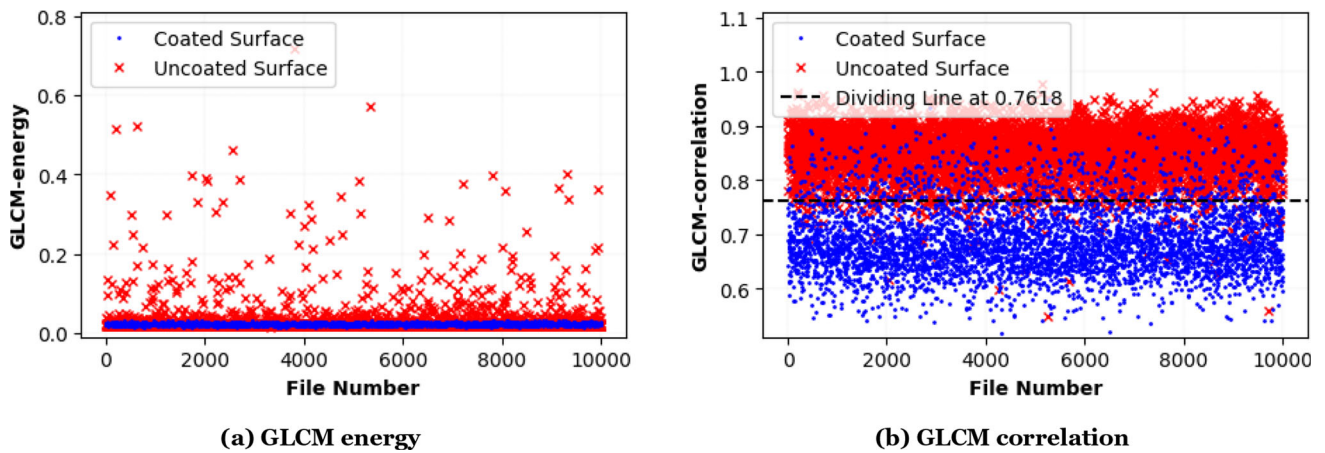
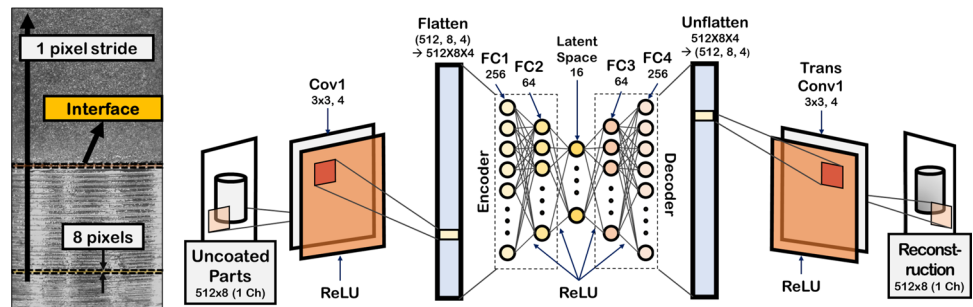


Fig. 6 GLCM analysis on coated and uncoated surfaces

Fig. 7 AE Model 2 configuration



defined as 8 pixels via a grid search, then cropped to dimensions of  $512 \times 8$  pixels in the vertical direction, employing a 1-pixel stride on the uncoated nozzle parts. Similar to AE1, the Adam optimizer was deployed with a learning rate of 0.001, and the MSE loss function was applied to the training process. The number of epochs for the training was 100. Figure 7 outlines the configuration of AE2, which was determined through a grid search.

In a manner analogous to AE-based anomaly detection (Yun et al., 2023a, 2023b), interface locations were estimated at regions displaying the highest reconstruction errors during prediction on a coated nozzle part. The mean average reconstruction error was 0.515 and the mean maximum reconstruction error was 1.972. In addition, the mean standard deviation of the reconstruction error was 0.226. AE2 located the interfaces with an accuracy of 84.38%. This study defined the success criteria as the capability to pinpoint the interface location in an image within a  $\pm 6$ -pixel margin of error, identically corresponding to the standard error range of  $\pm 0.127$  mm ( $\pm 0.005$  inches). The reference interface locations were determined by manual visual examination.

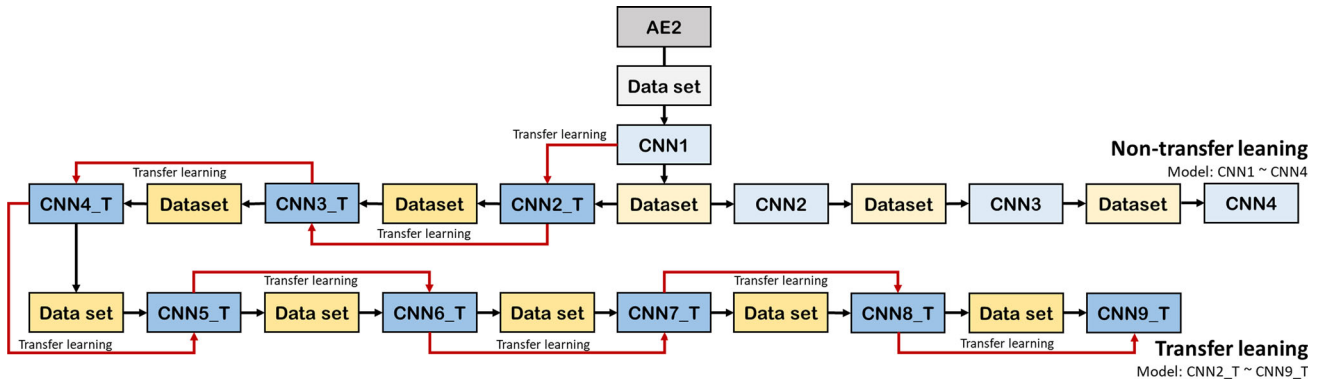
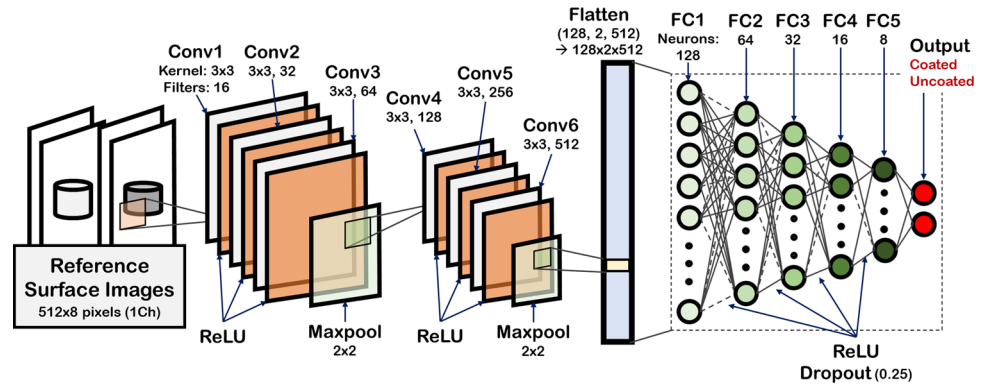
### CNN-based detection model—improved interface region detection

To develop a CNN-based detection model, partial images of a size of  $512 \times 8$  pixels were extracted from the coated nozzles with a 1-pixel stride. According to the estimated interface location by AE2, the upper side and the lower side of the interface were categorized autonomously as coated and uncoated image sets, respectively. Figure 8 depicts the configuration of a CNN-based detection model, determined through a grid search. The model deployed an Adam optimizer with a learning rate of 0.0001 and the cross-entropy loss function. The training process included the early stopping method. Specifically, the training was terminated when the training loss was below 0.0001.

The initial CNN-based detection model achieved 90.00% accuracy within a  $\pm 6$ -pixel error range. Further training, both with and without transfer learning, was conducted on a new dataset based on the estimated interface location by a prior CNN-based detection model. This process was repeated, resulting in 13 CNN-based detection models as described in Fig. 9. The layers were not frozen during the transfer learning processes. Overall, the models with transfer learning exhibited superior performance as shown in Fig. 10. The best model, CNN6\_T, achieved 95.33% accuracy within the  $\pm 6$ -pixel error range.

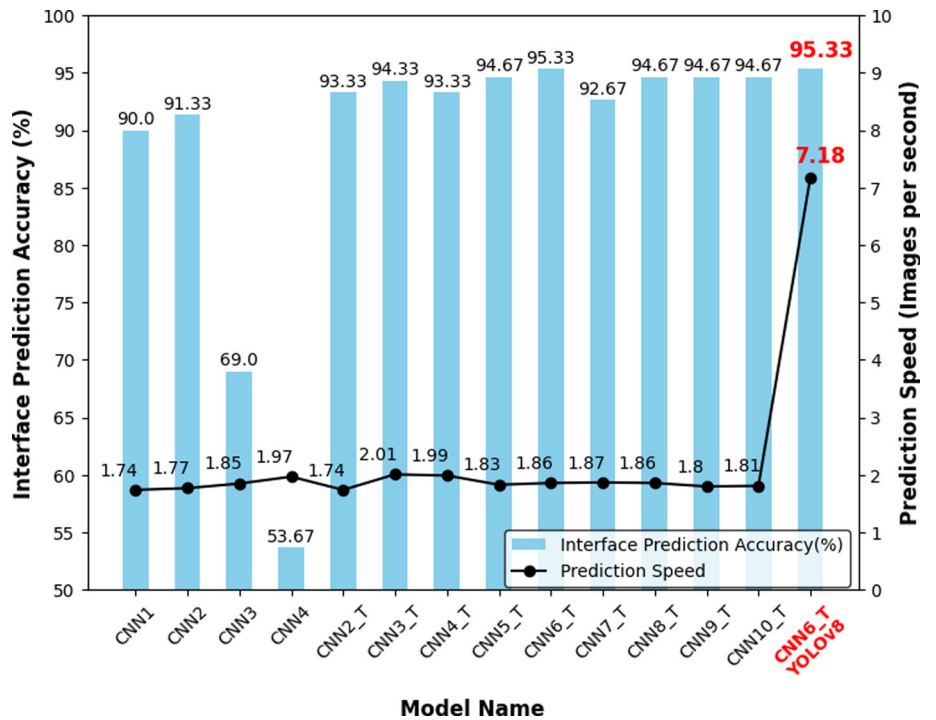


**Fig. 8** Configuration of the reference CNN-based detection model configuration & detection outline



**Fig. 9** Pipeline of transfer and non-transfer learning

**Fig. 10** CNN-based Detection Model Interface Detection Accuracy & Speed Comparison (Higher is better)



## CNN-based detection model + YOLO —improved detection speed

The CNN-based Detection Models, especially CNN6\_T, demonstrated high accuracy, yet their detection speed was approximately 1.5 to 1.8 images per second as shown in Fig. 10. As the developed framework executed detections from the bottom of a part image with a 1-pixel stride, it led to numerous detections to ascertain the estimated interface location. This is a typical detection speed issue regarding object detection problems when using a stride for detection. To address speed constraints in object detection, CNN-based algorithms such as Region-based Convolutional Neural Networks (R-CNN) (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), and Faster R-CNN (Ren et al., 2015) have been developed.

You Only Look Once (YOLO) (Redmon et al., 2016) is a one-stage real-time object detection algorithm that provides a balanced and acceptable combination of speed and accuracy for object detection (Terven & Cordova-Esparaza, 2023). A study (Kim et al., 2020) comparing Faster R-CNN, YOLO, and Single Shot MultiBox Detector (SSD) (Liu et al., 2016) demonstrated that while SSD was the fastest, its F1-score, Precision, Recall, and mAP was up to 11% lower. On the contrary, YOLO exhibited the best performance with a speed only 20% slower than SSD. Consequently, YOLO was chosen for this investigation requiring a balanced and acceptable combination of speed and accuracy.

Although employing YOLO, as the initial approach yielded 0% accuracy for pinpointing interfaces within a few pixels error range, it showed reasonable accuracy in identifying broader interface regions. To harness the advantage, this study devised a two-step prediction strategy, which integrated a YOLO model with a CNN-based prediction model, aimed to accelerate the prediction speed while maintaining high accuracy: Initially, YOLO suggests a probable interface area, which is then followed by a more focused CNN-based prediction within the narrowed-down region.

The YOLOv8n model was used to propose a region containing the interface by 0.25 of the default confidence threshold with the CNN6\_T model predicting within the proposed region. The datasets for training the YOLO model were extracted by a height of 128 pixels according to the same interface location used for developing the CNN6\_T model. The sample images of proposed interface regions by the trained YOLO model are depicted in Fig. 11. As Fig. 10 demonstrates, the detection speed with the region proposal from the trained YOLO model was approximately 4 times faster, reaching 7.18 images per second in comparison to other outcomes while maintaining the highest accuracy at 95.33%.

## Results and discussions

Annotating datasets is inevitable for supervised learning and results in human intervention. Hence, additional resources are consumed to build a model for an inspection application. The proposed framework overcame the issue by implementing multi-stage deep learning strategies. Without pre-defined datasets, manual labeling, and extensive feature engineering processes, it pinpointed the interfaces of the coated nozzles as well as categorized the types of the nozzles from a single uncoated nozzle. Consequently, this framework automated the inspection process thoroughly. In this section, the explanation and interpretation of the developed models are described based on t-distributed stochastic neighbor embedding (t-SNE) and integrated gradient methods to validate the reliability of the models. Moreover, the detection results of the CNN-based detection models are analyzed.

Figure 12 presents the latent space of AE2, the most compact dimensional region of partially cropped images corresponding to the three areas of uncoated, coated, and interface regions. It interprets the phenomena in the latent space and elucidates how AE2 discriminated between coated and uncoated surfaces and identifies the interface. Inputs for AE2 were partially cropped images ( $512 \times 8$  pixels) compressed in latent space features via the encoder. Subsequently, each latent space feature, represented by 16-dimensional values per AE2 architecture, was further reduced to a two-dimensional representation using t-SNE (Maaten & Hinton, 2008). The t-SNE calculates pairwise similarities between data points in the high-dimensional space, assigning higher probabilities to pairs of points that are close together and lower probabilities to those further apart. It achieves this by minimizing the Kullback–Leibler (KL) divergence between the high-dimensional and low-dimensional distributions of pairwise similarities, effectively making similar objects appear close together and dissimilar objects far apart in the lower-dimensional space; this optimization problem could be solved through gradient descent.

This reduction enables each partially cropped image to be represented as a single point in a two-dimensional plane, visualizing the latent space of AE2 for 1017 partial images extracted from a single image. The resulting visualization displays clusters within the latent space of AE2 according to uncoated, coated, and interface regions. Remarkably, the interface area, despite its adjacency to the uncoated region in the original image, appears farther from the uncoated area in the latent space, where AE2 underwent training. This consistent pattern across multiple sample images, as depicted in Fig. 13, affirms the ability of AE2 to estimate interfaces.

The initial CNN-based detection model, CNN1, trained on the AE2-generated dataset, yielded an accuracy of 90.00%. Accuracy increased to 95.33% via iterative learning with transfer learning and ultimately stabilized at 94.67% as

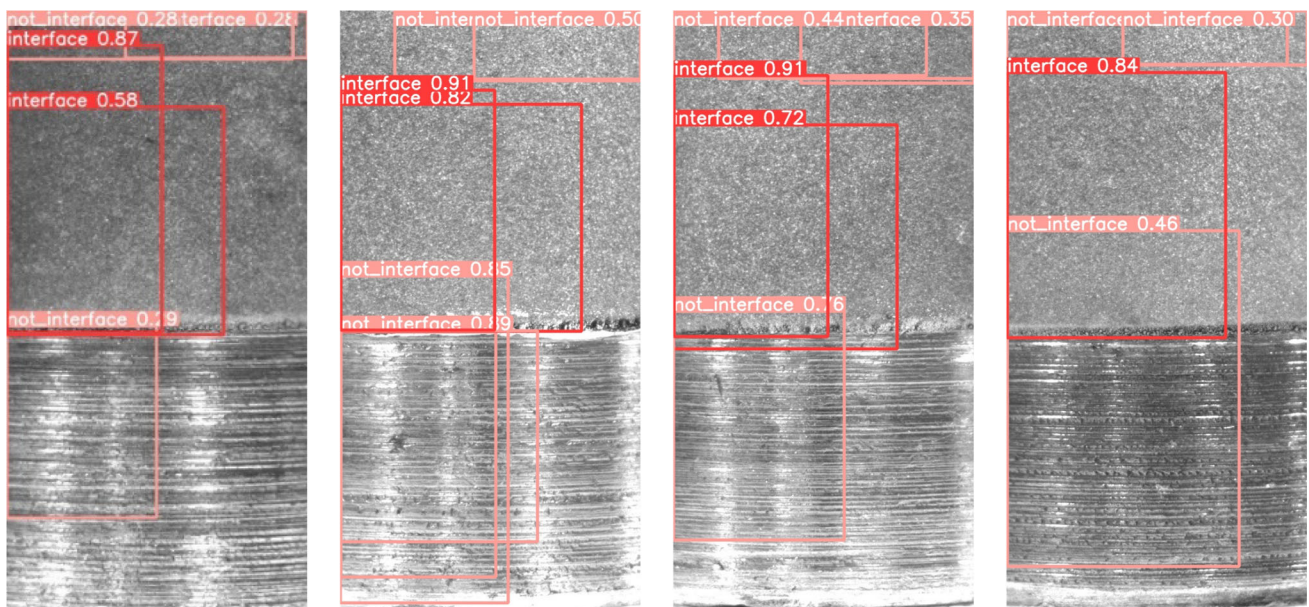


Fig. 11 Sample results by YOLOv8 detections

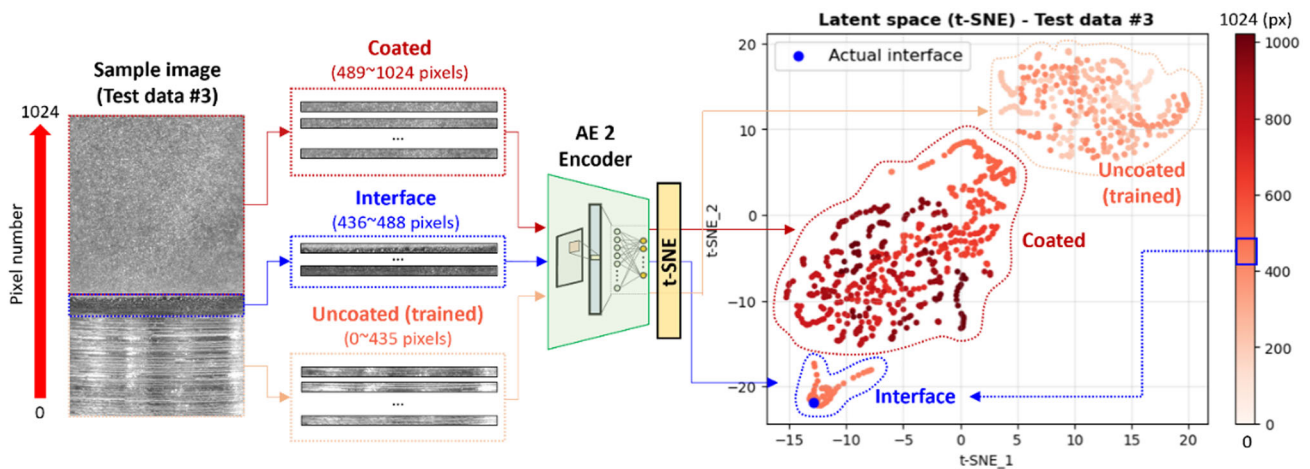


Fig. 12 AE2 latent space visualization result for a sample coated image (cropped) based on t-SNE

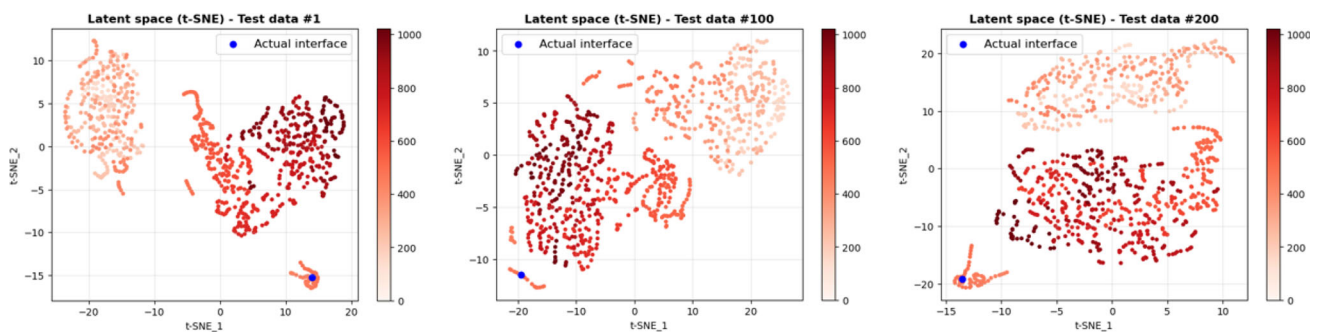
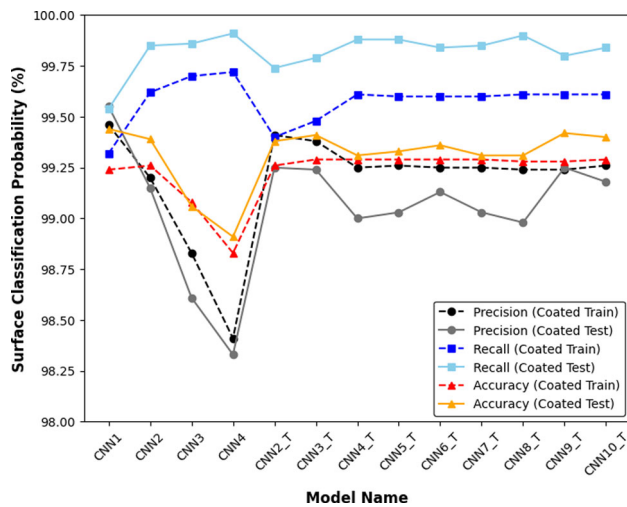


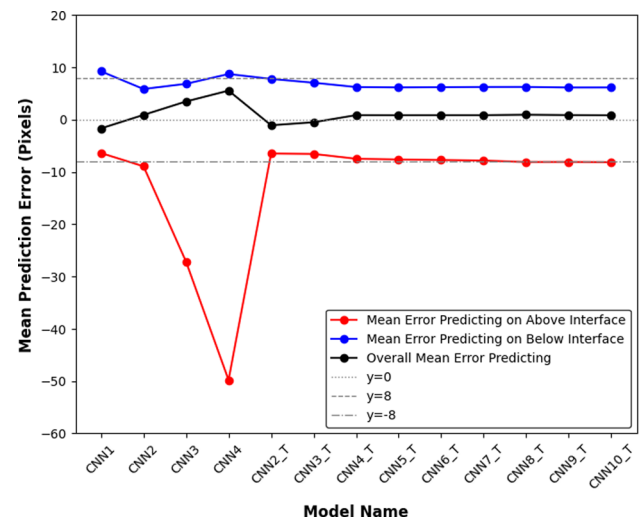
Fig. 13 AE2 latent space visualization results for sample coated images (cropped)



**Fig. 14** Detection precision, recall, and accuracy of surface classification

shown in Fig. 10. Conversely, iteration without transfer learning precipitously reduced accuracy from 91.33% to 69.00% by the third cycle. Figure 14 demonstrates precision, recall, and accuracy in surface classification detections, surpassing all 98.00%, yet plummeting precision and accuracy without transfer learning. Precision, recall, and accuracy became stable from CNN4\_T onwards in both training and testing datasets when applying transfer learning, suggesting their consistency contributes to interface detection accuracy convergence, as similar training datasets were utilized from CNN4\_T. In contrast, without transfer learning, the variation in training datasets, not utilizing pre-trained neural network weights, led to inconsistent results. Especially, there are metric discrepancies between surface classification detection and interface location. Given the objective of this study, the detection accuracy of interface locations was deemed the most vital metric.

Figure 15 shows the mean error, defined as the discrepancy between estimated by the CNN-based detection models and actual interface locations on coated training parts. The blue line represents the mean error that predicted interface locations were above the actual interface, and vice versa for the red line with the black line demonstrating the overall mean error. Notably, with iterative learning without transfer learning from CNN1 to CNN2, the red line shows a decrease from 6.37 pixels to 8.90 pixels, while the blue and black lines remain consistent. Given that the height of the partially cropped images was 8 pixels, it is inferred that dataset labeling along the interfaces in the red line case for CNN2 was inaccurate since an overall error range exceeded 8 pixels. Consequently, numerous coated part images estimated by CNN2, CNN3, and CNN4 included uncoated surfaces, causing a continuous decrease in accuracy. These models, trained without transfer learning and hence without any pre-trained

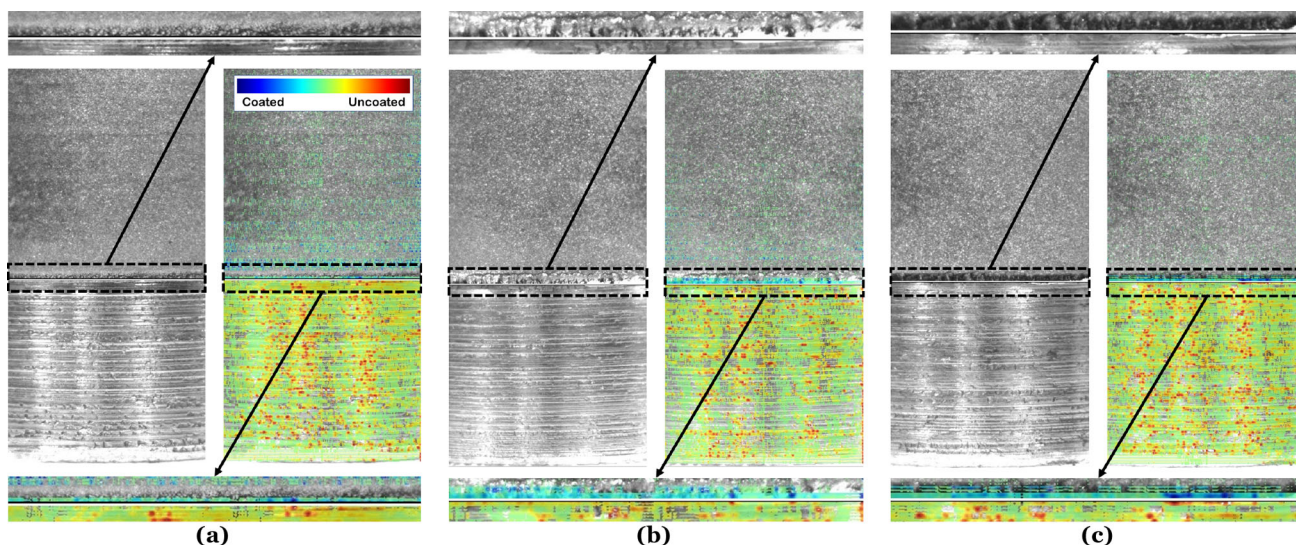


**Fig. 15** Detection mean error from the actual interface location (for coated train part)

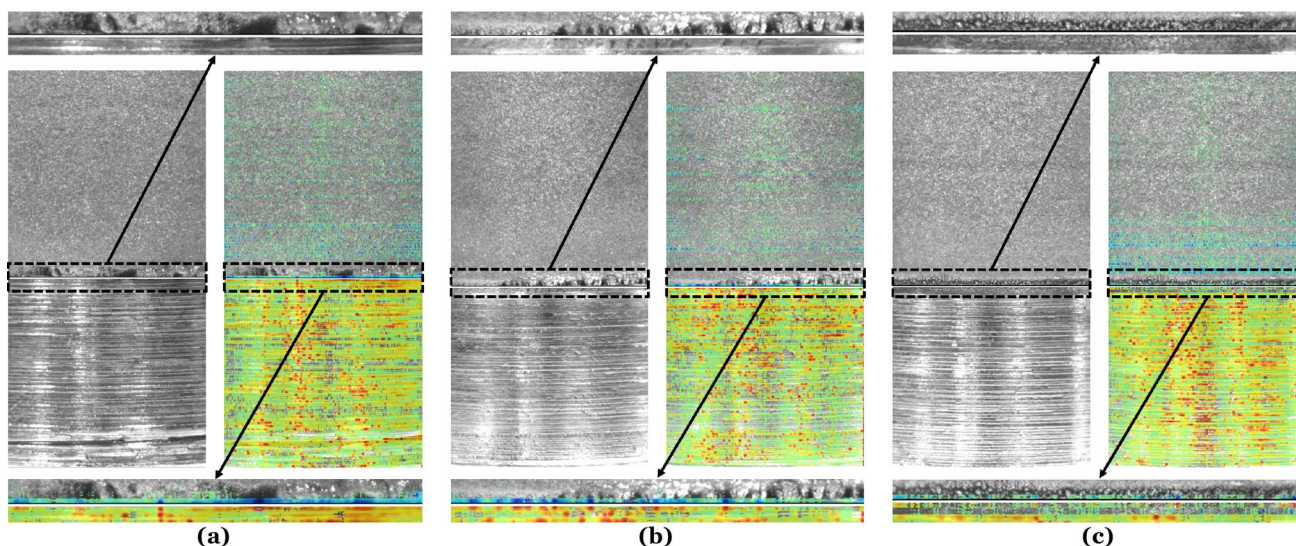
weights or bias, resulted in deteriorating accuracy during training and were built with high errors. On the contrary, the models utilizing transfer learning displayed convergence in mean error as they were built upon a pre-established network with pre-trained weights and bias, reducing vulnerability due to misclassified images.

To interpret and validate the reliability of the optimal CNN-based detection model, CNN\_6T, integrated gradients were used to generate heatmaps, and the deletion metric was applied (Qi et al., 2020; Selvaraju et al., 2020). In the jet colormap, red signifies key pixels for uncoated class detection, and blue represents those crucial for coated areas. Figure 16 depicts the heatmap examples of correctly predicted images and Fig. 17 illustrates heatmaps for correctly predicted images, including those with complex interface geometry. The black and white lines indicate the actual and estimated interface locations, respectively. The highlighted heatmaps confirm the clear separation between two classes along the interface, even in instances of complex geometry with diverse shapes as in Fig. 17a, b, or blurred boundaries as in Fig. 17c that make interface identification challenging. Furthermore, the heatmap examples of incorrectly predicted images are displayed in Fig. 18. Figure 18a suggests excessive light reflection obscuring the interface feature, Fig. 18b demonstrates the failure of the detection model to extract features near the interface, and Fig. 18c shows wrong detection due to the existence of multiple candidates of potential interface location. Therefore, future research involves developing a multiple-layered and configured deep learning model to resolve these issues and enhance accuracy.

The deletion metric was implemented to examine the effects of the weight of each pixel in the CNN6\_T detection model. As in Fig. 19, pixels with high weights were firstly



**Fig. 16** Heatmap examples of correctly predicted images



**Fig. 17** Heatmap examples of correctly predicted images having complicated geometry of interface

deleted as per a specified deletion ratio, with zero-weighted pixels being disregarded during this operation. Figure 20 illustrates that the detection accuracy declined steeply upon the removal of the important pixels. This observation validates the proper construction of the CNN6\_T detection model and the weighing of the pixels during interface detection. Furthermore, the area under the curve (AUC) was assessed, yielding a score of 0.1135, confirming the reliability of the developed predictive model according to the reference (Pet-siuk et al., 2018).

Lastly, a YOLO model was employed to improve the detection speed for industrial applications. The developed CNN-based detection model predicted from the bottom of

an image with a 1-pixel stride, and it resulted in a time-consuming detection speed of 1.5 to 1.8 images per second. This study applied YOLO alone at first, showing 0% accuracy for the detection within the  $\pm 6$ -pixel error range. Therefore, a two-stage detection pipeline was devised proposing the broad range of an interface region by a YOLO model and predicting from the bottom of the proposed region by the CNN model. Even though YOLO could not predict accurately the interface within the error range, it successfully proposed an interface region with a wide range of an area which led to up to 4 times faster detection speed, 7.18 images per second, without a reduction in accuracy.

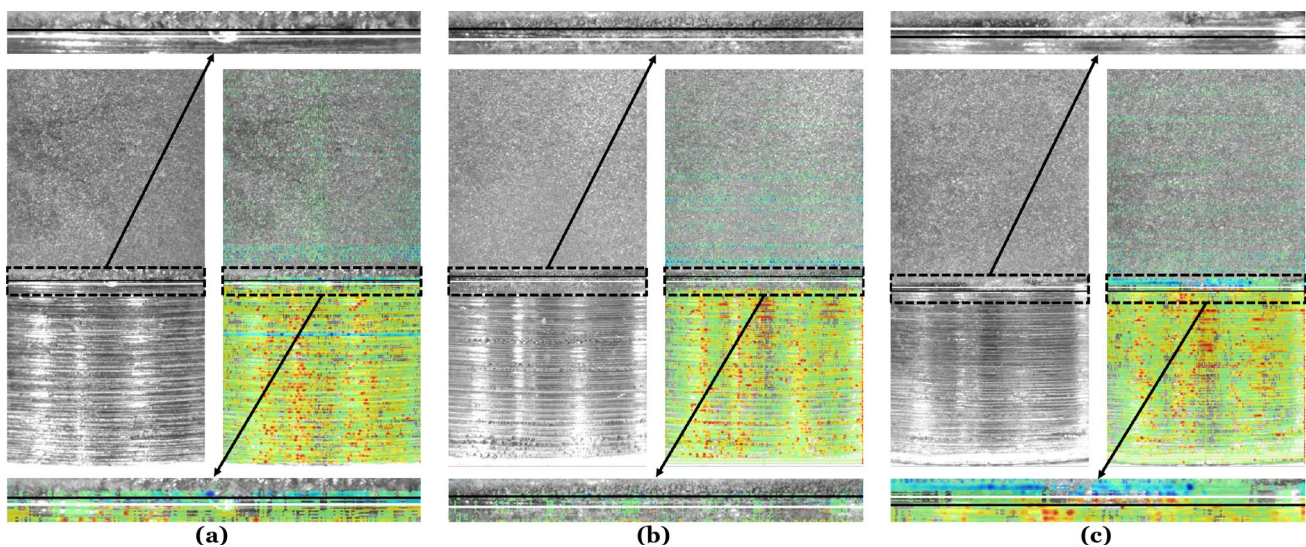


Fig. 18 Heatmap examples of incorrectly predicted images

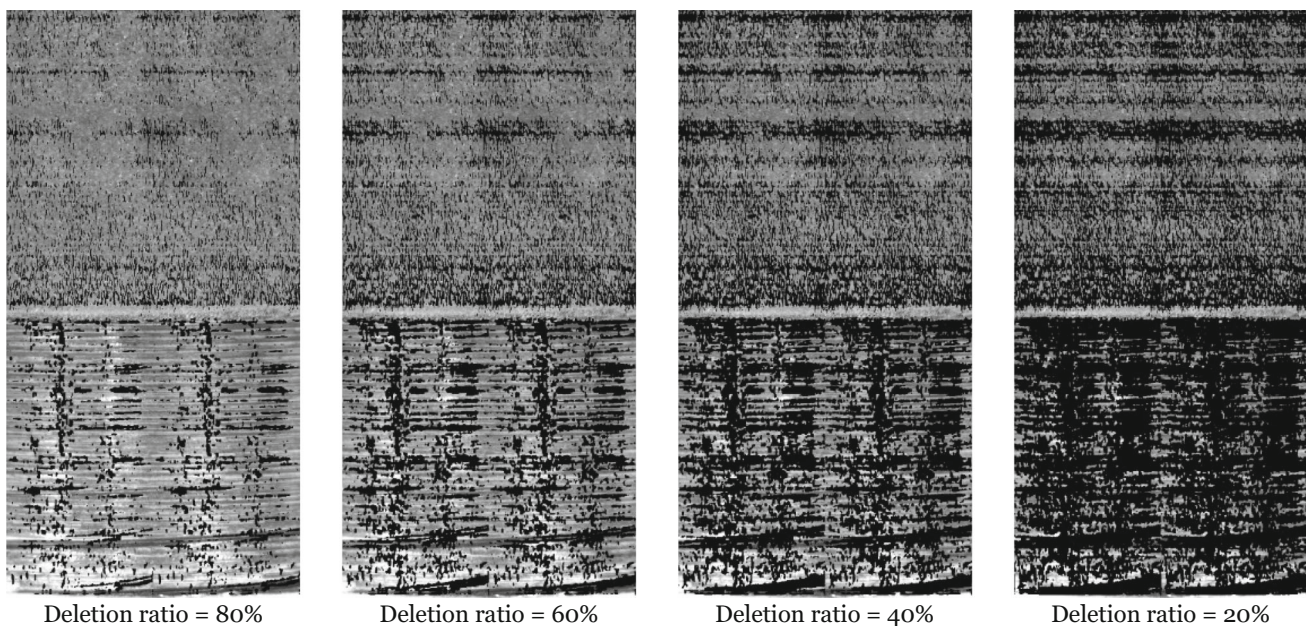


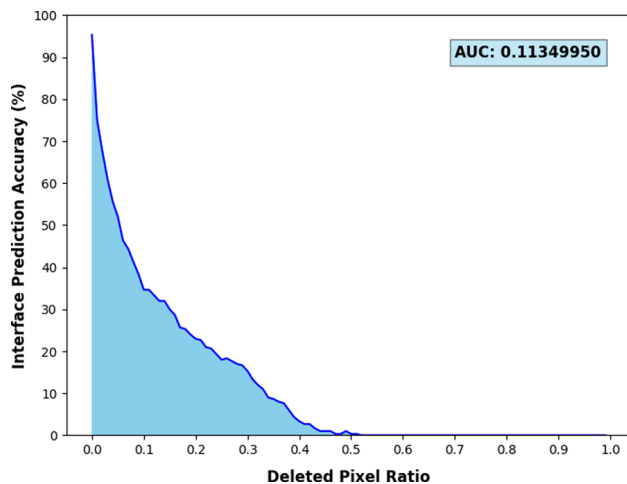
Fig. 19 Pixel deleted sample images

## Conclusion

This study proposed the framework for distinguishing the types of fuel injection nozzles and locating interfaces between coated and uncoated surfaces with autonomous data annotation. While computer vision techniques reveal the progress in the visual inspection domain, it suggests that an elaborate algorithm would be necessary to detect complex objects. The features of the fuel injection nozzles varied across different parts and even the viewing angles. As a result, this research implemented multi-stage deep learning strategies to develop an automated visual inspection

method. By conjugating autoencoder and CNN configurations, it addressed challenges in the application of coating inspection to classify the coated and uncoated nozzles and pinpoint the interface locations. Furthermore, the application of the YOLOv8 improved the detection speed of the CNN-based detection model. Finally, the interpretation and explanation of the deep learning models were described and validated their robustness.

Nonetheless, the framework still has limitations. Although the GLCM approach identified coated and uncoated nozzles without pre-defined datasets, it required at least two parts to be distinguished for developing an AE model.



**Fig. 20** Result of deletion metric

Additionally, this approach is not applicable to parts with more than two surfaces. The detection speed still requires optimization to exceed 30 images per second for real-time applications. Lastly, this framework is not validated for images from different sources and environments. Therefore, future research will focus on the adaptability of the framework concerning other coatings, materials, various inspection conditions, and intricate geometries, along with investigating further enhancements in detection speed and accuracy. Furthermore, a more affordable image acquisition setup will be investigated.

Then, this framework is expected to be implemented in a wide range of industries such as aerospace, automotive, and electronics manufacturing, where the feature of surfaces plays an important role in enhancing the performance, durability, and protection of components. Eventually, it is anticipated that this investigation would be a practical solution for real-time applications in industries by the achieved reasonable detection accuracy and speed even with less effort in training deep learning models. In addition, the ability to generate training datasets with autonomous data annotation could lead to cost savings, increased efficiency, and reduced potential for human error. Consequently, it is expected that this framework will aid cost and reliability issues in the inspection process of manufacturing domains.

**Acknowledgements** This research was supported by Indiana Next Generation Manufacturing Competitiveness Center (IN-MaC).

**Author contributions** CH conceptualization, data acquisition, methodology, investigation, visualization, writing-original draft. JL methodology, visualization, validation, writing-review. MJ methodology, supervision, resources, project administration, funding acquisition, writing-review. SL supervision, writing-review. HY conceptualization, methodology, validation, supervision, writing-review & editing.

**Funding** Open Access funding enabled and organized by KAIST. This work was partially supported by the National Science Foundation under Grant No. NSF AM-2125826 and the Technology Innovation Program 20015060 funded by MOTIE, Korea.

**Data availability** The datasets generated and analyzed during the current study are not publicly available due the fact that they constitute an excerpt of research in progress but are available from the corresponding author on reasonable request.

## Declarations

**Conflict of interest** The authors have no competing interests to declare that are relevant to the content of this article.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, 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 changes were made. 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/4.0/>.

## References

- Al Hasan, Md. M., Tahsin Mostafiz, M., Le An, T., Julia, J., Vashistha, N., Taheri, S., & Asadizanjani, N. (2023). EVHA: explainable vision system for hardware testing and assurance—An overview. *ACM Journal on Emerging Technologies in Computing Systems*, 19(3), 25. <https://doi.org/10.1145/3590772>
- Alonso, V., Dacal-Nieto, A., Barreto, L., Amaral, A., & Rivero, E. (2019). Industry 4.0 implications in machine vision metrology: An overview. *Procedia Manufacturing*, 41, 359–366. <https://doi.org/10.1016/j.promfg.2019.09.020>
- Babic, M., Farahani, M. A., & Wuest, T. (2021). Image based quality inspection in smart manufacturing systems: A literature review. *Procedia CIRP*, 103, 262–267. <https://doi.org/10.1016/j.procir.2021.10.042>
- Bai, J., & Feng, X.-C. (2007). Fractional-order anisotropic diffusion for image denoising. *IEEE Transactions on Image Processing*, 16(10), 2492–2502. <https://doi.org/10.1109/TIP.2007.904971>
- Bengio, Y., Courville, A., & Vincent, P. (2014). *Representation Learning: A Review and New Perspectives* (arXiv:1206.5538). arXiv. <https://doi.org/10.48550/arXiv.1206.5538>
- Brigham, E. O., & Morrow, R. E. (1967). The fast Fourier transform. *IEEE Spectrum*, 4(12), 63–70. <https://doi.org/10.1109/MSPEC.1967.5217220>
- Chin, R. T., & Harlow, C. A. (1982). Automated visual inspection: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. <https://doi.org/10.1109/TPAMI.1982.4767309>
- Chow, J. K., Su, Z., Wu, J., Tan, P. S., Mao, X., & Wang, Y. H. (2020). Anomaly detection of defects on concrete structures with the convolutional autoencoder. *Advanced Engineering Informatics*, 45, 101105. <https://doi.org/10.1016/j.aei.2020.101105>
- Cooper, C., Zhang, J., Huang, J., Bennett, J., Cao, J., & Gao, R. X. (2023). Tensile strength prediction in directed energy deposition through physics-informed machine learning and Shapley additive explanations. *Journal of Materials Processing Technology*, 315, 117908. <https://doi.org/10.1016/J.JMATPROTEC.2023.117908>
- Daniel, E. (2018). Optimum wavelet-based homomorphic medical image fusion using hybrid genetic-grey wolf optimization algorithm. *IEEE Sensors Journal*, 18(16), 6804–6811. <https://doi.org/10.1109/JSEN.2018.2822712>

- Doering, E. R., Havrilla, G. J., & Miller, T. C. (2004). Disilicide diffusion coating inspection by micro X-ray fluorescence imaging. *Journal of Nondestructive Evaluation*, 23(3), 95–105. <https://doi.org/10.1023/B:JONE.0000048865.96417.BC>
- Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., & Vincent, P. (2010). Why does unsupervised pre-training help deep learning? *The Journal of Machine Learning Research*, 11, 625–660. <https://doi.org/10.5555/1756006.1756025>
- Feng, S., Yu, H., & Duarte, M. F. (2020). Autoencoder based sample selection for self-taught learning. *Knowledge-Based Systems*, 192, 105343. <https://doi.org/10.1016/J.KNOSYS.2019.105343>
- Ficzere, M., Mészáros, L. A., Kállai-Szabó, N., Kovács, A., Antal, I., Nagy, Z. K., & Galata, D. L. (2022). Real-time coating thickness measurement and defect recognition of film coated tablets with machine vision and deep learning. *International Journal of Pharmaceutics*. <https://doi.org/10.1016/J.IJPHARM.2022.121957>
- Gadkari, D. (2004). Image Quality Analysis Using GLCM. *Electronic Theses and Dissertations, University of Central Florida*. <https://stars.library.ucf.edu/etd/187>
- Girshick, R. (2015). Fast R-CNN. *IEEE International Conference on Computer Vision (ICCV)*, 2015, 1440–1448. <https://doi.org/10.1109/ICCV.2015.169>
- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, 580–587. <https://doi.org/10.1109/CVPR.2014.81>
- Golnabi, H., & Asadpour, A. (2007). Design and application of industrial machine vision systems. *Robotics and Computer-Integrated Manufacturing*, 23(6), 630–637. <https://doi.org/10.1016/j.rcim.2007.02.005>
- Gonzalez, R., & Faisal, Z. (2019). *Digital Image Processing Second Edition*.
- Graps, A. (1995). An Introduction to Wavelets. *IEEE Computational Science and Engineering*, 2(2), 50–61. <https://doi.org/10.1109/99.388960>
- Gryzgoridis, J. (2012). Laser based nondestructive inspection techniques. *Journal of Nondestructive Evaluation*, 31(4), 295–302. <https://doi.org/10.1007/S10921-012-0144-X>
- Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, 77, 354–377. <https://doi.org/10.1016/j.patcog.2017.10.013>
- Guan, Y., Ghorbani, A. A., & Belacel, N. (2003). Y-means: A clustering method for intrusion detection. *Canadian Conference on Electrical and Computer Engineering*, 2, 1083–1086. <https://doi.org/10.1109/CCECE.2003.1226084>
- Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI-Explainable artificial intelligence. *Science Robotics*. <https://doi.org/10.1126/scirobotics.aay7120>
- Gunraj, H., Guerrier, P., Fernandez, S., & Wong, A. (2023). Solder-Net: Towards trustworthy visual inspection of solder joints in electronics manufacturing using explainable artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*. <https://doi.org/10.1609/aaai.v37i113.26858>
- Haralick, R. M. (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5), 786–804. <https://doi.org/10.1109/PROC.1979.11328>
- Harris, D. H. (1969). The Nature of Industrial Inspection. *Human Factors*, 11(2), 139–148. <https://doi.org/10.1177/001872086901100207>
- Jacob, A., Windhuber, K., Ranke, D., & Lanza, G. (2018). Planning, evaluation and optimization of product design and manufacturing technology chains for new product and production technologies on the example of additive manufacturing. *Procedia CIRP*, 70, 108–113. <https://doi.org/10.1016/J.PROCIR.2018.02.049>
- Jain, P., & Tyagi, V. (2015). LAPB: Locally adaptive patch-based wavelet domain edge-preserving image denoising. *Information Sciences*, 294, 164–181. <https://doi.org/10.1016/J.INS.2014.09.060>
- Joher, G., Chaurasia, A., & Qiu, J. (2023). *YOLO by Ultralytics*. <https://github.com/ultralytics/ultralytics>
- Jung, C., Yang, Q., Sun, T., Fu, Q., & Song, H. (2017). Low light image enhancement with dual-tree complex wavelet transform. *Journal of Visual Communication and Image Representation*, 42, 28–36. <https://doi.org/10.1016/J.JVCIR.2016.11.001>
- Katirci, R., Yılmaz, E. K., Kaynar, O., & Zontul, M. (2021). Automated evaluation of Cr-III coated parts using Mask RCNN and ML methods. *Surface and Coatings Technology*. <https://doi.org/10.1016/J.SURFCOAT.2021.127571>
- Khatami, A., Khosravi, A., Nguyen, T., Lim, C. P., & Nahavandi, S. (2017). Medical image analysis using wavelet transform and deep belief networks. *Expert Systems with Applications*, 86, 190–198. <https://doi.org/10.1016/J.ESWA.2017.05.073>
- Kim, J., Sung, J.-Y., & Park, S. (2020). Comparison of Faster-RCNN, YOLO, and SSD for Real-Time Vehicle Type Recognition. *2020 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*, 1–4. <https://doi.org/10.1109/ICCE-Asia49877.2020.9277040>
- Kozamernik, N., & Bračun, D. (2020). Visual inspection system for anomaly detection on KTL coatings using variational autoencoders. *Procedia CIRP*, 93, 1558–1563. <https://doi.org/10.1016/j.procir.2020.04.114>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*. <https://doi.org/10.1038/nature14539>
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2323. <https://doi.org/10.1109/5.726791>
- Lee, J., Noh, I., Lee, J., & Lee, S. W. (2022). development of an explainable fault diagnosis framework based on sensor data imagification: A case study of the robotic spot-welding process. *IEEE Transactions on Industrial Informatics*, 18(10), 6895–6904. <https://doi.org/10.1109/TII.2021.3134250>
- Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12), 6999–7019. <https://doi.org/10.1109/TNNLS.2021.3084827>
- Liu, T., Lough, C. S., Sehhat, H., Ren, Y. M., Christofides, P. D., Kinzel, E. C., & Leu, M. C. (2022). In-situ infrared thermographic inspection for local powder layer thickness measurement in laser powder bed fusion. *Additive Manufacturing*, 55, 102873. <https://doi.org/10.1016/J.ADDMA.2022.102873>
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9905 LNCS, 21–37. [https://doi.org/10.1007/978-3-319-46448-0\\_2](https://doi.org/10.1007/978-3-319-46448-0_2)
- Luisier, F., Blu, T., & Unser, M. (2007). A new SURE approach to image denoising: Interscale orthonormal wavelet thresholding. *IEEE Transactions on Image Processing : A Publication of the IEEE Signal Processing Society*, 16(3), 593–606. <https://doi.org/10.1109/TIP.2007.891064>
- Lundstrom, D. D., Huang, T., & Razaviyayn, M. (2022). A Rigorous Study of Integrated Gradients Method and Extensions to Internal Neuron Attributions. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *Proceedings of the 39th International Conference on Machine Learning*. PMLR. <https://proceedings.mlr.press/v162/lundstrom22a.html>
- McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* (arXiv: 1802.03426). arXiv. <https://doi.org/10.48550/arXiv.1802.03426>



- Megaw, E. D. (1979). Factors affecting visual inspection accuracy. *Applied Ergonomics*, 10(1), 27–32. [https://doi.org/10.1016/0003-6870\(79\)90006-1](https://doi.org/10.1016/0003-6870(79)90006-1)
- Nasimi, R., & Moreu, F. (2021). A methodology for measuring the total displacements of structures using a laser–camera system. *Computer-Aided Civil and Infrastructure Engineering*, 36(4), 421–437. <https://doi.org/10.1111/MICE.12652>
- Nason, G. P., & Silverman, B. W. (1995). The stationary wavelet transform and some statistical applications. In A. Antoniadis & G. Oppenheim (Eds.), *Wavelets and Statistics*. Springer. [https://doi.org/10.1007/978-1-4612-2544-7\\_17](https://doi.org/10.1007/978-1-4612-2544-7_17)
- Noble, J. A. (1995). From inspection to process understanding and monitoring: A view on computer vision in manufacturing. *Image and Vision Computing*, 13(3), 197–214. [https://doi.org/10.1016/0262-8856\(95\)90840-5](https://doi.org/10.1016/0262-8856(95)90840-5)
- Oztemel, E., & Gursev, S. (2020). Literature review of Industry 4.0 and related technologies. *Journal of Intelligent Manufacturing*, 31(1), 127–182. <https://doi.org/10.1007/S10845-018-1433-8>
- Park, J., Han, C., Jun, M. B. G., & Yun, H. (2023). Autonomous robotic bin picking platform generated from human demonstration and YOLOv5. *Journal of Manufacturing Science and Engineering*. <https://doi.org/10.1115/1.4063107>
- Park, J. K., Kwon, B. K., Park, J. H., & Kang, D. J. (2016). Machine learning-based imaging system for surface defect inspection. *International Journal of Precision Engineering and Manufacturing - Green Technology*, 3(3), 303–310. <https://doi.org/10.1007/S40684-016-0039-X>
- Park, M., Jin, J. S., Au, S. L., & Luo, S. (2008). Pattern recognition from segmented images in automated inspection systems. *Proceedings - 2008 International Symposium on Ubiquitous Multimedia Computing, UMC*, 87–92. <https://doi.org/10.1109/UMC.2008.26>
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates Inc.
- Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized Input Sampling for Explanation of Black-box Models. *British Machine Vision Conference 2018, BMVC 2018*. <https://arxiv.org/abs/1806.07421v3>
- Psarommatis, F., May, G., Dreyfus, P.-A., & Kiritsis, D. (2019). Zero defect manufacturing: State-of-the-art review, shortcomings and future directions in research. *Article in International Journal of Production Research*, 58(1), 1–17. <https://doi.org/10.1080/00207543.2019.1605228>
- Psarommatis, F., Sousa, J., Mendonça, J. P., & Kiritsis, D. (2022). Zero-defect manufacturing the approach for higher manufacturing sustainability in the era of industry 4.0: A position paper. *International Journal of Production Research*, 60(1), 73–91. <https://doi.org/10.1080/00207543.2021.1987551>
- Qi, Z., Khorram, S., & Fuxin, L. (2020). Visualizing deep networks by optimizing with integrated gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07), 11890–11898. <https://doi.org/10.1609/AAAI.V34I07.6863>
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 779–788. <https://doi.org/10.1109/CVPR.2016.91>
- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates Inc.
- Ren, Z., Fang, F., Yan, N., & Wu, Y. (2022). State of the art in defect detection based on machine vision. *International Journal of Precision Engineering and Manufacturing-Green Technology*, 9(2), 661–691. <https://doi.org/10.1007/s40684-021-00343-6>
- Rusk, N. (2016). Deep learning. *Nature Methods*, 13(1), 35. <https://doi.org/10.1038/nmeth.3707>
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2), 336–359. <https://doi.org/10.1007/S11263-019-01228-7>
- Shinde, P. P., Pai, P. P., & Adiga, S. P. (2022). Wafer defect localization and classification using deep learning techniques. *IEEE Access*, 10, 39969–39974. <https://doi.org/10.1109/ACCESS.2022.3166512>
- Singh, S. A., & Desai, K. A. (2022). Automated surface defect detection framework using machine vision and convolutional neural networks. *Journal of Intelligent Manufacturing*, 34(4), 1995–2011. <https://doi.org/10.1007/S10845-021-01878-W>
- Terven, J. R., & Cordova-Esparaza, D. M. (2023). A Comprehensive Review of YOLO: From YOLOv1 to YOLOv8 and Beyond. <https://arxiv.org/abs/2304.00501v1>
- van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86), 2579–2605.
- Wang, D., & Shang, Y. (2014). A new active labeling method for deep learning. *International Joint Conference on Neural Networks (IJCNN)*, 2014, 112–119. <https://doi.org/10.1109/IJCNN.2014.6889457>
- Wang, F., Zhao, Z., Zhai, Z., Shang, Z., Yan, R., & Chen, X. (2023). Explainability-driven model improvement for SOH estimation of lithium-ion battery. *Reliability Engineering & System Safety*, 232, 109046. <https://doi.org/10.1016/J.RESS.2022.109046>
- Wang, J., Fu, P., & Gao, R. X. (2019). Machine vision intelligence for product defect inspection based on deep learning and Hough transform. *Journal of Manufacturing Systems*, 51, 52–60. <https://doi.org/10.1016/j.jmsy.2019.03.002>
- Wang, X. Y., Yang, H. Y., & Fu, Z. K. (2010). A New Wavelet-based image denoising using undecimated discrete wavelet transform and least squares support vector machine. *Expert Systems with Applications*, 37(10), 7040–7049. <https://doi.org/10.1016/J.ESWA.2010.03.014>
- Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1), 37–52. [https://doi.org/10.1016/0169-7439\(87\)80084-9](https://doi.org/10.1016/0169-7439(87)80084-9)
- Xu, X., Wang, Y., & Chen, S. (2016). Medical image fusion using discrete fractional wavelet transform. *Biomedical Signal Processing and Control*, 27, 103–111. <https://doi.org/10.1016/J.BSPC.2016.02.008>
- Yang, Y., Su, Z., & Sun, L. (2010). Medical image enhancement algorithm based on wavelet transform. *Electronics Letters*, 46(2), 120–121. <https://doi.org/10.1049/EL.2010.2063>
- Yun, H., Kim, E., Kim, D. M., Park, H. W., & Jun, M. B. G. (2023a). Machine learning for object recognition in manufacturing applications. *International Journal of Precision Engineering and Manufacturing*, 24(4), 683–712. <https://doi.org/10.1007/S12541-022-00764-6>

Yun, H., Kim, H., Jeong, Y. H., & Jun, M. B. G. (2023b). Autoencoder-based anomaly detection of industrial robot arm using stethoscope based internal sound sensor. *Journal of Intelligent Manufacturing*, 34(3), 1427–1444. <https://doi.org/10.1007/s10845-021-01862-4>

Yun, J. P., Shin, W. C., Koo, G., Kim, M. S., Lee, C., & Lee, S. J. (2020). Automated defect inspection system for metal surfaces based on

deep learning and data augmentation. *Journal of Manufacturing Systems*, 55, 317–324. <https://doi.org/10.1016/j.jmsy.2020.03.009>

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