This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2024.3395316, IEEE Transactions on Instrumentation and Measurement

Class Imbalance Wafer Defect Pattern Recognition Based on Shared-database Decentralized Federated Learning Framework

Yong Zhang, Rukai Lan, Xianhe Li, Jingzhong Fang, Zuowei Ping, Weibo Liu and Zidong Wang

***Abstract*—In this paper, a novel shared-database decentralized federated learning framework (SDeceFL) is developed for wafer defect pattern recognition (DPR). Specifically, a differential privacy shared-database strategy is proposed to overcome the inter-class heterogeneity problem of different clients and enhance data privacy. A deformable convolutional auto-encoder (DCAE) is designed for data augmentation for handling class imbalance. The vision transformer (ViT) is employed for wafer DPR. The proposed DCAE-ViT-SDeceFL framework is validated on three public data sets (e.g., WM-811K, NEU-CLS-64, and CIFAR-100). Experimental results show the superiority of the SDeceFL framework over Ratio Loss-FedAvg, MOON, FedNH, BalanceFL, FedAvg, DeceFL and swarm learning. Compared with some deep learning methods, experimental results exhibit the effectiveness of the proposed DCAE-ViT-SDeceFL method for wafer DPR on WM-811K.**

***Index Terms*—Defect pattern recognition, class imbalance, deformable convolutional auto-encoder, vision transformer, decentralized federated learning, differential privacy.**

I. Introduction

With the breakthrough of the internet of things, the integrated circuit (IC) industry has developed rapidly in the past few years. As an important and fundamental material of the IC, the wafer plays a critical role in affecting the quality of the IC. It is known that Wafers are made from high-purity semiconductors through grinding, slicing and other extremely demanding processes which are complex and costly. During the wafer manufacturing process, wafers are easily damaged or influenced by the environment which cause defects. In recent years, many researchers have focused on improving the wafer manufacturing process as well as the manufacturing environment to fabricate qualified wafer products. Unfortunately, it is nearly impossible to produce non-defect wafers based on existing manufacturing processes.

This work was supported in part by the National Natural Science Foundation of China under Grants 62273264 and 61933007, the BRIEF funding of Brunel University London, the Royal Society of the UK, and the Alexander von Humboldt Foundation of Germany.

Y. Zhang, R. Lan and X. Li are with the Department of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, 430081, China. (emails: zhangyong77@wust.edu.cn, iakrulan@wust.edu.cn, 1210070395@qq.com)

J. Fang, W. Liu and Z. Wang are with the Department of Computer Science, Brunel University London, Uxbridge, Middlesex, UB8 3PH, United Kingdom. (emails: {Jingzhong.Fang, Weibo.Liu2, Zidong.Wang}@brunel.ac.uk) (*Corresponding author: Weibo Liu*)

Z. Ping is with the National Key Laboratory of Science and Technology on Vessel Integrated Power System, Naval University of Engineering, Wuhan, 430030, China. (email: pingzuowei@hust.edu.cn)

Serving as an important quality control technique in wafer manufacturing, a number of wafer pattern recognition (DPR) methods have been proposed to identify/analyze the defects of the wafers. By using DPR, the source of defects (e.g., materials, manufacturing processes, and equipment) can be properly identified. By doing so, wafer manufacturers are able to take actions in time to adjust the industrial process, which could reduce the manufacturing costs, improve wafer quality and thus strengthen the competitive position of manufacturers in the market [21], [47]. It is known that using a proper wafer DPR technology can reduce the wafer defect rate effectively and bring profits to the manufacturer.

With the increasing focus on information protection, in the past few decades, data privacy has become an important topic in wafer manufacturing. Federated learning can assist in the maintenance of the performance of machine learning algorithms while protecting the data privacy of local data. Federated learning has been extensively applied in data protection to enhance data transmission under the premise of data privacy and security [59]. For example, in [30], a new federated learning method has been developed for deep neural networks based on iterative model averaging. Recently, a transfer learning-based federated learning approach has been proposed for fault diagnosis in [59], where different models are utilized by different users in order to enhance data privacy. In addition, in [60], a federated transfer learning framework for machinery fault diagnosis has been introduced, where the prior distributions are employed for bridging the domain gap indirectly. As a popular federated learning method, decentralized federated learning has attracted various attention in the past few years [54]. In [54], a decentralized federated learning (DeceFL) framework has been developed to simulate the communication connections of clients. As a potential research direction of federated learning, the DeceFL framework uses topological graphs to simulate the communication connections, which protects the network from external network attacks and effectively reduces the impact of clients' communication disruption on the overall federated learning system. Thus, the DeceFL framework seems to be a suitable candidate for protecting sensitive data in wafer manufacturing.

Within federated learning, the non-independently and identically distributed (Non-IID) characteristic observed across heterogeneous clients results in substantial divergence when updating local clients, thereby presenting a fundamental challenge to the global model aggregation. A few studies have attempted to address the Non-IID problem in federated learn-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2024.3395316, IEEE Transactions on Instrumentation and Measurement

ing [8], [24], [43]. Such approaches focus on developing algorithmic-level strategies (such as class-balanced loss functions [24], [43], [56] and class-balanced training strategies [8]) to allow the global model to learn the data distribution characteristics of each client. For example, in [56], a fine-grained calibrated cross-entropy loss has been applied in local updating. In [8], the class semantics have been infused into class prototypes. In fact, data augmentation is a suitable candidate to solve the Non-IID problem in federated learning from the perspective of data processing. In this case, a seemingly natural idea is to put forward a new data augmentation strategy to handle the class imbalance problem in federated learning. In this paper, we not only develop an algorithmic-level approach to solve the heterogeneity problem in federated learning but also investigate the class imbalance problem at the level of data source management.

In the past, wafer DPR was carried out based on visual inspection by experienced engineers, which is expensive and time consuming. As an emerging topic in machine learning, deep learning has attracted an ever-increasing research interest due to its powerful feature extraction ability [41], [42], [49]. Recognizing as a powerful family of deep learning techniques, the convolutional neural network (CNN) has been widely used in wafer DPR due to its strong feature extraction ability [21]. For example, a CNN-based method has been introduced in [21] for DPR of wafers with mixed defects, where the CNN models are built based on each type of defect.

It should be noticed that the quality and quantity of training data are of vital importance in deep learning [10], [11], [19], [25]. Due to the uncertainty of the manufacturing site, the defect categories caused by the wafer manufacturing process are often imbalanced. Recently, data augmentation has been proven to be an effective way to solve the class imbalance problem. Some popular data augmentation methods include the generative adversarial network (GAN), variational auto-encoder, PixelCNN. In this situation, it seems natural to adopt the data augmentation technique to tackle the class imbalance problem in wafer DPR. Inspired by the above discussions, we aim to develop a new data augmentation method which is capable of extracting and learning the features with various scales and shapes adaptively.

Note that the convolution kernel plays a critical role in extracting features, which also affects the computational cost of the CNN. In recent years, various convolution kernels have been developed to improve the feature extraction ability of the CNN by expanding the receptive field [7], [51], [62]. An object detection method has been introduced in [62] based on the atrous convolution kernel with different expansion rates. Compared with the traditional convolution kernel, the dilated convolution kernel is able to adjust the dilation rate according to different data sets so as to modify the receptive field [51]. In [7], the deformable convolution (DC) has been proposed to adaptively learn the features of the object with various scales and shapes. Considering the characteristics of DC in expanding the receptive field, we propose a deformable convolutional auto-encoder (DCAE) module to reconstruct the raw data for DPR.

Motivated by the above discussions, this paper aims to develop a new data protection wafer DPR framework. Specifically, to tackle the privacy protection issue in the wafer DPR, a novel shared-database decentralized federated learning framework (SDeceFL) is proposed, where a differential privacy shared-database (DPS) strategy is put forward to overcome the inter-class heterogeneity problem of different clients and enhance data privacy. A DCAE module is proposed and is integrated with the vision transformer (ViT) for data augmentation in wafer DPR to solve the class imbalance problem, which introduces offsets to expand the receptive field by integrating the DC kernel in the traditional convolutional auto-encoder (CAE). The SDeceFL framework is embedded into DCAE-ViT for data transmission and global model development with the hope of alleviating the impact of fragmented data provided by wafer manufacturers. The main contributions of this paper can be summarized in the following three aspects:

(a) A new decentralized federated learning framework is put forward where a DPS strategy is developed to tackle the data privacy leaks and inter-class heterogeneity problems;
(b) The DCAE module is developed to expand the receptive field of the convolution and alleviate the class imbalance problem by generating new samples belonging to the minority class; and
(c) The proposed DCAE-ViT-SDeceFL framework is applied to analyze the public wafer image data for DPR with promising results, which could benefit wafer manufacturers by providing a reliable and efficient defect identification approach with guaranteed data privacy and security.

The remaining sections of this paper are organized as follows. The background of wafer manufacturing, wafer DPR, and federated learning are introduced in Section II. Then, a novel wafer DPR framework is introduced in Section III. In Section IV, the experiment setting, data description, and data pre-processing are discussed. Section V presents the experimental results, where the results for the proposed framework and some selected methods are compared and discussed in detail. Finally, conclusions and some possible future research directions are provided in Section VI.

II. Background

In this section, the general background of wafer manufacturing and some widely used wafer defect defection techniques are introduced. In addition, the wafer DPR techniques and data augmentation techniques for wafer DPR are presented.

A. Wafer Manufacturing

Wafer manufacturing includes various technological sectors (e.g., chemical mechanical polishing, exposure, post-exposure bake, and ion implantation). In general, wafers will be made into nano-chips after four core processes (i.e., photolithography, etching, deposition, and ion implantation).

Owing to the increasingly demanding requirements (e.g., low power consumption, good performance, and tiny scale) of wafers, a number of processes (such as evaporation deposition, dual-ion beam sputtering deposition, and plasma enhanced chemical vapor deposition) have been deployed in wafer manufacturing to overcome the problems. Nevertheless, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2024.3395316, IEEE Transactions on Instrumentation and Measurement

deployment of such processes may cause defects which are difficult to be detected based on previous experience. To tackle the defect detection challenges, a variety of wafer defect detection methods have been developed.

B. Wafer Defect Detection

In wafer manufacturing, uncertainties (e.g., complex environments and variations of processing parameters) may easily cause defects in the manufactured wafers. By analyzing the electron microscope images of the wafer, the defects exhibit different spatial patterns. It is known that there are nine types of defects in wafer manufacturing including center, donut, edge-loc, edge-ring, local, near-full, random, and scratch [47]. Specifically, the anomaly of film deposition leads to the edge-loc defect, and the anomaly of etching leads to the edge-ring defect. Uneven cleaning would cause the local defects, and human mistakes result in the near-full defects in wafers. To reduce the defective rate of chips, a number of wafer defect detection methods (including both model-based methods and data-driven methods) have been developed to monitor the production process in real-time by analyzing the electron microscope images of wafers.

One of the most well-known model-based wafer defect detection methods is template matching. In [2], the sequential similarity detection algorithm (which is recognized as one of the best matching criteria) has been proposed for accurate defect detection. Nevertheless, the template matching methods may suffer from the uncertainty, especially the randomness of template selection, which is time-consuming for real-world deployment.

Recently, data-driven methods have been widely used to assess the quality of wafers [39]. Among existing data-driven methods, image processing methods have been successfully adopted in detecting possible defects by analyzing the defect of images. In [39], a spatial attention bilinear CNN has been proposed to classify defective castings and non-defective ones. It is known that many existing defect detection methods can detect wafer defects effectively [40]. Nevertheless, considering the potential problems (e.g., lacking wafer defect data, high labeling cost, unsatisfactory model), there is a need to develop some advanced wafer DPR methods in wafer manufacturing.

C. Wafer Defect Pattern Recognition

In wafer defect detection, pattern recognition plays a critical role in extracting and analyzing the features of various wafer defects. The wafer DPR indicates the accurate recognition of wafer defect patterns, which aims to identify the anomalies through the manufacturing process. In recent years, the CNN has become one of the most preferred models in the field of DPR [21].

The feature learning performance of the CNN is highly dependent on the features extracted through the continuous accumulation of convolutional layers. A number of CNNs with specifically designed network architectures have been proposed (e.g., the AlexNet, the GoogLeNet, the VGGNet, and the ResNet [13]). With the development of the attention mechanisms, the transformer has become the state-of-the-art deep learning architecture for natural language processing and computer vision [9], [23]. In [9], the vision transformer (ViT) has been proposed where a standard transformer is directly applied to handle the sequences of image patches for image processing, which shows competitive or even superior performance against the CNNs.

D. Federated Learning

Owing to the increasing importance of data security and privacy, a variety of data protection mechanisms have been designed by many wafer manufacturing enterprises. In this situation, the wafer data is fragmented into several *data islands*, which not only enhances user privacy but also causes the data isolation problem. The *data islands* across enterprises makes it difficult to establish big data mechanisms for data sharing. In other words, it is nearly impossible to build an effective machine learning model only based on the fragmented data.

To tackle the aforementioned problem, it seems reasonable to apply the federated learning technique to balance the data privacy/security and communication cost [54]. Different from centralized learning where the whole data set is trained on a single node, a generalized model is established on distributed devices collaboratively via federated learning, which makes it feasible to solve the *data islands* problem. Generally, federated learning techniques are capable of decreasing communication costs and ensuring data privacy during the data transmission/transfer by training and aggregating local models into a *global* model.

Aggregating local models into a global model is a critical issue in federated learning. In recent years, a large number of federated learning methods have been put forward [30], [59]. For example, the federated averaging (FedAvg) algorithm has been proposed in [30] where the centralized federated learning is deployed to enhance the generalization ability of the model and alleviate the overfitting problem. In [59], a federated transfer learning algorithm has been proposed for fault diagnosis, where a federal initialization stage is introduced to keep similar data structures during the distributed feature extraction stage, and a federated communication stage is further implemented using deep adversarial learning. Unfortunately, the centralized federated learning-based algorithms would face a high communication burden and vulnerability once the central client fails or is affected by a cyber-attack. Very recently, in [54], a principled DeceFL is proposed, which relies only on local information transmission between clients and their neighbors instead of using a central client for sharing all the acquired data. In wafer manufacturing, data security and privacy are of critical importance for enterprises. Based on the above discussions, the federated learning techniques could tackle the data sharing problem for solving the *data islands* in wafer DPR.

Federated learning faces a major challenge due to the class imbalance in the training data of each client, leading to a notable impact on the performance of model learning. In the past few years, numerous studies have been conducted in order to improve the performance of federated learning on Non-IID data. In [43], the ratio loss has been introduced

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2024.3395316, IEEE Transactions on Instrumentation and Measurement

in the FedAvg framework to mitigate the effect of the class imbalance problem. In [24], the model-contrastive loss has been designed and added to the local model in the FedAvg framework to resolve the heterogeneity between clients (Non-IID). In [8], a federated learning framework, the FedNH, has been proposed, which uniformly distributes class prototypes in the latent space and then infuses the semantics into prototypes smoothly. In [37], a long-tail federated learning framework named BalaceFL has been proposed, which can robustly learn both common and rare classes.

E. Data Augmentation

Data augmentation is one of the popular avenues for handling class imbalance in federated learning. As a popular family of data augmentation methods, data generation methods are widely employed in both academia and industry, which can be classified into statistical-based methods, probabilistic-based methods, and deep learning-based methods [33].

During the past few years, a number of probabilistic image generation methods have been proposed [32]. In [32], several probability-based models have been put forward to generate different defect patterns (e.g., annular defect patterns, mixed defect patterns, repetitive defect patterns, and random defect patterns). It should be noticed that the performance of the probabilistic-based methods and statistical-based methods is heavily dependent on expert knowledge. For some high-dimensional and irregularly shaped defects, it would be extremely difficult for experts to recognize defects only based on empirical experience, thereby influencing the quality of the data generation process.

With the rapid development of hardware equipment, a variety of deep learning-based methods have been put forward for data generation. For example, in [26], a generative adversarial network has been put forward to obtain high-quality thermal images based on data augmentation.

As a well-known branch of deep learning, the auto-encoder (AE) has been widely used in unsupervised learning for feature extraction, which demonstrates competitive or better feature extraction performance than the probabilistic-based methods [3]. The standard AE contains an encoder and a decoder, where the encoder is used to sample and learn the original features, and the decoder is employed to decode the features so as to reconstruct the original input. In the past few years, many AE-based methods have been developed for representation learning. For instance, the CAE has been presented in [29] which combines the convolution and pooling operations of the CNN with AE for hierarchical feature extraction. The CAE has been extensively used in noise reduction and data reconstruction. Nevertheless, the *limited* receptive field of the traditional convolution operator becomes a bottleneck for the development of the CAE.

III. Methodology

In this paper, a novel DCAE-ViT-SDeceFL framework is developed for wafer DPR where 1) a novel SDeceFL framework with DPS strategy is proposed; 2) a DCAE-based data augmentation approach is designed to solve the class imbalance problem; 3) the ViT is utilized to extract the defect features by using the multi-head self-attention (MSA) mechanism for DPR. The overall scheme of the developed wafer DPR framework is shown in Fig. 1.

A. The DCAE-based Data Augmentation Approach

It is known that the convolution kernel is capable of studying the geometric changes of objects by using sliding windows and scale-invariant feature transformations. The kernels are employed to extract the features of the input data with different shapes. With a certain number of convolution kernels, stacking the convolutional layers could expand the receptive field. Specifically, the receptive field expands linearly as the number of convolutional layers increases. Nevertheless, the corresponding computational cost of the convolution process increases exponentially. Related studies have shown that the performance of the CNNs is constrained by the receptive field to some extent [51].

Compared with the traditional convolution kernel, the DC has the offset field which could enlarge the receptive field and improve the sparse spatial sampling capability. The DC learns from offsets by an additional convolution kernel with the same size as the input feature map x, where the number of channels is $2N$ corresponding to N two-dimensional offsets. After that, the input feature map x and offsets are jointly used as the input of the next layer.

By introducing the offsets to the traditional convolution kernel, DC is designed to expand the receptive field. For the ith pixel location of image p_i on the output feature map y_{out} of the DC is calculated as follows:

$$y_{out}\left(p_i\right)=\sum_{p_n\in R} w\left(p_n\right)\cdot x\left(p_i+p_n+\triangle p_n\right) \tag{1}$$

where R denotes the convolution kernel; $w\left(\cdot\right)$ is a function to calculate the weights and biases; p_n represents the nth position in the convolution kernel; $\{\triangle p_n|n=1,...,|R|\}$ denotes the offset of the nth position in the convolution kernel R; and x represents the input feature map. Note that the size of the convolution kernel R is the size of the receptive field.

It is worth mentioning that the bilinear interpolation method is used to solve the non-integer offset $\triangle p_n$ problem in DC. The interpolated coordinates can be expressed as follows:

$$X\left(\beta\right)=\sum_{\alpha} G\left(\alpha,\beta\right)\cdot X\left(\alpha\right) \tag{2}$$

where β denotes an arbitrary fractional location ($\beta=p_0+p_n+\triangle p_n$); $X\left(\cdot\right)$ represents the feature map after convolution; α is the enumeration of all integral spatial locations in X; $G(\cdot,\cdot)$ is the two-dimensional bilinear interpolation kernel, which can be expressed as follows:

$$G\left(\alpha,\beta\right)=g\left(\alpha_x,\beta_x\right)\cdot g\left(\alpha_y,\beta_y\right) \tag{3}$$

where $g\left(a,b\right)=\max\left(0,1-|a-b|\right)$.

In this paper, the DCAE is proposed where the DC kernel is integrated with the traditional CAE to further improve the receptive field. The developed DCAE is a sparse spatial

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2024.3395316, IEEE Transactions on Instrumentation and Measurement

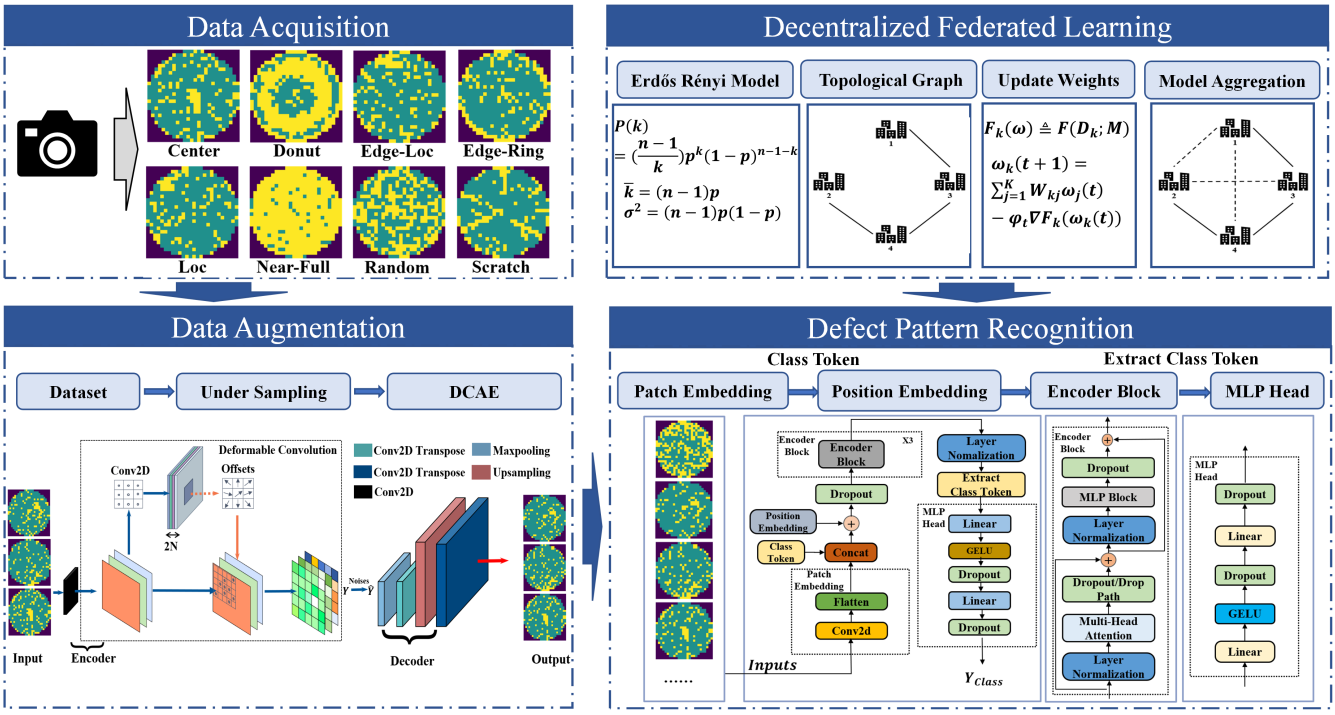


Fig. 1. Flowchart of the proposed wafer DPR framework

sampling technique, which could compress the structure information of image neighboring locations into a fixed grid, resulting in better feature capturing performance compared with the traditional convolution. Based on the feature capturing capability and the self-learning capability of the DCAE, the class imbalance wafer images are reconstructed and generated. The loss function of the proposed DCAE is given below:

$$L_{DCAE} = \frac{\sum_{i=1}^{n}\left(y_i - y_i^{'}\right)^2}{n} \tag{4}$$

where y represents the actual label; $y_i^{'}$ represents the predicted label; and n represents the number of categories.

1) Structure of the DCAE: The data generation process is summarized into four steps:

(a) Input the real samples into the DCAE;
(b) Encode the input to the intermediate variable Y;
(c) Add noises into Y as $\hat{Y}$;
(d) Decode the integrated variable $\hat{Y}$ for generating new samples.

It is worth noting that the integration of random noises into input images would lead to the generation of different samples. In this paper, more samples are obtained for the minority class by employing the DCAE so as to alleviate the class imbalance problem.

In the proposed DCAE module, the encoder includes the DC layer and the pooling layer, which is used to capture the core feature and map the image to a high-dimensional space. The decoder is made up of a transpose convolutional layer and an upsampling layer for decoding the high-dimensional feature map. In addition, the "MAE" loss function is used in the proposed DCAE module for parameter optimization.

It should be noted that the deep neural network with a large convolution kernel (DNNLCK) may include a large number of training parameters, which requires high computational cost [38]. Compared with the DNNLCK, the proposed method could effectively reduce the model parameters and extract the defect feature at different scales by employing DC. A single convolutional layer may result in a simple model structure that is unable to capture the underlying patterns in the training data, leading to underfitting. In contrast, DCAE, compared to CAE under the same conditions, can leverage DC layers for more flexible receptive field control, capturing defect features at varying scales and improving model performance. To further enhance the model's feature representation capability and alleviate underfitting problems, we can increase the channel numbers of DC kernels to improve the expressive power of the convolutional layers.

B. The ViT-based Defect Pattern Recognition

In wafer defect classification, the number of defect samples in certain categories may be very small due to the fact that there may be a significant imbalance in the number of defect samples in different categories. In this case, the trained model would perform poorly for minority classes at the defect classification stage, which results in classification bias. ViT is a defect classification model that utilizes the self-attention (SA) mechanism and the MSA mechanism.

Specifically, the augmented data obtained from the DCAE are fed into the ViT. During the "Patch Embedding" process, the images are chunked into blocks with the same size. Then, each block is flattened into a vector. The positional information is concatenated to the flattened vectors during the "Position Embedding" process. After the "Patch Embedding" process

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2024.3395316, IEEE Transactions on Instrumentation and Measurement

and "Position Embedding" process, a special character class token is concatenated to each vector. Subsequently, the features of the vectors are extracted from the "Encoder Blocks" and performed global pooling. Finally, the obtained features are fed into the "MLP Head" for image classification where the MLP is the multilayer perceptron.

It is known that the standard transformer is designed for text-based tasks. In order to process the two-dimensional image data, the patch embedding module is designed in the ViT to flatten the two-dimensional image data to one-dimensional text-like data.

The encoder of the ViT proposed in [9] can be treated as a feature extraction module, which consists of an MSA mechanism and an MLP block. The MSA mechanism is an extension of the SA mechanism. The MSA mechanism can be described as follows:

$$Attention\left(Q,K,V\right)=softmax\left(QK^{\mathsf{T}}/\sqrt{d}\right)V \quad (5)$$

where Q, K, V represent the query vector, key vector, and value vector, respectively; $d = D/k$ guarantees that the number of computations and parameters remains unchanged when the number of heads k is changed; D represents the dimensionality of patch embedding; and $softmax(\cdot)$ represents the sum of normalized probabilities of 1.

The image classification process of the ViT [9] can be summarized as follows:

$$z_0=\left[M_{class};M_p^1E;M_p^2E;\ldots;M_p^NE\right]+E_{pos} \quad (6)$$

$$z_l^{'}=\mathrm{MSA}\left(\mathrm{LN}\left(z_{l-1}\right)\right)+z_{l-1} \quad (7)$$

$$z_l=\mathrm{MLP}\left(\mathrm{LN}\left(z_l^{'}\right)\right)+z_l^{'} \quad (8)$$

$$y=\mathrm{LN}\left(z_L^0\right) \quad (9)$$

where $M_p \in \mathbb{R}^{N\times(P^2\cdot C)}$ denotes sequences of image patches; P represents the height and weight of each image patch; $N=(H\times W)/P^2$ denotes the resulting number of patches; $E \in \mathbb{R}^{\left(P^2\cdot C\right)\times D}$ denotes the fully connected layer; $E_{pos} \in \mathbb{R}^{(N+1)\times D}$ represents the position embedding; $z_l^{'}$ denotes the output feature map of the lth encoder block; l denotes the number of encoder block $(l=1,\ldots,L)$; $\mathrm{MSA}\left(\cdot\right)$ represents the calculation process of the MSA mechanism; $\mathrm{MLP}\left(\cdot\right)$ is the multi-layer perceptron (MLP) block; $\mathrm{LN}\left(\cdot\right)$ denotes the layer norm (LN); z_L^0 denotes the feature map of class token; and y is the output of the ViT algorithm. Note that in ViT, the learnable embedded class tokens y need to pass through an MLP layer with the number of classes as the dimension, followed by a softmax activation function, in order to obtain the probability information for each defect category. The category prediction label is then determined by selecting the class with the highest probability.

It is worth mentioning that the original image $M \in \mathbb{R}^{W\times H\times C}$ (whose height is H, weight is W, and number of channels is C) is reshaped into a sequence of flattened 2D patches $M_p \in \mathbb{R}^{N\times(P^2\cdot C)}$. The ViT introduces the learnable position embedding to record the position and update the information E_{pos} in order to make up for the loss of the image position information during reshaping the images. In addition, the ViT introduces the class token to record the classification information M_{class} of the image. The input dimension $\left(P^2\cdot C\right)$ is embedded into the D dimension by a trainable linear projection, which is called patch embedding. The patch embedding with dimension (N,D) is the input of the encoder block. After the MSA combined with linear layers to capture important feature information, the final output is a feature map with a size of $(N+1,D)$, where the MLP consists of fully connected layers and two dropout layers with the Gaussian Error Linear Units (GELU) activation function. Compared with the ReLU and ELU activation functions, GELU exhibits better smoothness, continuity, and convergence rate. GELU can effectively adapt to the distribution of various image features. Therefore, GELU is selected as the activation function of ViT [34].

The loss function of the ViT is computed by sparse categorical cross-entropy, which is described as follows:

$$L_{ViT}=-\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N}y_j^i\log\left(\hat{y}_j^i\right) \quad (10)$$

where $\hat{y}_j^i$ is the predicted probability that sample i belongs to j, and y_j^i is a sign function that takes 1 when the true class of i is equal to j and 0 otherwise; N is the number of classes; and M is the number of samples.

Remark 1: LN and BatchNorm (BN) are the two most commonly used data normalization methods [50]. Compared with LN, BN will mix all samples in a batch together when calculating the mean and variance, which may destroy the correlation relationship of the data. During the training process of the ViT model, the input images are divided into multiple patches through a process called Patchification. LN, which performs normalization along the feature dimension, is suitable for handling variable-length data like patches. LN helps maintain the relative order between positions and ensures the stability and reliability of the normalization process. By applying LN in the feature dimension, the ViT model can effectively handle the patches and capture meaningful representations from the input data. This enables the model to maintain the positional information and achieve stable and reliable normalization throughout the training process. In this paper, in order to retain the time relationship of data during normalization, LN is employed as the normalization method.

C. The SDeceFL Framework

Considering the aforementioned weaknesses of the centralized federated learning, the decentralized federated learning framework has been proposed in [54] for decentralized training and parameter aggregation (based on the local information transferring between clients and their neighbors). Here, the clients identify their "neighbors" through a time-invariant/time-varying topology graph. The SDeceFL framework reduces the burden of the central server caused by excessive client communication. The topology graph of the different federated learning and swarm learning (SL) frameworks is shown in Fig. 2.

By using the Erdos Renyi method, we randomly generate an undirected Erdos Renyi connectivity graph with n nodes and the corresponding connection probability p [54]. Whether

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2024.3395316, IEEE Transactions on Instrumentation and Measurement

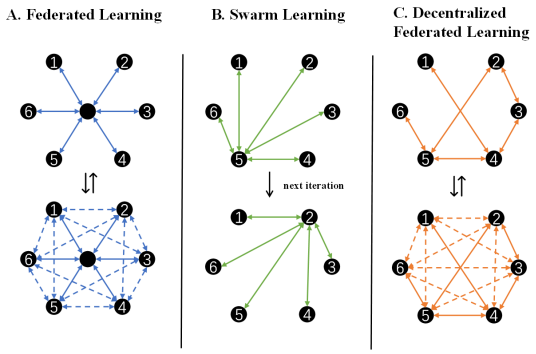


Fig. 2. The topology graph of the different federated learning and swarm learning framework

an edge exists between each pair of nodes in the graph is determined independently according to a specified probability. In this case, the SDeceFL framework can properly simulate the complex communication situations between clients in reality. In theory, all edges between nodes should communicate with each other as the number of generations increases. Unfortunately, due to some special reasons (e.g., business competition or network outage), the client can only choose to communicate with other clients in a few rounds of training. Compared with centralized federated learning, the random local communication between clients can effectively simulate complex data sharing in reality and alleviate the single node failure when training a global model.

By assigning different connection probabilities to the clients, the fragmented data sharing in the realistic scenario is simulated, where data sharing is no longer dependent on the central server. In federated learning, the connected clients need to build local models and then establish the global model. The global model aggregation and parameter updating of the SDeceFL framework can be summarized as follows:

$$\omega_k\left(t+1\right)=\sum_{j=1}^{K}W_{kj}\omega_j\left(t\right)-\varphi_t\nabla F_k\left(\omega_k\left(t\right)\right) \tag{11}$$

K indicates the number of SDeceFL clients; k is a client; t represents the t-th iteration; $\omega_k(t)$ is the estimated global optimum for the k-th client at t-th iteration; φ_t is the learning rate; $\nabla F_k\left(\omega_k\left(t\right)\right)$ represents the gradient calculated by the local client k in the t-th iteration; and W_{ij} indicates the connection between the clients i and j. To be specific, the information transmission between clients i and j occurs when $W_{ij}>0$. $W_{ij}=0$ indicates no information transmission between i and j. When $W_{ij}>0$, the client i is referred to as a neighboring client of client j. The set of all such clients j is denoted as $N_i=\{j|W_{ij}>0,\forall j\in N\}$.

In the SDeceFL framework, each federated client is updated locally, in which the weights are updated according to the undirected topology graph. In detail, the Erdos Renyi method is used in this paper to generate the topology graph G_{np} with n nodes and the edge connectivity probability p, and finally return the adjacency matrix. Then, the established global model is sent to each federated client to initialize the weights of each client.

The local loss function $F_k\left(\omega\right)\triangleq F\left(D_k;M\right)$ is the user-specified loss function on the data sets D_k where the model parameters are defined by ω in the model M. The $F\left(D;M\right)$ can be reformulated by $F\left(\omega\right)\triangleq\frac{1}{K}\sum_{k=1}^{K}F_k\left(\omega\right)$ where K denotes the number of connected clients, and $F\left(\omega\right)$ represents the average loss function on all data. The global loss function could be designed in a private and centralized manner from the global perspective. In this case, the global loss function is not given to any local client. A local loss function is then designed and distributed to the local client, which is part of the global loss function. In such cases, data are locally stored with guaranteed security, and the global loss function is protected even for the clients.

Based on the DeceFL, the SDeceFL framework is put forward where a DPS strategy is proposed for tackling the inter-class heterogeneity of different clients. The class imbalance problem in a single client is effectively tackled by data augmentation in the DCAE module, but there may exist label heterogeneity problems among different clients. The DPS strategy effectively deals with label heterogeneity by 1) constructing a shared database that contains the label distributions of the clients trained by the two participating clients; and 2) utilizing the obtained shared database to train the global model, which acquires prior knowledge of the label distributions for model training. The main procedure of the DPS strategy is divided into four steps:

(a) Randomly select a certain percentage of samples from each client to build a public database;
(b) Add Laplace noises into the obtained database known as the shared-database;
(c) Train the global model using the shared-database;
(d) Distribute the weights of the trained global model to the selected clients for local training.

It is worth pointing out that the Laplace noises are introduced into the sampled client data to achieve differential privacy, which ensures the non-disclosure of client data information within the public database. The shared-database is employed to train the global model whose weights are distributed to the local clients participating in data distribution. By doing so, the local client is aware of part of the data distribution information of its neighboring clients as priori knowledge, which would benefit the further local client training. The schematic diagram of the proposed DPS strategy for two local clients is shown in Fig. 3.

Each federated client runs the training algorithm locally, and the estimation of global parameters is transferred to its neighbors. The federated client calculates the average of the neighbors' weights/gradients and generates the aggregated weights in the next iteration when receiving additional weights/gradients from the neighbors. In the SDeceFL framework, each federated client completes the updating process when receiving/sending local weights to neighbors instead of aggregating and transmitting to the third-party central clients.

The procedure of the SDeceFL algorithm is provided in Algorithm 1.

It should be noted that the random variable $X_{i,j}$ of each pair of distinct nodes (i,j) is generated according to $Bernoulli\left(p\right)$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2024.3395316, IEEE Transactions on Instrumentation and Measurement

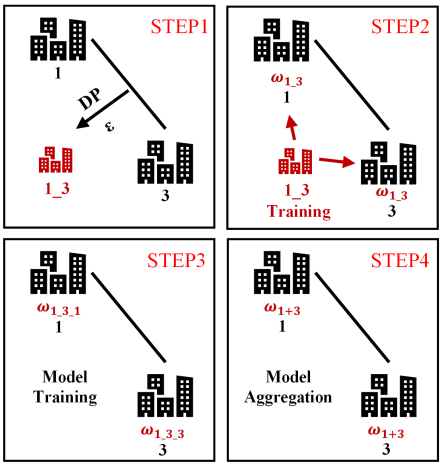


Fig. 3. Learning processes for public databases under differential privacy

probability distribution, which is shown as follows:

$$X_{i,j} = \begin{cases} 1 & \text{Node } i \text{ and } j \text{ are connected} \\ & \text{with probability } p \\ 0 & \text{Node } i \text{ and } j \text{ are not connected} \\ & \text{with probability } 1-p \end{cases} \tag{12}$$

where i and j are the node numbers; and p represents the edge connectivity probability. The normalized Laplace adjacency matrix W is updated and calculated based on the random variable $X_{i,j}$, which is demonstrated as follows:

$$w_{i,j} = \begin{cases} 1 & if \quad X_{i,j} = 1 \\ 0 & if \quad X_{i,j} = 0 \end{cases} \tag{13}$$

IV. Experiments and Analysis

In this paper, the proposed wafer DPR framework is applied to the public data sets, WM-811K. In this section, the description of the data sets and the settings of the related experimental parameters are first introduced. After that, the proposed DCAE data augmentation method is evaluated and compared with some existing data augmentation methods. Then, DCAE-ViT model is compared with some selected state-of-the-art defect classification methods. In order to verify the effectiveness of the proposed method under decentralized federated training, the centralized federated learning method (i.e., the FedAvg algorithm), the SL method and the federated learning framework that has been used to deal with class imbalance in recent years (i.e., Ratio Loss-FedAvg [43], MOON [24], FedNH [8] and BalanceFL [37]) are selected and compared. Ablation experiments are also conducted to verify the effectiveness of the proposed DCAE-ViT-SDeceFL framework.

A. Data Description

The WM-811K public data sets, which is the most widely used data sets, is adopted in this paper for wafer DPR [47]. The data sets contain information of the defect category, the production lot, the chip size and the image pixel. There are 811457 wafer images in the data sets, which are collected by

Algorithm 1: The Main Steps of the SDeceFL Framework

Input : The number of clients participating in federated training K;
each clients N_i, $(i = 1, 2, \ldots, K)$;
the data sets of the i-th federal client D_i, $(i = 1, 2, \ldots, K)$;
the client learning rate L_i;
the learning rate of public databases l;
the iteration number T;
the adjacency matrix W;
the collection ratio ε.

Output: Model parameters of client i at round n: M_n^i.

for $t \in [1, T]$ **do**
  for *each client* N_i *parally* **do**
    the clients expand and balance local wafer samples by training DCAE models with local data sets D_i
    construct a matrix W for $n \times n$, where each element is either 0 or 1 (1 means there is an edge between node i and node j, and 0 means there is no edge)
    for each pair of distinct nodes (i, j), generate a random variable $X_{i,j}$
    update and calculate the normalized Laplace adjacency matrix W according to the random variable $X_{i,j}$
    build a public database by selecting ε percentage of local data D_i from each local client N_i, and train the ViT using the acquired database with the learning rate set to l
    distribute the trained weights to the local client N_i involved in data sharing in previous step
    train the ViT model of each federated client N_i by using the local data D_i with the learning rate set to L_i
  end
  update each client's weight according to Eq. (11)
  get the global model M_n^i by global model aggregation of each client N_i
end

experts through the industrial production process. The wafer images can be divided into 9 classes based on different defect patterns which are Non-pattern, Center, Donut, Edge-local, Edge-ring, Local, Near-full, Random and Scratch. In the WM-811K data sets, 18.2% of the wafer images are Non-pattern, 3.1% of the wafer images have actual defects, and 78.7% of the wafer images are unlabelled. In order to demonstrate the generalizability of the proposed DCAE-ViT-SDeceFL framework, the NEU-CLS-64 public data sets [15] and the CIFAR-100 public data sets are also employed for performance evaluation. The NEU-CLS-64 public data sets assemble approximately 7000 tiny images with 9 defect classes, i.e., crazing (Cr), grooves and gouges (GG), inclusion (In), patches (Pa), pitted surface (PS), rolling dust (RD), rolled-in scale (RS), scratches

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2024.3395316, IEEE Transactions on Instrumentation and Measurement

(Sc), and spots (Sp). The CIFAR-100 public data sets assemble approximately 60000 images with 100 defect classes.

B. Data Pre-Processing and Data Augmentation

In the experiment, 14312 labeled wafer images from the WM-811K public data sets are selected randomly and formed the experimental data sets. In the experimental data sets, 93.87% of the wafer images (13436 images) are Non-pattern, and 6.13% of the wafer images have defects. It should be noticed that there is an obvious class imbalance problem in the experimental data sets. In this case, the Non-pattern wafer images are under-sampled from 13436 to 436. The wafer images with defects are reconstructed by the proposed DCAE for data augmentation. Specifically, the wafer images with defects which belong to Center, Donut, Edge-local, Edge-ring, Local, Near-full, Random and Scratch are 630, 508, 888, 558, 891, 528, 592, 639, respectively. Similarly, the NEU-CLS-64 public data sets with steel plate defects are reconstructed by the proposed DCAE for data augmentation, the steel plate images with defects which belong to SP, Sc, RD, PS, Pa, In, GG and Cr.

C. Data Processing

After data pre-processing and data augmentation, the ViT is employed for the wafer defect classification. The wafer images with the size of $96 \times 96 \times 3$ are first split into 256 patches with the size of 6×6, Then, the patches are linearly embedded and added with the position embeddings whose size are $256 \times 6 \times 6$. Similarly, we perform the same operation after resizing steel plate data from 64×64 to 96×96. It is worth mentioning that an extra learnable class token is also added in order to store the information of classification. The resulting sequence of vectors are then fed into the 3-layer encoder which consists of alternating layers of MSA and MLP blocks. In the resulting sequence, each vector incorporates information of other vectors. In this case, the MSA blocks operate as a feature extractor of the entire sequence where multiple attention heads are applied to different positions in the input sequence, therefore, ViT can focus on multiple positions of the sequence at the same time. in order to extract useful features and reduce irrelevant noises. The number of the heads in MSA is set to be 4. The classification information in the class token is extracted by the dense layer of the MLP blocks and is treated as the defect classification results.

D. SDeceFL-Based Model Training

In the experiment, the SDeceFL framework is utilized to train the model. The unprocessed data is divided into four groups and allocated to four clients randomly. Next, four clients use local data to train the DCAE model for data augmentation. The SDeceFL framework is utilized to train the model. The data set is divided into four groups and allocated to four clients randomly. The connections between clients in each iteration are simulated by a time-varying undirected topology graph $\mathcal{G}(t) = (N, \varepsilon, W)$, where $N = \{1, 2, 3, 4\}$ denotes the clients; ε is the time-varying boundaries; and W represents the corresponding adjacency matrix. Two random nodes are connected with the probability $p = 0.5$. The information transmission between the chosen two nodes is determined by the adjacency matrix W^c, which is shown as follows:

$$W^c = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}. \tag{14}$$

In each iteration, four nodes are divided into two groups randomly, which are the communicating nodes and unconnected nodes. The unconnected nodes don't participate in the data communication of this iteration, but they still train and update the local models. The weight matrix of two unconnected nodes W^{nc} is shown as follows:

$$W^{nc} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{15}$$

In each round of federated communication, half of the nodes are randomly selected from all nodes to participate in communication as a time-varying setting, and the corresponding adjacency matrix W^c is randomly generated by using the Erdos Renyi method. W^{nc} is used to describe the communication situation of the remaining half nodes which do not participate in communication. By sequentially drawing the update of the communication matrix in the simulation experiment, it can be found that only two nodes communicate with each other in each round. With the increase in global training rounds, all nodes will complete the mutual communication. The time-varying undirected topology for four nodes is updated according to the following steps, which are shown in Fig. 4.

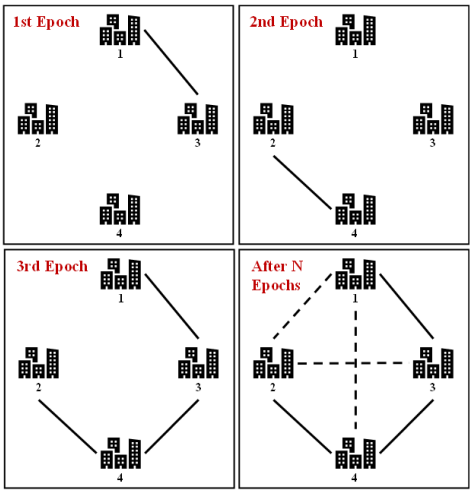


Fig. 4. The time-varying topology graph of the SDeceFL framework

Step 1 Node 2 and node 3 communicate. Node 1 and node 4 train locally. The corresponding weight matrix is described as follows:

$$W_1^{all} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{16}$$

Step 2 Node 1 and node 2 communicate. Node 3 and node 4 update locally. The corresponding weight matrix is described as follows:

$$W_2^{all} = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{17}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2024.3395316, IEEE Transactions on Instrumentation and Measurement

Step 3 Node 1 and node 3 communicate. Node 2 and node 4 update locally. The corresponding weight matrix is described as follows:

$$W_3^{all} = \begin{bmatrix} 0.5 & 0 & 0.5 & 0 \\ 0 & 1 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{18}$$

Step 4 Node 1 and node 4 communicate. Node 2 and node 3 update locally. The corresponding weight matrix is described as follows:

$$W_4^{all} = \begin{bmatrix} 0.5 & 0 & 0 & 0.5 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.5 & 0 & 0 & 0.5 \end{bmatrix}. \tag{19}$$

The steps are repeated until the maximum iteration number is reached or the model converges, which aims to obtain a server-like data model interaction in order to guarantee the information transmission between each group of nodes.

E. Experiment Setup

In order to achieve a comprehensive evaluation of the proposed DCAE-ViT-SDeceFL framework, two experiments are conducted: a comparison study and an ablation study. The details of the experimental platform are CUDA 11.4 and GTX 3080Ti GPU which has data parallelism, and is implemented with TensorFlow 1.5.0.

1) Comparison Study: The comparison study in this paper can be divided into data augmentation experiment, classification experiment and federated learning experiment.

(a) Data Augmentation Experiment: Resize the wafer image to the size of $26 \times 26 \times 3$ and feed it into the DCAE, where we set the kernel size in layer 1 and layer 2 of both encoders to 3×3, and the number of channels of both layer 1 and layer 2 are set to be 32. The Adam algorithm is chosen as the optimizer of the proposed DCAE. The batch size and the learning rate are set to be 32 and 0.001, respectively.

(b) Defect Classification Experiment: The input data of the ViT are the $96 \times 96 \times 3$ wafer images and the $64 \times 64 \times 3$ steel plate defect images. The stochastic gradient descent (SGD) optimizer with momentum is employed, where the momentum, the learning rate and the batch size are set as 0.9, 0.0001 and 32, respectively. When constructing the public database, the data collection ratio is set as 0.05, and the distribution of the noise is chosen as Laplace. The learning rate of the ViT is set by 0.00005, and the other parameters remain unchanged.

(c) Federated Learning Experiment: The Adam optimizer is employed in this experiment, where the momentum, the initial learning rate, and the batch size are set as 0.9, 0.001, and 32, respectively. The number of clients (i.e., num-users) is set to be 4.

2) Ablation Study: In order to verify the effectiveness of the proposed method, an ablation study is conducted on three public data sets in which each modification rule is implemented separately. We use ResNet50-DeceFL without DCAE data augmentation as the baseline model. Experimental settings (e.g., optimizer, data sets, and models) are consistent with the aforementioned comparison study.

V. Experiment Results and Discussions

A. Results and Discussions of DCAE

In this paper, two performance indicators (i.e., Entropy and Tenengrad) are employed to evaluate the quality of the reconstructed images. The proposed method is compared with some well-known data augmentation methods including the Wasserstein GAN (WGAN) [1], CAE [53], NL-CAE [45], GC-CAE [4], ADC-GAN [17], ReAC-GAN [20] and Simple Diffusion [16].

In the experiment, 2000 images generated by the selected data augmentation methods are randomly selected to compute the average value of the Entropy and Tenengrad. The assessment indicators of generated images using the proposed method and some selected methods are listed in Table I. As shown in Table I, the Entropy scores of the images generated by our proposed method are 6.54 and 6.97 on the WM-811K public data sets and the NEU-CLS-64 public data sets, respectively. The Tenengrad scores of the images generated by our proposed method are 10.38 and 6.14 on the WM-811K public data sets as well as the NEU-CLS-64 public data sets, respectively. The model size of the DCAE is also compared with the other methods in Table I. It can be found that the model size of DCAE is only 0.11 MB.

The DCAE-generated images and original images are depicted in Fig. 5. It can be found from Fig. 5 that the generated samples of different classes successfully exhibit the wafer defect characteristics of the current class. In this case, we can conclude that the quality of the images generated by the DCAE is better than that of the compared methods, which demonstrates the effectiveness of the proposed DCAE in the wafer image generation task.

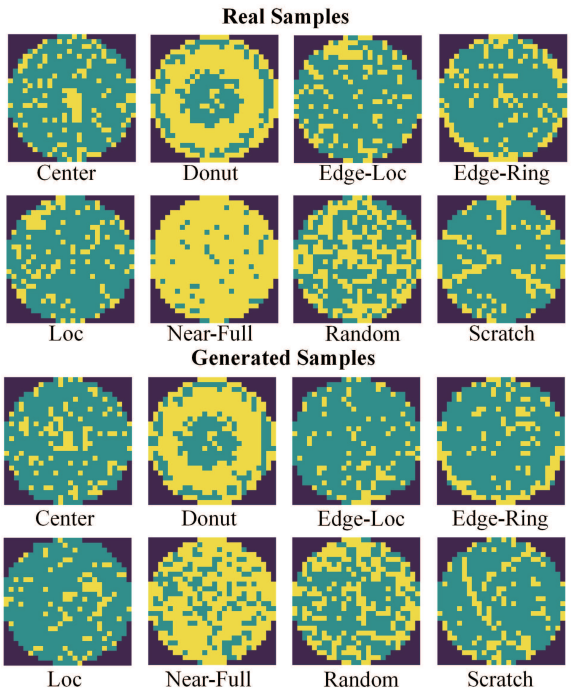


Fig. 5. DCAE generated images and original images

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2024.3395316, IEEE Transactions on Instrumentation and Measurement

The DCAE-based data augmentation results are shown in Fig. 6 and Fig. 7 for two different data sources. Fig. 6 displays the distribution of the raw data (WM-811K public data sets) and the data augmented by using the DCAE. In our work, a certain number of defect categories (e.g., Donut, Edge-Ring, and Near-Full) are specifically augmented to tackle the class imbalance problem. The distribution of the raw steel plate data (NEU-CLS-64 public data sets) and DCAE-augmented data is presented in Fig. 7. The number of samples in certain categories (including RD and GG) is augmented. It can be seen in Fig. 6 and Fig. 7 that the pre-processed data distribution of wafer and steel plate becomes uniform based on DCAE-based data augmentation, thus demonstrating the effectiveness of the DCAE module. The augmented data sets are then divided into the training set and the testing set with a ratio of 4:1.

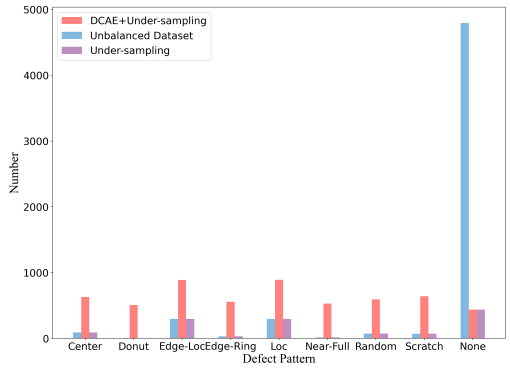


Fig. 6. Distribution of raw data and augmented data by using the DCAE

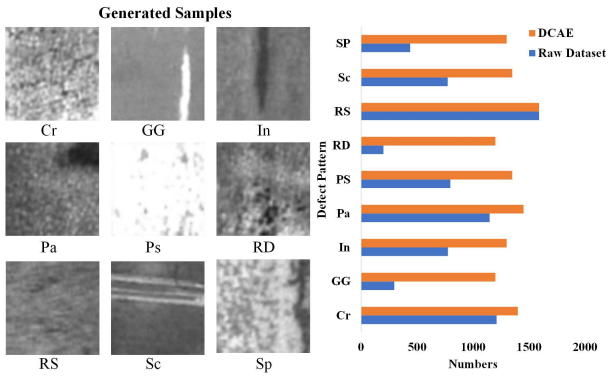


Fig. 7. Generated sample presentations and data distributions

TABLE I
COMPARISON OF EVALUATION INDICATORS OF DATA AUGMENTATION METHODS

		WM-811K		NEU-CLS-64	
Method	Model size(MB)	Entropy	Tenengrad	Entropy	Tenengrad
CAE	0.08	5.62	8.85	5.87	5.58
NL-CAE	1.61	6.21	9.68	6.43	5.89
GC-CAE	1.84	6.37	9.95	6.48	6.04
WGAN	9.92	5.96	9.01	6.01	5.76
ADC-GAN	20.45	6.28	9.94	6.51	5.98
ReAC-GAN	45.57	6.43	9.75	6.77	5.95
Simple Diffusion	196.32	6.49	10.12	6.86	6.09
DCAE	0.11	**6.54**	**10.38**	**6.97**	**6.14**

In this paper, a comparison experiment is conducted to verify the performance of the proposed model when dealing with data under different distribution scenarios. In the experiment, the distribution of the original data is analyzed. The most frequent defect type is defined as the majority class, and the other eight classes are considered as minority classes. After that, undersampling and oversampling techniques are employed to adjust the proportions of the classes. In the experiment, the distribution scenarios are set as $5:1\times 8$, $10:1\times 8$, and $20:1\times 8$. The ViT is integrated with the selected data augmentation models (e.g., WGAN, CAE, NL-CAE, GC-CAE, ADC-GAN, ReAC-GAN and Simple Diffusion) and the proposed DCAE for the classification task. Table II demonstrates the accuracy of each model under different distribution scenarios. According to the results, the proposed DCAE model achieves the highest accuracy (which are 0.933, 0.929, and 0.926) when combined with the ViT model and consistently outperforms other methods under three distribution scenarios. The results show the superior performance of DCAE in data augmentation for both non-independent and non-identically distributed data.

TABLE II
COMPARISON OF THE ACCURACY BETWEEN DIFFERENT MODELS

Method	$5:1\times 8$	$10:1\times 8$	$20:1\times 8$
CAE	0.843	0.823	0.821
NL-CAE	0.915	0.912	0.899
GC-CAE	0.901	0.897	0.883
WGAN	0.911	0.904	0.893
ADC-GAN	0.917	0.910	0.889
ReAC-GAN	0.922	0.919	0.908
Simple Diffusion	0.926	0.921	0.911
DCAE	**0.933**	**0.929**	**0.926**

Table III shows the performance evaluation of CAE and DCAE with the same kernel size. In Table III, the Entropy and Tenengrad scores of the DCAE are higher than those of the CAE with the same parameter setting, which indicates that the DC layer outperforms the convolutional layer in extracting features. Compared with the traditional convolutional layer, the DC layer has more parameters, which helps the model learn more complex features and can alleviate the underfitting problem to some extent.

Table IV shows the results of the comparison experiment which is used to investigate the impact of different channel numbers on the model performance. We can see that using only a two-layer 3×3 CNN in CAE cannot achieve the same image quality as DCAE incorporated with DCN. Under the condition of 32 channels, DCAE achieves an Entropy of 6.54 and a Tenengrad of 10.38, while CAE only reaches 6.13 and 9.84 under the same conditions. It can be seen that DCAE with 32 channels surpasses the performance of CAE with 64 channels, indicating that with the increase of parameter capacity, DCAE can alleviate underfitting to some extent and achieve better data augmentation results than the CAE.

Remark 2: Compared with the traditional CAE, the proposed DCAE shows better performance in controlling the receptive field. DC can adaptively adjust the shape and position of convolution kernels based on the degree of deformation in the target, allowing for flexible control over the receptive field.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2024.3395316, IEEE Transactions on Instrumentation and Measurement

TABLE III
PERFORMANCE EVALUATION OF CAE AND DCAE WITH THE SAME KERNEL SIZE

Method	Kernel Size (Layer 1)	Kernel Size (Layer 2)	Entropy	Tenengrad	Parameters
CAE	3×3	3×3	5.62	8.85	20288
	7×7	7×7	6.36	9.94	109888
	13×13	13×13	6.87	10.43	378688
DCAE	3×3	3×3	6.54	10.38	29539
	7×7	7×7	6.92	10.53	160096
	13×13	13×13	7.02	10.78	551776

TABLE IV
PERFORMANCE EVALUATION OF CAE AND DCAE WITH THE DIFFERENT CHANNELS

Method	Channel	Entropy	Tenengrad	Parameters
CAE	16	5.96	9.43	5536
	32	6.13	9.84	20288
	64	6.32	9.96	77440
DCAE	16	6.12	9.87	7856
	32	6.54	10.38	29539
	64	6.61	10.44	114368

Compared to the traditional convolutional layer, the DC layer has a larger number of parameters, which aids the model in learning more complex features and alleviates the underfitting problem to some extent.

B. Results and Discussions of Defect Classification

In order to demonstrate the effectiveness of the proposed framework, several popular classification models (including WMDPI [35], T-DenseNet [36], PeleeNet [44], ConvNeXt [27]), WDP-BNN [57], and WaferSegClassNet [31]) are selected for performance evaluation. The classification results of the proposed framework are compared with the classification results of selected models. In this paper, four different evaluation metrics (i.e., accuracy, recall, precision, and F1 score) are utilized to evaluate the classification results, which are described as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{20}$$

$$Recall = \frac{TP}{TP + FN} \tag{21}$$

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{23}$$

where TP denotes true positive; TN represents true negative; FP is false positive; and FN represents false negative.

In order to fully demonstrate the effectiveness of the ViT-based defect classifier, in this experiment, three public data sets (i.e., the original experimental data sets, the under-sampled experimental data sets and the augmented experimental data sets) are used to train the proposed model to evaluate the classification performance the model on data sets with different distribution. The classification samples are transformed to two-dimensional by using the t-distributed stochastic neighbor embedding (t-SNE) and are shown in Fig. 8.

As shown in Fig. 8, the model only learns the majority class when applying to the original experimental data sets, which is not satisfactory. For the under-sampled experimental data sets, the model can distinguish some defects which have obvious features. For the augmented experimental data sets, the model is able to separate various defect patterns with satisfactory accuracy, which achieves the best classification results in terms of overall performance. It can be found from the figure that all the models trained by the augmented experimental data sets have higher classification accuracy than the model trained by the original experimental data sets. It is worth mentioning that the classification of the model trained by the original experimental data sets converges fast, but the model falls into the local optimal solution for the majority class, and the model is overfitting. It can also be found in Fig. 8 that the proposed model trained by an augmented experimental data set converges fast and has the highest accuracy.

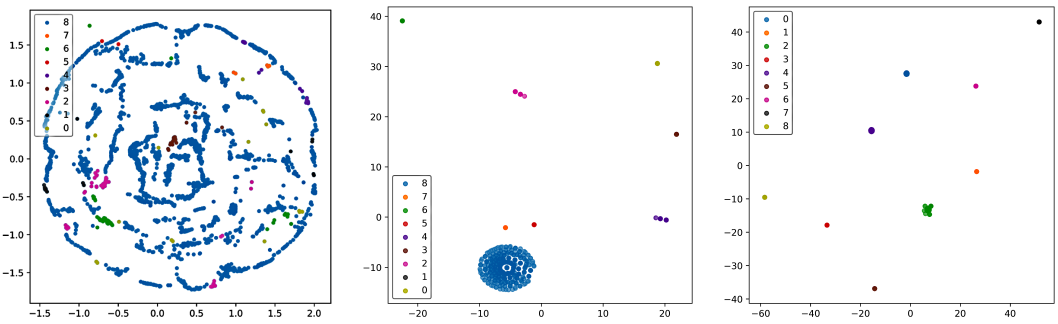
Table V shows the classification accuracy of all selected models on various defects in the test set. It can be found that compared with other classification models, the proposed method shows higher average recognition accuracy for all defects. Because the multi-head attention mechanism divides input features into multiple heads to learn different attention weights and capture different feature subspaces. The multi-head attention mechanism can make the model distinguish the features of different categories so as to improve the classification performance on imbalanced data with small sizes.

TABLE V
DPR METHODS

Classifier	WMDPI	PeleeNet	ConvNeXt	T-DenseNet	WDP-BNN	WaferSegClassNet	Our Method
Center	92.50%	95.50%	94.10%	84.50%	97.70%	97.10%	96.80%
Donut	91.50%	93.50%	94.40%	91.20%	94.60%	92.40%	95.10%
Edge-Loc	81.80%	90.90%	90.10%	81.50%	93.80%	91.50%	94.10%
Edge-Ring	97.90%	97.60%	95.50%	92.10%	95.80%	94.20%	98.70%
Loc	83.90%	88.50%	90.60%	81.80%	92.10%	92.80%	93.10%
Random	95.80%	95.40%	96.20%	85.30%	95.30%	93.40%	95.30%
Scratch	81.40%	88.90%	91.50%	84.60%	92.20%	93.70%	94.70%
Near-Full	93.30%	91.90%	92.10%	92.60%	94.40%	92.80%	95.60%
None-Pattern	97.90%	100%	99.50%	85.50%	97.50%	97.60%	100%
Average	90.70%	93.60%	93.80%	86.60%	94.80%	93.90%	95.90%

Table VI and Fig. 9 demonstrate the comparison evaluation of each model. According to Table VI, the proposed model has the best performance in terms of four evaluation indicators compared with other models. The accuracy and loss of the ViT on the training data sets and test data sets are shown in Fig. 10. According to Fig. 10, after convergence, the model has high precision in both the training data sets and test data sets, and there is no obvious overfitting or underfitting problem, which demonstrates that the training process is relatively smooth. Parallel computing techniques are used to build the proposed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2024.3395316, IEEE Transactions on Instrumentation and Measurement



(a) (b) (c)

Fig. 8. Visual analysis of classification performance t-SNE: (a) Original data; (b) Under sampling data; (c) Augmented data.

model on GPUs with the hope of saving computational time. It can be found in Fig. 10 that the proposed model exhibits satisfactory performance on both the training set and the test set.

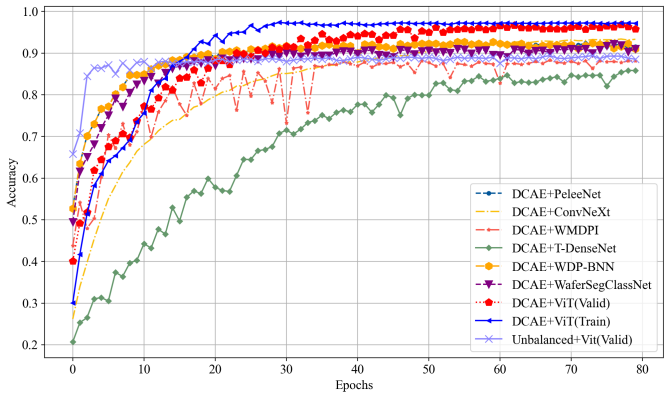
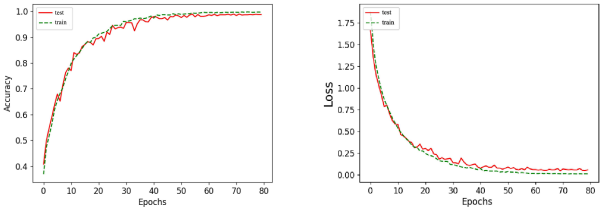


Fig. 9. Comparison of model before and after DCAE



(a) (b)

Fig. 10. ViT-based DPR performance (a) Accuracy; (b) Loss.

TABLE VI
COMPARISON OF EVALUATION INDICATORS OF EACH MODEL

Method	F1-Score	Precision	Recall	Accuracy
WMDPI	91.10	89.40	91.60	90.70
PeleeNet	86.80	89.20	85.10	93.60
ConvNeXt	92.40	92.60	91.30	93.80
T-DenseNet	77.90	79.00	76.20	80.80
WDP-BNN	92.20	93.10	93.20	94.80
WaferSegClassNet	92.60	92.90	92.10	93.90
Our Method	**94.70**	**95.60**	**94.10**	**95.90**

C. Results and Discussions of Federated Learning

Firstly, the raw data are categorized into four groups and separately allocated to four nodes which are used to simulate the clients. Secondly, local clients train and use the DCAE to increase data samples based on the local data. Thirdly, the node communication topology graph is obtained via Erdos Renyi. For connected clients, we then carry out 20 rounds of local training, and 80 rounds of data transmission and aggregation.

Experimental results of the selected methods and the proposed method on three public data sets (including WM-811K, NEU-CLS-64, and CIFAR-100) are presented in Table VII. Compared with selected federated learning frameworks (e.g., Ratio Loss-FedAvg, MOON, FedNH, BalanceFL, FedAvg, and DeceFL) and the swarm learning framework, the proposed framework achieves better performance in terms of average classification accuracy. The average classification accuracy of the proposed framework on WM-811K, NEU-CLS-64 and CIFAR-100 are 0.982, 0.990 and 0.935, respectively.

TABLE VII
RESULTS COMPARISON WITH FEDERATED LEARNING AND SWARM LEARNING METHODS

Dataset	WM-811K	NEU-CLS-64	CIFAR-100
Ratio Loss-FedAvg	0.941	0.964	0.523
MOON	0.954	0.969	0.675
FedNH	0.961	0.973	0.552
BalanceFL	0.967	0.979	0.726
DCAE+VIT+FedAvg	0.975	0.984	0.925
DCAE+VIT+SL	0.970	0.980	0.927
DCAE+VIT+DeceFL	0.972	0.981	0.923
Our Method	**0.982**	**0.990**	**0.935**

In the process of wafer production and processing, there are a number of unstable factors (e.g., commercial competition and communication outage) in different factories, which makes it difficult to realize the ideal global communication of centralized federated learning. For the centralized federated learning method, the client data need to be processed by the central server simultaneously, which leads to high computational costs [30]. As such, when handling a huge amount of data, the computational costs of the FedAvg would be extremely high, which could not be ignored.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2024.3395316, IEEE Transactions on Instrumentation and Measurement

In decentralized federated learning methods, the central client is eliminated without sacrificing the model convergence [54]. In the SDeceFL framework, the information is shared by every local client through the undirected connected graph. In this situation, completing the mutual communication of all clients requires sufficient time. Although the computational cost of model aggregation via SDeceFL may be higher than that of the FedAvg, we do not need to carry out data aggregation for establishing the global model of the SDeceFL, which means that the local client aggregation process is relatively fast, and small communication burden is brought. As a matter of fact, local communication and local data training can greatly reduce the time and cost of aggregation. Compared with FedAvg, the communication pressure of the SDeceFL framework during data transmission is lower. By using the flexible training strategy, the SDeceFL framework is more appropriate for wafer enterprises than the FedAvg framework considering the real-world practical cases.

In comparison with the FedAvg framework, the SDeceFL framework takes a longer time to complete the communication process because only topological communication between random local clients is established instead of obtaining the global communication connection. Note that the model aggregation of the SDeceFL is not to aggregate a global model. Although the local client aggregation process may take a relatively long time, the aggregation speed is fast and the communication pressure is low. After completing the model aggregation, the running time of the actual DPR is short, which is acceptable for the inspection of wafer products produced in the assembly line. As shown in Table VIII, the required accumulated communication traffic of the FedAvg, the SL, the DeceFL and the SDeceFL to achieve 97% accuracy on the WM-881K public data sets are 35283.92 MB, 31814.27 MB, 1590.16 MB and 1732.53 MB, respectively. The required accumulated communication traffic of the DeceFL and the SDeceFL frameworks is much smaller than that of other frameworks. The computational time for SDeceFL to reach the specified accuracy is only 782.15s. In addition, the decentralized federation is not only suitable for the reality of the client's communication mechanism and the prevention of privacy leakage, but also of great significance in reducing the communication bandwidth and the risk of attacks on the center.

TABLE VIII
THEORETICAL COMMUNICATION TRAFFIC AND TIME TO REACH THE TARGET ACCURACY

	WM-811K(97%)	
Method	Traffic (MB)	Time (s)
DCAE+VIT+FedAvg	35283.92	2136.67
DCAE+VIT+SL	31814.27	1972.81
DCAE+VIT+DeceFL	1590.16	1087.41
DCAE+VIT+SDeceFL	1732.53	**782.15**

The trained model with the SDeceFL framework, the FedAvg algorithm, and the SL algorithm on the test data sets are shown in Fig. 11. As shown in Fig. 11, the accuracy of the proposed model with SDeceFL is higher than that of the others on each node. It is worth mentioning that the SDeceFL framework is able to obtain reliable results without collecting data from the central server directly, which protects the original data effectively.

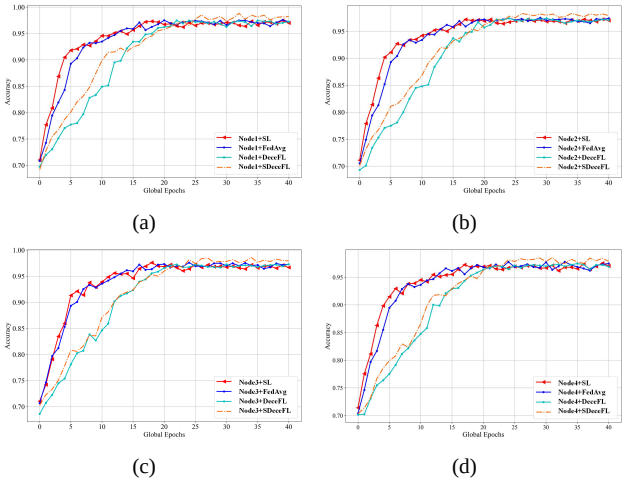


Fig. 11. Accuracy of each node in the test set: (a) Node 1; (b) Node 2; (c) Node 3; (d) Node 4.

Remark 3: Compared to traditional centralized learning, the SDeceFL framework can promote data sharing and cooperation among different clients and effectively protect the privacy of their information. Additionally, the SDeceFL framework takes 1) the characteristics of data; 2) the needs of different regions and communities in order to build inclusive models. The SDeceFL framework could effectively prevent the problem of data monopolization resulting from centralized learning brought by some large technology companies or organizations. In wafer manufacturing, the proposed DCAE-ViT-SDeceFL framework could realize the data cooperation with privacy protection among different clients and guarantee the reliability of the produced data.

D. Ablation Study

In this paper, the influence of the proposed modules in our proposed DCAE-ViT-SDeceFL framework is validated via the ablation study on WM-811K, NEU-CLS-64 and CIFAR-100 public data sets. Table IX shows the results of the ablation experiments. The influence of each module is summarized below:

(a) The influence of the data augmentation module for classification is presented in the first column of Table IX. By employing the DCAE for data augmentation, the DCAE-ResNet50-DeceFL method achieves better classification accuracy than the baseline by 3.6%, 5.7%, and 3.8% on the WM-811K, NEU-CLS-64, and CIFAR-100 public data sets, respectively.
(b) The influence of the ViT module is shown in the second column of Table IX. By comparing with the baseline, the ViT-DeceFL improves classification accuracy on WM-811K, NEU-CLS-64, and CIFAR-100 by 4.8%, 6.6%, and 11.9%, respectively.
(c) The influence of the SDeceFL module is displayed in the third column of Table IX. Compared with the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2024.3395316, IEEE Transactions on Instrumentation and Measurement

baseline, the ResNet50-SDeceFL method obtains higher classification accuracy on WM-811K, NEU-CLS-64, and CIFAR-100 public data sets by 1.4%, 1.6%, and 3.1%, respectively.

According to the aforementioned experimental results and the ablation study, we can conclude that the designed DCAE module for data augmentation improves the classifier by producing balanced data to tackle the class imbalance problem. Owing to the strong feature extraction ability, the ViT module performs as an outstanding classifier which meets the requirements of complex classification tasks (such as object detection, wafer DPR, and steel plate defect detection). Based on the proposed DPS strategy, the SDeceFL demonstrates superiority over the DeceFL method, which effectively alleviates the label heterogeneity problem among clients so as to improve the classification performance. To conclude, the proposed DCAE-ViT-SDeceFL framework achieves optimal results on WM-811K, NEU-CLS-64, and CIFAR-100 public data sets with the designed modules.

TABLE IX
RESULTS OF ABLATION EXPERIMENTS FOR DCAE-VIT-SDECEFL FRAMEWORK

DCAE	ViT	SDeceFL	WM-811K	NEU-CLS-64	CIFAR-100
✕	✕	✕	0.897	0.901	0.737
✓	✕	✕	0.933	0.958	0.775
✕	✓	✕	0.945	0.967	0.856
✕	✕	✓	0.911	0.927	0.768
✓	✓	✕	0.972	0.981	0.923
✓	✕	✓	0.941	0.965	0.829
✕	✓	✓	0.957	0.971	0.901
✓	✓	✓	**0.982**	**0.990**	**0.935**

VI. CONCLUSION AND FUTURE WORKS

In this paper, a novel privacy protection framework has been proposed for wafer DPR. A novel SDeceFL framework has been proposed to tackle the inter-class heterogeneity problem of different clients and enhance data privacy. A new data augmentation module, the DCAE, has been developed to tackle the class imbalance problem. The ViT has been combined with the DCAE module for DPR of the wafer data under class imbalance. The proposed DCAE-ViT-SDeceFL framework has been evaluated and tested on the WM-811K public data sets. Experiment results have shown the superiority of the proposed framework over some existing frameworks. The proposed framework can provide high-quality data for clients which have sparse or imbalanced data. In addition, the proposed DCAE-ViT-SDeceFL framework could allow different clients to train a reliable model while preventing data leakage, which has unique advantages in data generation and privacy protection. In the future, we aim to: 1) apply the proposed framework to other data analysis and fault detection tasks [6], [14], [18], [55]; 2) employ evolutionary computation methods to choose the hyperparameters of the ViT-based DPR network [12], [22], [46], [48]; 3) develop new deep learning-based wafer DPR methods based on federated learning [28]; 4) design a lower communication cost and more private parameter fusion sharing strategy based on the SDeceFL framework [5], [52], [58], [61].

REFERENCES

[1] M. Arjovsky, S. Chintala and L. Bottou, Wasserstein GAN, In: *Proceedings of the 2018 International Conference on Machine Learning (ICML)*, Long Beach, USA, Aug. 2018, pp. 214-223.
[2] D. Barnea and H. F. Silverman, A class of algorithms for fast digital image registration, *IEEE Transactions on Computers*, vol. C-21, no. 2, pp. 179-186, 1972.
[3] F. Bi, T. He and X. Luo, A fast nonnegative autoencoder-based approach to latent feature analysis on high-dimensional and incomplete data, *IEEE Transactions on Services Computing*, in press, DOI: 10.1109/TSC.2023.3319713.
[4] Y. Cao, F. Xu and S. Lin, GCNet: Non-local networks meet squeeze-excitation networks and beyond, In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Los Angeles, USA, Jun. 2019, pp. 6788-6798.
[5] H. Chen, C. Li, M. Mafarja, A. A. Heidari, Y. Chen and Z. Cai, Slime mould algorithm: a comprehensive review of recent variants and applications, *International Journal of Systems Science*, vol. 54, no. 1, pp. 204-235, 2023.
[6] Q. Cui, K. Liu, Z. Ji and W. Song, Sampling-data-based distributed optimisation of second-order multi-agent systems with PI strategy, *International Journal of Systems Science*, vol. 54, no. 6, pp. 1299-1312, 2023.
[7] J. Dai, H. Qi and Y. Li, Deformable convolutional networks, In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, Jul. 2017, pp. 764-773.
[8] Y. Dai, Z. Chen, J. Li, S. Heinecke, S. Li and R. Xu, Tackling data heterogeneity in federated learning with class prototypes, In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Washington, USA, Feb. 2023, pp. 7314-7322.
[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, An image is worth 16×16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929v2*, 2020.
[10] J. Dou, Z. Gao, G. Wei, Y. Song and M. Li, Switching synthesizing-incorporated and cluster-based synthetic oversampling for imbalanced binary classification, *Engineering Applications of Artificial Intelligence*, vol. 123, art. no. 106193, 2023.
[11] J. Dou and Y. Song, An improved generative adversarial network with feature filtering for imbalanced data, *International Journal of Network Dynamics and Intelligence*, vol. 2, no. 4, art. no. 100017, 2023.
[12] J. Fang, W. Liu, L. Chen, S. Lauria, A. Miron and X. Liu, A survey of algorithms, applications and trends for particle swarm optimization, *International Journal of Network Dynamics and Intelligence*, vol. 2, no. 1, pp. 24-50, 2023.
[13] S. Gaba, I. Budhiraja, V. Kumar, S. Garg, G. Kaddoum and M. M. Hassan, A federated calibration scheme for convolutional neural networks: Models, applications and challenges, *Computer Communications*, vol. 192, pp. 144-162, 2022.
[14] F Han, J. Liu, J. Li, J. Song, M. Wang and Y. Zhang, Consensus control for multi-rate multi-agent systems with fading measurements: The dynamic event-triggered case, *Systems Science & Control Engineering*, vol. 11, no. 1, art. no. 2158959, 2023.
[15] Y. He, X. Wen and J. Xu, A Semi-Supervised Inspection Approach of Textured Surface Defects under Limited Labeled Samples, *Coatings*, vol. 12, pp. 1707, 2022.
[16] E. Hoogeboom, J. Heek and T. Salimans, Simple Diffusion: End-to-end diffusion for high resolution images, In:*Proceedings of the 40th International Conference on Machine Learning*, Hawaii, USA, Jul. 2023, 13213-13232.
[17] L. Hou, Q. Cao, H. Shen, S. Pan, X. Li and X. Cheng, Conditional gans with auxiliary discriminative classifier, In:*Proceedings of the 39th International Conference on Machine Learning*, Baltimore, USA, Jul. 2022, 8888-8902.
[18] Y. Hou, Y. Zhang, J. Lu, N. Hou and D. Yang, Application of improved multi-strategy MPA-VMD in pipeline leakage detection, *Systems Science & Control Engineering*, vol. 11, no. 1, art. no. 2177771, 2023.
[19] J. Hu, K. Pan, Y. Song, G. Wei and C. Shen, An improved feature selection method for classification on incomplete data: Non-negative latent factor-incorporated duplicate MIC, *Expert Systems with Applications*, vol. 212, art. no. 118654, 2023.
[20] M. Kang, W. Shim, M. Cho and J. Park, Rebooting acgan: Auxiliary classifier GANs with stable training, *Advances in Neural Information Processing Systems*, vol. 34, pp. 23505-23518, 2021.

[21] K. Kyeong and H. Kim, Classification of mixed-type defect patterns in wafer bin maps using convolutional neural networks, *IEEE Transactions on Semiconductor Manufacturing*, vol. 31, no. 3, pp. 395-402, 2018.

[22] H. Li, H. Liu, C. Lan, Y. Yin, P. Wu, C. Yan and N. Zeng, SMWO/D: A decomposition-based switching multi-objective whale optimiser for structural optimisation of Turbine disk in aero-engines, *International Journal of Systems Science*, vol. 54, no. 8, pp. 1713-1728, 2023.

[23] J. Li, F. Tan, C. He, Z. Wang, H. Song, P. Hu and X. Luo, Saliency-aware dual embedded attention network for multivariate time-series forecasting in information technology operations, *IEEE Transactions on Industrial Informatics*, in press, DOI: 10.1109/TII.2023.3315369.

[24] Q. Li, B. He, and D. Song, Model-contrastive federated learning, In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Online, June. 2021, pp. 10713-10722.

[25] J. Liao, H.-K. Lam, S. Gulati and B. Hayee, Improved computer-aided diagnosis system for nonerosive reflux disease using contrastive self-supervised learning with transfer learning, *International Journal of Network Dynamics and Intelligence*, vol. 2, no. 3, art. no. 100010, 2023.

[26] W. Liu, Z. Wang, L. Tian, S. Lauria and X. Liu, Melt pool segmentation for additive manufacturing: A generative adversarial network approach, *Computers & Electrical Engineering*, vol. 92, art. no. 107183, 2021.

[27] Z. Liu, H. Mao and C. Y. Wu, A convnet for the 2020s, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, New Orleans, USA, Jun. 2022, pp. 11976-11986.

[28] G. Ma, Z. Wang, W. Liu, J. Fang, Y. Zhang, H. Ding and Y. Yuan, Estimating the state of health for lithium-ion batteries: A particle swarm optimization-assisted deep domain adaptation approach, *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 7, pp. 1530-1543, 2023.

[29] J. Masci, U. Meier, D. Ciresan and J. Schmidhuber, Stacked convolutional auto-encoders for hierachucal feature extraction, In: *Proceedings of the 21st International Conference on Artificial Neural Networks*, Espoo, Finland, Jun. 2011, pp. 52-59.

[30] B. McMahan, E. Moore and D. Ramage, Communication-efficient learning of deep networks from decentralized data, In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Lauderdale, USA, Apr. 2017, pp. 1273-1282.

[31] S. Nag, D. Makwana, S. C. Teja R, S. Mittal and C. K. Mohan, WaferSegClassNet - A light-weight network for classification and segmentation of semiconductor wafer defects, *Computers in Industry*, vol. 142, art. no. 103720, 2022.

[32] F. D. Palma, G. D. Nicolao and G. Miraglia, Unsupervised spatial pattern classification of electrical-wafer-sorting maps in semiconductor manufacturing, *Pattern Recognition Letters*, vol. 26, no. 12, pp. 1857-1865, 2005.

[33] C. Qin, R. Yang, M. Huang, W. Liu and Z. Wang, Spatial variation generation algorithm for motor imagery data augmentation: Increasing the density of sample vicinity, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 3675-3686, 2023.

[34] P. Ramachandran, B. Zoph and Q. V. Le, Searching for activation functions, *arXiv preprint arXiv:1710.05941*, 2017.

[35] M. Saqlain, B. Jargalsaikhan and B. Lee, A voting ensemble classifier for wafer map defect patterns identification in semiconductor manufacturing, *IEEE Transactions on Semiconductor Manufacturing*, vol. 33, no. 2, pp. 196-205, 2020.

[36] Z. Shen, J. Wu, W. Zhang and H. Huang, Wafer map defect recognition based on deep transfer learning, *IEEE Access*, vol. 8, pp. 194015-194026, 2020.

[37] X. Shuai, Y. Shen, S. Jiang, Z. Zhao, Z. Yan and G. Xing, BalanceFL: Addressing class imbalance in long-tail federated learning, In: *ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, Milan, Italy, May. 2022, pp. 271-284.

[38] L. Sun, Y. Chen, B. Wang and Y. Tang, Large kernel matters improve semantic segmentation by global convolutional network, In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, USA, Jun. 2019, pp. 1594-1603.

[39] Z. Tang, E. Tian and Y. Wang, Nondestructive defect detection in castings by using spatial attention bilinear convolutional neural network, *IEEE Transactions on Industrial Informatics*, vol. 17, no. 1, pp. 82-89, 2021.

[40] G. Tong, Q. Li and Y. Song, Two-stage reverse knowledge distillation incorporated and Self-Supervised Masking strategy for industrial anomaly detection, *Knowledge-Based Systems*, vol. 273, art. no. 110611, 2023.

[41] D. Wang, S. Shi, J. Lu, Z. Hu and J. Chen, Research on gas pipeline leakage model identification driven by digital twin, *Systems Science & Control Engineering*, vol. 11, no. 1, art. no. 2180687, 2023.

[42] J. Wang, Y. Zhuang and Y. Liu, FSS-Net: A fast search structure for 3D point clouds in deep learning, *International Journal of Network Dynamics and Intelligence*, vol. 2, no. 2, art. no. 100005, 2023.

[43] L. Wang, S. Xu, X Wang and Q. Zhu, Addressing class imbalance in federated learning, In: *Proceedings of the AAAI Conference on Artificial Intelligence*, California, USA, Feb. 2021, pp. 10165-10173.

[44] R. J. Wang, X. Li and C. X. Ling. Pelee: A real-time object detection system on mobile devices, *Advances in neural information processing systems*, vol. 31, 2018.

[45] X .Wang, R .Girshick, A .Gupta and K. He, Non-local neural networks, In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, Jun. 2018, pp. 7794-7803.

[46] Y. Wang, W. Liu, C. Wang, F. Fadzil, S. Lauria and X. Liu, A novel multi-objective optimization approach with flexible operation planning strategy for truck scheduling, *International Journal of Network Dynamics and Intelligence*, vol. 2, no. 2, art. no. 100002, 2023.

[47] M. Wu, J. S. R. Jang and J. Chen, Wafer map failure pattern recognition and similarity ranking for large-scale data sets, *IEEE Transactions on Semiconductor Manufacturing*, vol. 28, no. 1, pp. 1-12, 2015.

[48] L. Xu, J. Du, B. Song and M. Cao, A combined backstepping and fractional-order PID controller to trajectory tracking of mobile robots, *Systems Science & Control Engineering*, vol. 10, no. 1, pp. 134-141, 2022.

[49] Y. Xue, R. Yang, X. Chen, Z. Tian and Z. Wang, A novel local binary temporal convolutional neural network for bearing fault diagnosis, *IEEE Transactions on Instrumentation and Measurement*, vol. 72, art. no. 3525013, 2023.

[50] Z. Yao, Y. Cao, Y. Lin, Z. Liu, Z. Zhang and H. Hu, Leveraging batch normalization for vision transformers, In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Montreal, Canada, Oct. 2021, pp. 413-422.

[51] F. Yu, V. Koltun and T. A. Funkhouser, Dilated residual networks, In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Hawaii, USA, Jul. 2017, pp. 472-480.

[52] H. Yu, J. Hu, B. Song, H. Liu and X. Yi, Resilient energy-to-peak filtering for linear parameter-varying systems under random access protocol, *International Journal of Systems Science*, vol. 53, no. 11, pp. 2421-2436, 2022.

[53] N. Yu, H. Chen, Q. Xu, M. M. Hasan and O. Sie, Wafer map defect patterns classification based on a lightweight network and data augmentation, *CAAI Transactions on Intelligence Technology*, in press, DOI: 10.1049/cit2.12126.

[54] Y. Yuan, J. Liu and D. Jin, DeceFL: A principled decentralized federated learning framework, *arXiv preprint arXiv:2107.07171*, 2021.

[55] X. Yue, J. Chen and G. Zhong, Metal surface defect detection based on Metal-YOLOX, *International Journal of Network Dynamics and Intelligence*, vol. 2, no. 4, art. no. 100020, 2023.

[56] J. Zhang, Z. Li, B. Li, J. Xu, S. Wu, S. Ding and C. Wu, Federated learning with label distribution skew via logits calibration, In:*Proceedings of the 39th International Conference on Machine Learning*, Baltimore, USA, Jul. 2022, 26311–26329.

[57] Q. Zhang, Y. Zhang, J. Li and Y. Li, WDP-BNN: Efficient wafer defect pattern classification via binarized neural network, *Integration*, vol. 85, pp. 76-86, 2022.

[58] T. Zhang, Q. Liu, J. Liu, Z. Wang and H. Li, Multiple-bipartite consensus for networked Lagrangian systems without using neighbours' velocity information in the directed graph, *Systems Science & Control Engineering*, vol. 11, no. 1, art. no. 2210185, 2023.

[59] W. Zhang and X. Li, Federated Transfer Learning for Intelligent Fault Diagnostics Using Deep Adversarial Networks With Data Privacy, *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 1, pp. 430-439, 2022.

[60] W. Zhang and X. Li, Data privacy-preserving federated transfer learning in machinery fault diagnostics using prior distributions, *Structural Health Monitoring*, vol. 21, pp. 1329–1344, 2022.

[61] X. Zhang, J. Song, P. Cheng, K. Shi and S. He, Mean square exponential stabilisation for directional 2D Roesser hidden Markov model, *International Journal of Systems Science*, vol. 54, no. 4, pp. 867-879, 2023.

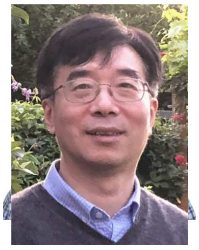
[62] N. Zeng, P. Wu, Z. Wang, H. Li, W. Liu and X. Liu, A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection, *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1-14, 2022.

Yong Zhang received the Ph.D. degree in control theory and control engineering from Huazhong University of Science and Technology, Wuhan, China, in 2010. From 2014 to 2015, he was a Visiting Scholar with the Department of Information Systems and Computing, Brunel University London, Uxbridge, U.K. He is currently a Professor with the School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, China. He has authored over 50 papers in refereed international journals. His current research interests include health state assessment of industrial equipment/product with multimodal data. Dr. Zhang is a very active Reviewer for many international journals.

Rukai Lan received the B.E. in the School of Mechanical and Electrical Engineering from Anhui Jianzhu University, Anhui, China. Since 2021, he has been pursuing his doctorate in the School of Information Science and Engineering, Wuhan University of Science and Technology. His current research interests include computer vision, deep learning, and autonomous driving.

Xianhe Li received the B.E. and M.Sc. degrees from the School of Information Science and Engineering, Wuhan University of Science and Technology, in 2019 and 2023, respectively. His research interests include computer vision and defect detection.

Jingzhong Fang received his B.Eng. degree in automation from Shandong University of Science and Technology, Qingdao, China, in 2020, and the M.Sc. degree in data science and analytics from Brunel University London, Uxbridge, U.K., in 2021. He is currently pursuing the Ph.D. degree in Computer Science at Brunel University London, Uxbridge, U.K. His research interests include data analysis and deep learning techniques.



Zidong Wang (S'03–M'04–SM'08–F'14) received the B.Sc. degree in mathematics in 1986 from Suzhou University, Suzhou, China, and the M.Sc. degree in applied mathematics in 1990 and the Ph.D. degree in electrical engineering in 1994, both from Nanjing University of Science and Technology, Nanjing, China.

He is currently a Professor of Dynamical Systems and Computing in the Department of Computer Science, Brunel University London, Uxbridge, U.K. From 1990 to 2002, he held teaching and research appointments in universities in China, Germany and the U.K. Prof. Wang's research interests include dynamical systems, signal processing, bioinformatics, control theory and applications. He has published a number of papers in international journals. He is a holder of the Alexander von Humboldt Research Fellowship of Germany, the JSPS Research Fellowship of Japan, William Mong Visiting Research Fellowship of Hong Kong.

Prof. Wang serves (or has served) as the Editor-in-Chief for *International Journal of Systems Sciences*, the Editor-in-Chief for *Neurocomputing*, the Editor-in-Chief for *Systems Science & Control Engineering*, and an Associate Editor for 12 international journals including *IEEE Transactions on Automatic Control*, *IEEE Transactions on Control Systems Technology*, *IEEE Transactions on Neural Networks*, *IEEE Transactions on Signal Processing*, and *IEEE Transactions on Systems, Man, and Cybernetics Part C*. He is a Member of the Academia Europaea, a Member of the European Academy of Sciences and Arts, an Academician of the International Academy for Systems and Cybernetic Sciences, a Fellow of the IEEE, a Fellow of the Royal Statistical Society and a member of program committee for many international conferences.

Weibo Liu received the B.Eng. degree in electrical engineering from the Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, U.K., in 2015, and the Ph.D. degree in artificial intelligence in 2020 from the Department of Computer Science, Brunel University London, Uxbridge, U.K. He is currently a Lecturer in the Department of Computer Science, Brunel University London, Uxbridge, U.K. His research interests include intelligent data analysis, evolutionary computation, transfer learning, machine learning, and deep learning. He serves as an Associate Editor for the *Journal of Ambient Intelligence and Humanized Computing* and the *Journal of Cognitive Computation*. He is a very active reviewer for many international journals and conferences.