Strojniški vestnik - Journal of Mechanical Engineering 70(2024)11-12, pp1-pp15
DOI:10.5545/sv-jme.2023.900
Original Scientific Paper
Received for review: 2023-12-18
Received revised form: 2024-10-08
Accepted for publication: 2024-10-09

Improving the Efficiency of Steel Plate Surface Defect Classification by Reducing the Labelling Cost Using Deep Active Learning

Wenjia Yang[1] – Youhang Zhou[1,2,*] – Gaolei Meng[1] – Yuze Li[1] – Tianyu Gong[1]

[1] Xiangtan University, School of Mechanical Engineering and Mechanics, China
[2] Xiangtan University, Engineering Research Center of Complex Tracks Processing Technology and Equipment of Ministry of Education, China

Efficient surface defects classification is one of the research hotpots in steel plate defect recognition. Compared with traditional methods, deep learning methods have been effective in improving classification accuracy and efficiency, but require a large amount of labeled data, resulting in limited improvement of detection efficiency. To reduce the labeling effort under the premise of satisfying the classification accuracy, a deep active learning method is proposed for steel plate surface defects classification. Firstly, a lightweight convolutional neural network is designed, which speeds up the training process and enhances the model regularization. Secondly, a novel uncertainty-based sampling strategy, which calculates Kullback-Leibler (KL) divergence between two kinds of distributions, is used as an uncertainty measure to select new samples for labeling. Finally, the performance of the proposed method is validated using the steel surface defects dataset from Northeastern University (NEU-CLS) and the milling steel surface defects dataset from a local laboratory. The proposed global pooling-based classifier with global average pooling (GAPC) network model combined with the Kullback-Leibler divergence sampling (KLS) strategy has the best performance in the classification of steel plate surface defects. This method achieves 97 % classification accuracy with 44 % labeled data on the NEU-CLS dataset and 92.3 % classification accuracy with 50 % labeled data on the milling steel surface defects dataset. The experimental results show that the proposed method can achieve steel surface defects classification accuracy of not less than 92 % with no more than 50 % of the dataset to be labeled, which indicates that this method has potential application in surface defect classification of industrial products.

Keywords: surface defect classification, convolutional neural network, active learning, global pooling

Highlights

- *A deep active learning method for improving the efficiency of steel plate surface defect classification by reducing labeling cost is proposed.*
- *Proposed GAPC-based CNN model can speed up the training process and enhance the model regularization.*
- *The KLS uncertainty sampling strategy can effectively reduce the amount of label data required.*
- *The proposed method can achieve classification accuracy of not less than 92 % with no more than 50 % of the dataset requiring labeling.*

0 INTRODUCTION

Steel plate is widely utilized in aerospace production [1] and [2], architecture industry [3] and [4], and machinery manufacturing [5] and [6]. Its surface defects are the key factors affecting the quality of steel products. However, different categories of steel surface defects often occur during production. Surface defects of the steel plate, such as rolled-in scale, patches, crazing, pitted surface, inclusion, and scratches, are unavoidable. The defects not only affect production quality but also incur economic losses and give rise to safety concerns. An efficient classification of steel plate surface defects can contribute to a better understanding of the causes of defect formation, optimize production processes, enhance product quality and improve economic efficiency. Therefore, efficient and accurate classification of surface defects has become an indispensable function in the iron and steel industry.

The traditional methods of steel plate surface defect inspection mainly include manual visual inspection, eddy current inspection [7] and magnetic flux leakage testing [8], etc. Owing to the influence of subjective factors and a high error inspection rate, these methods have been unable to meet the current inspection requirements of the iron and steel industry. In recent years, with the development of science and technology, the inspection technology based on deep learning and machine vision, as a kind of non-contact inspection method, has become a research hotspot in the field of surface defect inspection [9] to [11]. As a kind of deep learning model, convolutional neural network (CNN) [12] has outstanding performance in many classification tasks in industry. The success of CNN models for classification tasks brings a rapid

*Corr. Author's Address: School of mechanical Engineering and mechanics, Xiangtan University, Xiangtan, China, zhouyouhang@xtu.edu.cn

development of CNN-based steel defect classification methods. Zhou et al. [13] proposed a CNN model for effective and robust classification of surface defects in hot rolled steel sheets. This model achieved a classification accuracy of over 97 % through 500 iterations using 60 % of the dataset as labeled data for training. He et al. [14] proposed a new method for defects detection and classification of low carbon steel wire arc additive manufacturing (WAAM) products using an improved cost-sensitive convolutional neural network. This method achieved a classification accuracy of over 92 % through 800 iterations using 75 % of the dataset as labeled data for training. However, although the above studies can obtain ideal classification accuracy, they need more than 60 % of the dataset as labeled data for training, which will inevitably produce high labeling cost and affect the classification efficiency.

Currently, among the methods that can effectively alleviate the labeling cost, four of them have shown great potential: transfer learning [15], data augmentation by generative models [16] and [17], semi-supervised learning [18] to [20], and active learning [21].

Transfer learning is developed on the assumption that earlier layers in the convolutional base learn generic, reusable local patterns like curves and edges, while higher layers learn task-specific features. Hence the lower layers in an existing model trained on one big dataset can be reused on a relatively small-sized target dataset to improve the generalization ability of the model. Both Fu [22] and Yang [23] adopted pre-trained SqueezeNet [24] as backbone architecture. Although the high classification accuracy had been achieved, all available data still needs to be labeled. In addition, surface defects have different image contexts compared to most large datasets, so it is hard to find the right number of layers to reuse [25]. This means the training time will be inevitably extended. As shown in [22], the model was trained on NEU-CLS dataset [26] in 20 minutes by using a NVDIA TITAN X GPU (12G memory).

Generative models, like variational autoencoders (VAE) [27], generative adversarial networks (GAN) [28], and their variants, provide a different way to solve the problem. Instead of manually collecting more training data, the existing samples can be used to guide the generation of new artificial samples. Yun et al. [29] used conditional convolutional VAE [30] to generate images for each kind of defect and then used a CNN-based model for classification. Tang et al. [31] took a similar approach to classify photovoltaic module defects. However, the generative model they adopted was GAN. This kind of method has the disadvantage of generating many samples with less information because the generation process does not take sample informativeness into account [32]. Consequently, these methods may prolong the training time and waste computational resources.

Semi-supervised learning uses both labeled and unlabeled data for model training. Gao et al. [33] proposed combined pseudo-label CNN (PLCNN). However, PLCNN abandoned the unsupervised pretraining process that plays an essential role in the original paper. This may harm the model's classification ability. The accuracy of PLCNN on NEU-CLS dataset is 90.7 %, which is inferior to other methods. He et al. [34] and He et al. [35] both utilized semi-supervised GAN (SGAN) to perform defects classification. The major difference between their works is the former used a trained convolutional autoencoder [36] to initialize the discriminator in SGAN with identical topology, whereas the latter trained another residual network and combined it with SGAN to form a multi-training algorithm. Using generative models may help to learn the latent structure of defects, but it will take more time and computational resources to complete the training.

Compared with the above three methods, the uncertainty-based active learning method is an effective approach to reduce both labeling cost and computing resource, where the most informative samples are incrementally selected for labeling to improve the model classification ability at low labeling budgets. Yang et al. [37] presented a new framework that combines a fully convolutional network and an uncertainty method in active learning to reduce biomedical image analysis annotation effort by making judicious suggestions on the most effective annotation areas. This method can achieve state-of-the-art segmentation performance using 50 % of the training data. There are three widely used uncertainty sample strategies, namely: least confident (LC) [38], margin sampling (MS) [39] and entropy (EN) [40]. These strategies assume that model's prediction on an unlabeled data pool obtains the model's uncertainty over the unlabeled data. By applying different uncertainty measures, the most informative samples can be selected for labeling. However, these methods only utilize the model predictions on the unlabeled data, ignoring the uncertainty information of the model on the labeled data, which is considered useful. By taking both types of uncertainty into consideration, the uncertainty of model can be better measured, and the most informative samples can be screened out for labeling.

In this work, a deep active learning method for steel plate defect classification is proposed. To enhance the learning efficiency and reduce the computational cost, a simplified convolutional network is designed based on simple features of hot rolled steel plate surface defects to expedite the training process. A global pooling layer is adopted to improve the model's generalization ability. Experiments are carried out on both global average pooling and global max pooling to find which is more suitable for active learning. Then, the average probability distribution over classes (PDC) calculated from labeled data for a specific class is considered as the best performance of the model on this class to integrate two kinds of uncertainty. By quantifying the difference between the PDC of an unlabeled sample and the optimal model performance on the predicted sample label, a new uncertainty index is obtained to guide sample selection. Based on experiment results on the NEU-CLS dataset and milling steel plate surface defects dataset, the proposed method can achieve superior classification results with less labeled data.

1 OVERALL FRAMEWORK

The framework (Fig. 1) consists of two key components: a convolutional neural network for model training and a sample strategy for data collection. A small portion of the existing dataset is randomly selected for labeling. The selected samples from the labeled data pool which is denoted as D^{L}, and the rest of the samples compose the unlabeled data pool which is denoted as D^{U}. As is shown in Fig. 1, the model is firstly trained by the initial labeled data pool. Then the trained model is used to predict the labeled and unlabeled data respectively, and the defect images are selected from the unlabeled data pool for labeling according to the proposed sampling strategy. At this point, the two data pools are updated. The training process is restarted and repeated until the model classification performance is satisfied.

2 METHOD

2.1 Model Design

2.1.1 Feature Extractor

Steel plate surface defects are not as complex as human faces or other objects with lots of features. There is no need to use a big and complex convolutional base which can be hard to train. Hence, a shallow network is utilized to reduce training time. The convolutional base adopted in this paper can be considered as a shallow version of visual geometry group (VGG) network [41] and [42], where only four convolutional layers are kept, and every two convolutional layers

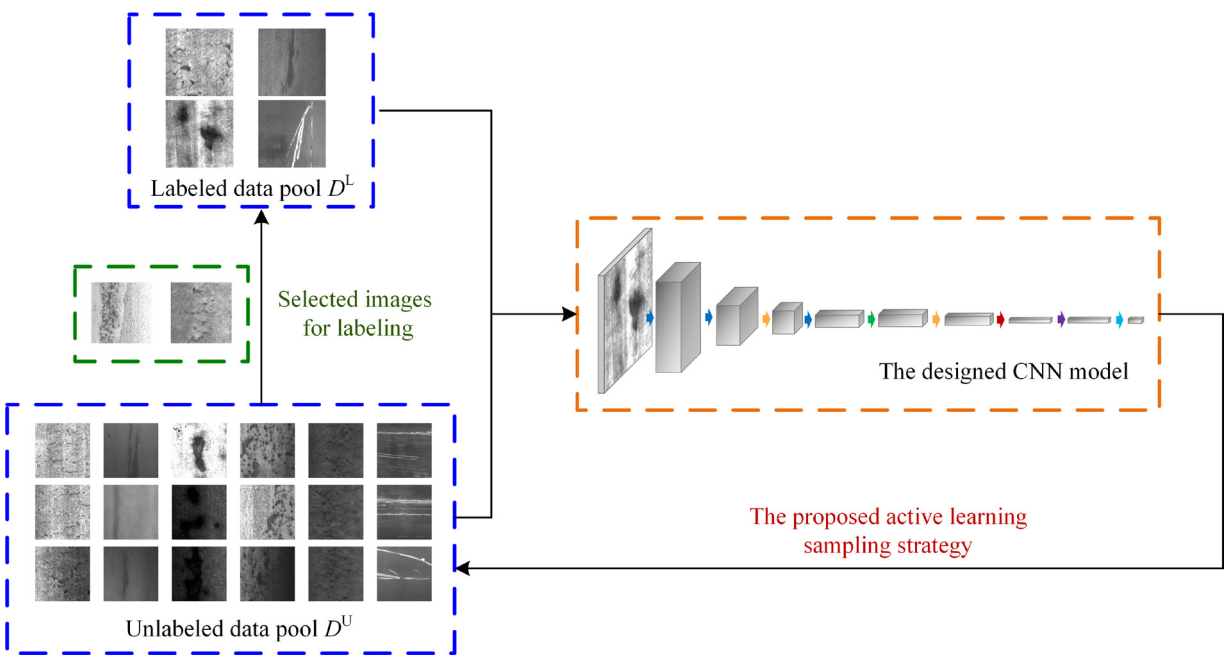

Fig. 1. *Overview of the proposed method framework*

are followed by a max-pooling layer. All batch normalization layers are removed to speed up training.

2.1.2 Global Pooling as Structural Regularizer

The traditional classifier (TRC), followed by the convolutional base, is composed of two hidden layers and a dense output layer [43] and [44]. In recent years, a new type of classifier has emerged. Scholars [45] and [46] have replaced the two hidden layers in traditional classifier with global average pooling layer (GAP). Szegedy claimed that this replacement has boosted the top-1 accuracy by about 0.6 %. The outputs of the feature extractor are multiple feature maps. In TRC setting, the feature maps need to go through the flatten layer to be expanded into a one-dimensional feature vector before they can be passed into the classifier. However, GAP and global max pooling (GMP) calculate the average and maximum value of each feature map as the output. Fig. 2 shows the difference between flatten and global pooling.

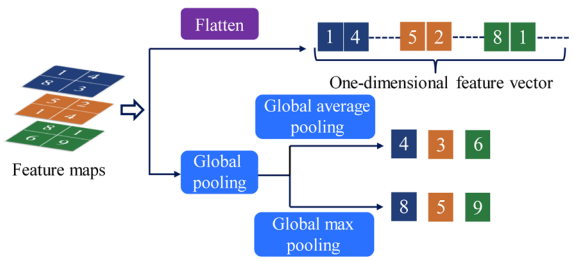

Fig. 2. *The difference between flatten and global pooling*

To explore the performance of classifiers based on global pooling, this paper carried out a comparative analysis of global pooling-based classifier (GPC) and TRC based on commonly used regularization methods, and clarified which classifier is the most effective in the subsequent experimental results. The designed CNN model structure is shown in Fig. 3. Compared with traditional CNN, the main difference is the use of global pooling to replace the hidden layers of the final densely connected classifier.

2.2 Sampling Strategy

The labeled data pool D^{L} can not only be used to train the model, but also contains the model's uncertainty information about the dataset. Unlike traditional uncertainty-based sample strategies which only utilize model predictions on the unlabeled data, The Kullback-Leibler divergence sampling (KLS) is proposed to consider the uncertainty of the model on the labeled data and incorporate it into the sampling process.

The single sample in the labeled data pool D^{L} and the unlabeled data pool D^{U} is denoted as x, and the corresponding label is y. The Initial model is trained by D^{L}. After predicting every sample in D^{U}, a prediction array is available, each row of which is the PDC for a specific sample. For sample $x_u \in D^{\mathrm{U}}$, its PDC is denoted as:

$$p\left(y=y' \mid x_u, W\right)=\left\{\begin{array}{l} p\left(y^1 \mid x_u, W\right), \\ p\left(y^2 \mid x_u, W\right), \\ \ldots, \\ p\left(y^C \mid x_u, W\right) \end{array}\right\}_{1\times C}, \qquad (1)$$

where $p(y^c \mid x_u, W)$, $c \in (1, \ldots, C)$ represents the probability that the label of sample x_u is y^c, with C being the number of classes. W represents the model parameters. y' is the predicted label which can be calculated by:

$$y' = \arg\max_{y} p\left(y \mid x_u, W\right). \qquad (2)$$

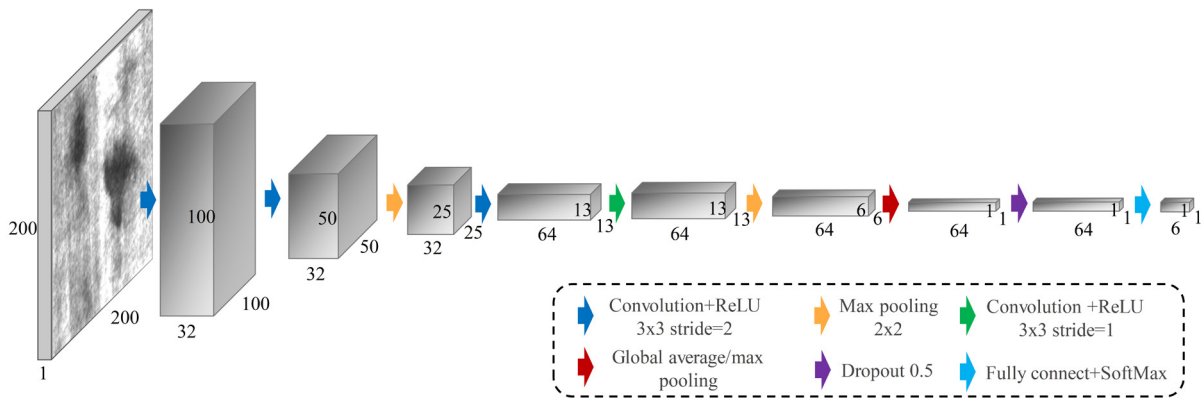

Fig. 3. *Architecture of the designed network*

The generated prediction array can be split into separate PDC subgroups based on the sample's predicted label. Suppose $y'=y^c$, the PDC of x_u will be contained in group y^c, which is:

$$p\left(y=y^c \mid x_u, W\right)=\left\{\begin{array}{l} p\left(y^1 \mid x_u, W\right), \\ p\left(y^2 \mid x_u, W\right), \\ \ldots, \\ p\left(y^C \mid x_u, W\right) \end{array}\right\}_{1 \times C}. \quad (3)$$

The next step is to predict the labeled data. For a trained model with strong regularization, it is optimized for the trained data pool D^{L}. Therefore, its predictions on D^{L} are considered as its best performance. For sample $x_1 \in D^{\mathrm{L}}$, its PDC and predicted label can be calculated by Eqs. (1) and (2). For label y^c, the average PDC $p(y=y^c \mid X_1^c, W)_{\mathrm{avg}}$ can be calculated by:

$$p\left(y=y^c \mid X_l^c, W\right)_{avg}=\frac{1}{N}\left\{\begin{array}{l} \sum_{n=1}^{N} p\left(y^1 \mid x_l^n, W\right), \\ \sum_{n=1}^{N} p\left(y^2 \mid x_l^n, W\right), \\ \ldots, \\ \sum_{n=1}^{N} p\left(y^C \mid x_l^n, W\right) \end{array}\right\}, \quad (4)$$

where X_1^c, represents the set of samples whose predicted labels are y^c, N is the size of X_1^c.

$p(y=y^c \mid X_1^c, W)_{\mathrm{avg}}$ is taken as the best performance of model on label y^c. In the PDC group belonging to y^c, any normal sample's PDC should be close to this average distribution. If $p(y=y^c \mid x_u, W)$ diverges too far from $p(y=y^c \mid X_1^c, W)_{\mathrm{avg}}$, the sample x_u is considered as abnormal. In other words, the model is uncertain about the class of sample x_u.

To measure the difference between two distributions, the Kullback-Leibler (KL) divergence [47] is introduced:

$$KL(P \| Q)=\sum_i P(i) \log \frac{P(i)}{Q(i)}. \quad (5)$$

The KL divergence between these two distributions (*kls*) as the model's uncertainty about sample x_u can be calculated by:

$$kls_u=KL\left(p\left(y=y^c \mid X_l^c, W\right)_{avg} \| \, p\left(y=y^c \mid x_u, W\right)\right), \quad (6)$$

$p(y=y^c \mid X_1^c, W)_{\mathrm{avg}}$ is calculated for every label y^c as the performance baseline. Then, for every sample's PDC in every PDC subgroup, the corresponding *kls* value is calculated. In every subgroup, the top k samples with highest *kls* value will be selected for labeling. This strategy naturally guarantees the diversity in each selected batch by sampling an equal number of samples in every subgroup. This means that the total number of selected samples will be $k \times C$.

2.3 Stopping Criterion

Considering that active learning is iterative, a stopping criterion is needed to stop the training process when the model performance is reached. Therefore, the stopping criterion should be highly connected to the model performance. As the designed model is strongly regularized, validation accuracy (VA) is adopted as the criterion. According to the experiments, the validation accuracy of the network is always less than or equal to the test accuracy.

The original data set is split into three parts: D^{L} for training, D^{V} for validation, D^{U} for sampling. Therefore, all labeling work will be costed by the labeling of D^{L}, D^{V} and the sampled data. The pseudo-code of the proposed method is shown in Algorithm I. The implementation process of this method is shown in Fig. 4.

Algorithm I: Deep active learning method for steel plate surface defect classification

Input: initial training data pool D^{L}, validation data pool D^{V}, unlabeled data pool D^{U}, random initialized model parameters W', stopping criterion VA, number of sampled images for each class k.

Output: optimized model parameters W.

Repeat
  Make predictions on D^{U};
  The predicted labels are calculated by Eq. (2);
  The PDC of all samples is counted and grouped according to the predicted label;
  Make predictions on D^{L};
  The performance benchmark of the model on each label is calculated according to Eq. (4);
  The KLS value corresponding to each PDC is calculated by Eqs. (5) and (6);
  In each PDC group, k samples with the highest KLS value are selected.
  The samples are labeled, $D^{\mathrm{L}}$and D^{U} are updated;
  The model is trained with D^{L};
  D^{U} is used to verify the performance of the model and VA' is calculated;
Until $VA' > VA$.

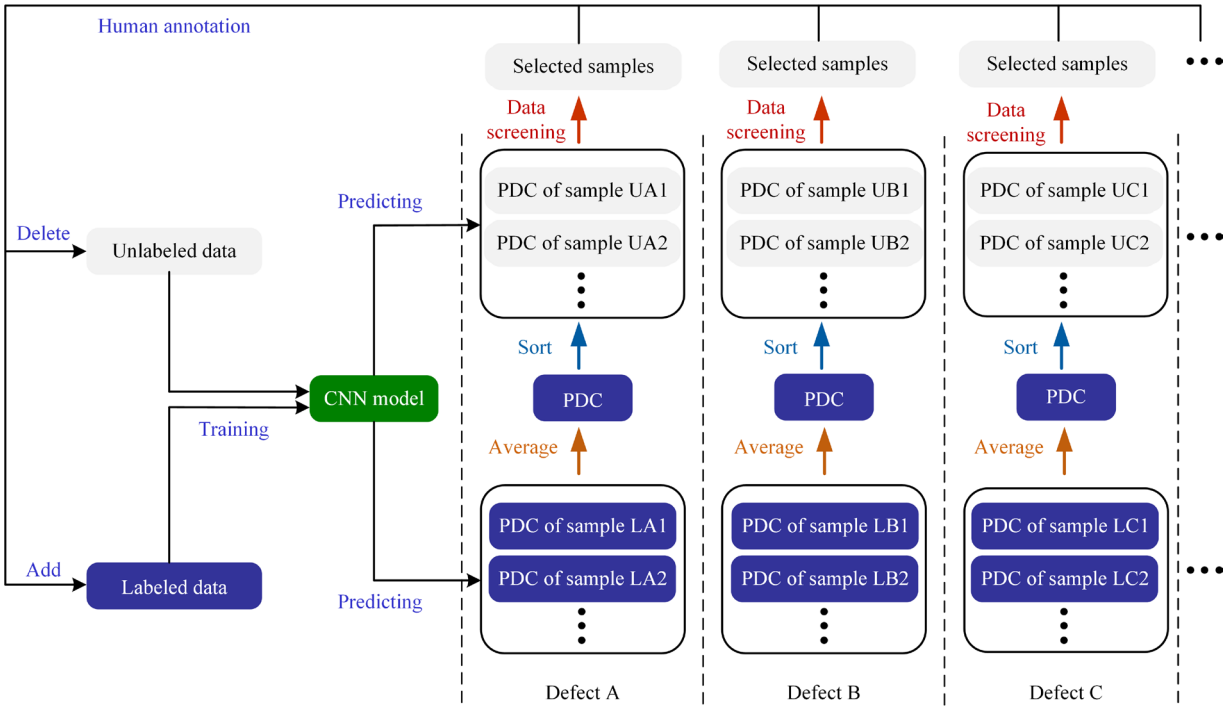

Fig. 4. *The implementation process of this method*

3 EXPERIMENTAL SETUP

The public hot-rolled steel strip database from Northeastern University (NEU-CLS) [26] is used to evaluate the effectiveness and applicability of the proposed method for the defect classification of a steel plate surface.

3.1 NEU-CLS Dataset

NEU-CLS is a standard high-quality database collecting the typical defects of hot-rolled steel plate surfaces. This dataset includes six types of defects: rolled-in scale, patch, crazing, pitted surface, inclusion, and scratches. There are 300 grayscale images for each type of defect (with a total number of 1800 images). The resolution of each image is 200×200 pixels. A selection of defect images is shown in Fig. 5.

3.2 Implementation Details

The experiments are performed on a work computer with Intel(R) Core i5-11400F CPU, 16GB memory and NVIDIA GeForce RTX 3060 GPU (12G memory). The dataset of NEU-CLS is divided into three parts, specifically 60 % for training, 20 % for validation, and 20 % for testing. The 5 % of training data (9 images per class) are sampled at random as initial labeled data pool D^{L}. The rest of the training data make up the unlabeled data pool D^{U}. Batch size is set to 8 and SGD optimizer is used with a momentum of 0.9. Early stopping with a patience of 10 is used for all experiments to stop the training process when overfitting begins. The results of the experiment are averaged over 5 runs. The detailed information of the designed network is shown in Table 1. The network takes 200×200×1 image as input and outputs 6 class predictions.

The experiment is mainly carried out in two parts.

1. The regularization effect of GPC is tested in a supervised learning condition. All training data are used to train the model and the learning rate is set to 0.005. TRC is mixed with dropout and geometric transformation-based data augmentation to form four methods as comparison. The dropout rate is set to 0.5. The geometric transformations include horizontal flipping, −30˚ to 30˚ rotations, random shearing and zooming. 0.1 of the width and height of the image .s shifted at random in the shifting operation. The shearing range is set to 0.2 and the zooming range is set to [0.8, 1.2]. The GPC with global average pooling and global max pooling are named as GAPC and GMPC in the following experiments, respectively.

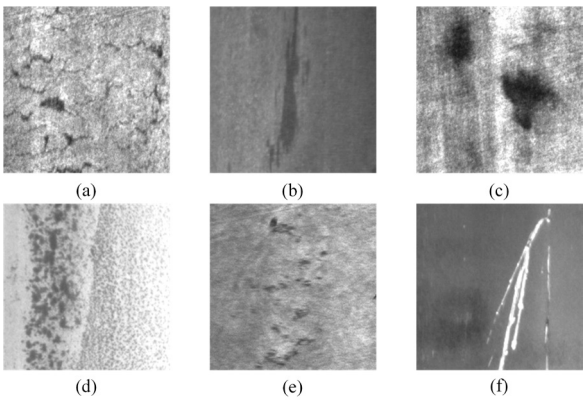

Fig. 5. *Images in NEU-CLS dataset; a) crazing, b) inclusion, c) patch, d) pitted surface, e) rolled-in scale, and f) scratch*

Table 1. *Detailed configuration of the designed network architecture*

Layer	Kernel size / Stride	Output size
Convolution + ReLu	3×3/2	100×100×32
Convolution + ReLu	3×3/2	50×50×32
Maxpool	2×2/1	25×25×32
Convolution + ReLu	3×3/2	13×13×64
Convolution + ReLu	3×3/1	13×13×64
Maxpool	2×2/1	6×6×64
Global average/Maxpool		1×1×64
Dropout 50 %		1×1×64
FC + Softmax		1×1×6

2. The effect of the proposed KLS sampling strategy in active learning is compared with the traditional uncertainty-based sampling method including LC, MS, and EN. Random Sampling (RS) is used as a performance base line, which discards all uncertainty strategies and randomly selects samples from D^{U} in each training cycle. The stopping criterion *VA* is set to 0.95. In the experiment, GPC is combined with dropout to further strengthen the regularization effect of the model. Meanwhile, inverse-time-decay is used to gradually decrease the learning rate. The decay strategy of the learning rate is defined as:

$$l_r = l_{ini} \times \frac{1}{1+\frac{d_r \times N_{epoch}}{d_s}}, \quad (7)$$

where l_{ini} is the initial learning rate. d_r is the decay rate and its value is set to 0.96. d_s is the decay step and its value is set to 162. N_{epoch} is the epoch number of the current training iteration, which is reset at the beginning of each iteration. By using learning rate decay, the initial learning rate can be set to a large value to speed up the model training and avoid local minima [48]. Therefore, in this part of the experiments, l_{ini} is set to 0.01.

The accuracy, precision, recall, and F1-score are used as the metrics to evaluate the classification performance of the proposed method comprehensively. After the model prediction, the defect image will be defined as one of four cases: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). The aforementioned metrics are defined as:

$$Accuracy = \frac{Num_{\mathrm{TP}} + Num_{\mathrm{TN}}}{Num_{\mathrm{TP}} + Num_{\mathrm{FP}} + Num_{\mathrm{TN}} + Num_{\mathrm{FN}}}, \quad (8)$$

$$Precision = \frac{Num_{\mathrm{TP}}}{Num_{\mathrm{TP}} + Num_{\mathrm{FP}}}, \quad (9)$$

$$Recall = \frac{Num_{\mathrm{TP}}}{Num_{\mathrm{TP}} + Num_{\mathrm{FN}}}, \quad (10)$$

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}. \quad (11)$$

where the Num_{TP}, Num_{FP}, Num_{TN}, Num_{FN} represent the number of defects that are defined as TP, FP, TN, FN, respectively.

4 RESULTS AND DISCUSSION

4.1 Comparison of Regularization Methods

As shown in Table 2, TRC without any regularization method has a short training time, but its classification accuracy is relatively low: only 91.4 % of the test samples are correctly classified. The method of combining TRC and data augmentation has the highest classification accuracy, reaching 96.4 %, but its training time is also relatively long, indicating that the augmentation process and the enlarged data set have a great impact on the model training time. The combination of TRC and dropout reduces the training time by about 20 % compared with data augmentation, but its classification accuracy is the lowest, only 90.8 %. The classification accuracy of the GPC (GAPC and GMPC) method proposed in this paper reaches 96.2 %, which is only 0.2 % lower than that of the data augmentation method, but the GPC method greatly reduces the training time. Especially, the GMPC method reduces the model training time by about 50 % compared with the data augmentation method, which greatly improves the training efficiency and enhances the generalization ability of the model in a short time.

Table 2. *The performance of GPC and TRC in supervised learning setting*

Methods	Accuracy [%]	Precision [%]	Recall [%]	F1-score [%]	Training time [s]
TRC	91.4	91.6	91.4	91.2	69.78
TRC+ Augmentation	96.4	96.6	96.4	96.4	102.89
TRC + Dropout	90.8	91.2	90.8	90.8	85.07
TRC + Augmentation + Dropout	95.6	95.6	95.6	95.6	142.37
GMPC	96.2	96.2	96.2	96.2	51.83
GAPC	96.2	96.2	96.2	96.2	62.31

4.2 Comparison of Sampling Strategies

As shown in Fig. 6, under the GAPC-based classifier, the classification performance of RS is better than traditional methods by more than 5% in a low label budget (<35 % of the dataset to be labeled). When the number of labeled data increases to more than 35 %, the performance of RS stagnates, and traditional methods start to catch up and finally outperform RS. Fig. 7 shows that traditional methods perform similarly under the GMPC-based classifier, and the classification performance of RS is the worst.

However, the proposed KLS method constantly outperforms RS and traditional methods in both classifier architectures. Fig. 7 shows that traditional methods perform similarly under the GMPC-based classifier, and the classification performance of RS is the worst. Taking the performance of the GAPC-based classifier as an example, KLS sampling strategy can achieve 91.8 % classification accuracy with 32 % of the data set to be labeled, which can reduce the label cost by more than 3 % compared with the traditional methods. Moreover, KLS can achieve 97 % classification accuracy with 44 % of the dataset to be labeled. Compared with the traditional uncertainty sampling method, KLS sampling strategy is more efficient for the use of labeled data.

4.3 Comparison of Classifier Performance

To determine the best classifier, the performances of GPC-based classifiers (GAPC and GMPC) are compared and analyzed. Fig. 8 shows that the gap

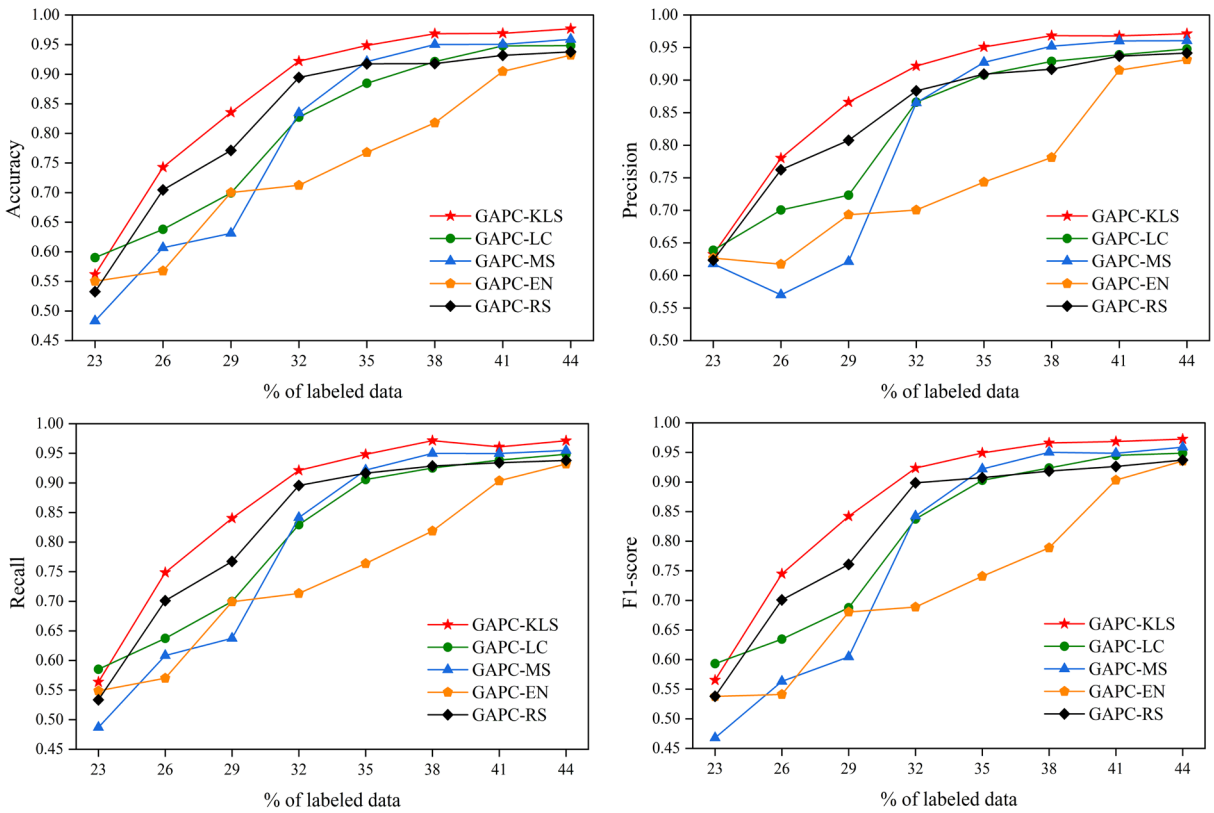

Fig. 6. *The performance of different sample strategies with GAPC*

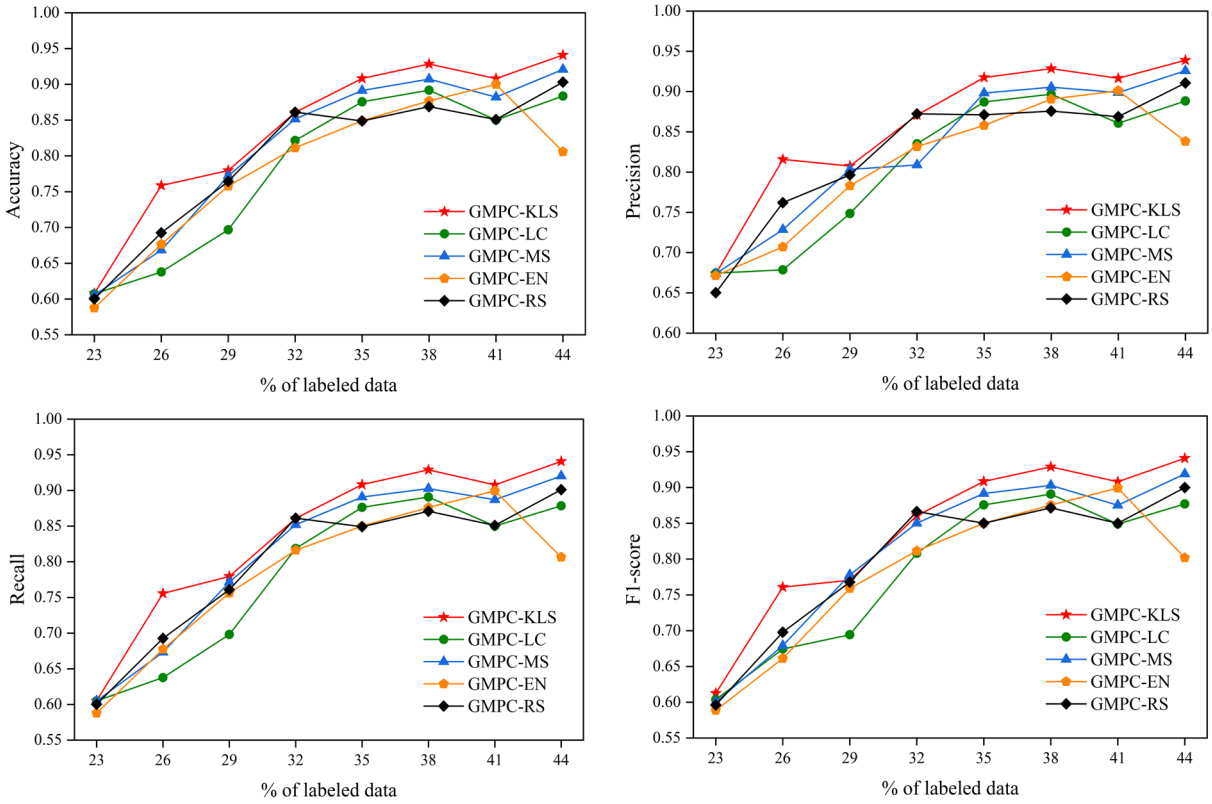

Fig. 7. *The performance of different sample strategies with GMPC*

between validation loss and training loss of the proposed model does not increase with the training time, indicating that the regularization effect of GPC is significant. Specifically, the accuracy and loss curves of the GMPC-based method converge faster in the first 150 epochs, but after 200 epochs, the performance of the GMPC-based method stagnates and starts to fluctuate in a wide range. The above situation does not appear in the GAPC-based method. Since the number of labeled samples increases with time in active learning, it can be inferred that the GMPC-based method is easy to converge only under the condition of a few labeled samples. When the number of labeled samples increases to a certain extent, the convergence becomes more difficult. Obviously, the GMPC-based method is more susceptible to labeled samples. As shown in Table 3, under the premise of achieving the same classification accuracy (90 %), the GAPC-based classifiers require 3 % to 15 % less labeled data than the GMPC-based classifiers except for EN. With 44 % of the dataset to be labeled (Table 4), the classification performance metrics of GAPC-based classifiers are 3 % to 14 % higher than that of GMPC-based classifiers. Therefore, considering the accuracy and model stability, the GAPC-based classifier is more suitable than the GMPC-based classifier for steel plate surface defect classification.

The accuracy and amount of labeled data on the NEU-CLS dataset of the proposed model and other Deep-learning based approaches are shown in Table 5. Compared with the end to end (ETE) method [49] and Supervised learning method mentioned in section 3.2, the proposed method achieves 2 % and 1.4 % higher accuracy on the NEU-CLS dataset, respectively, and the data that need to be labeled is reduced by 16 % and 36 %, respectively. Although the classification accuracy of SDC-SN-ELF+MRF method [13] is 0.3 % higher than that of the proposed method, the amount of labeled data required by this method is significantly higher (36 % higher than that of the proposed method). The results confirm that the proposed model can obtain good accuracy with less labeled data.

Additionally, the average training time of the proposed method (omitting the labeling time) is 180 s, and the final model size is 578.7 KB, which has application prospects in improving the efficiency of steel plate surface defect classification.

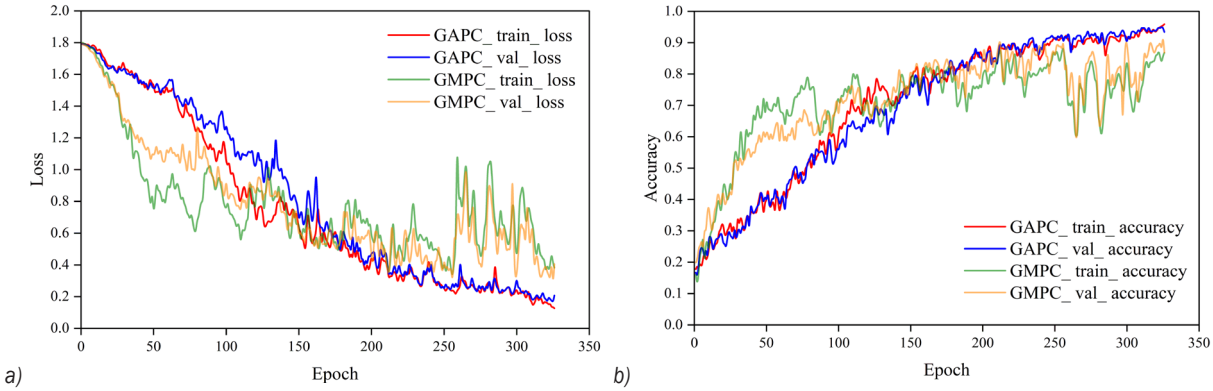

Fig. 8. *a) The loss, and b) accuracy of proposed GPC-based method over time*

Table 3. *Percentage of labeled samples needed to reach 90 % of corresponding performance metric*

Metric	GAPC					GMPC				
	KLS	LC	MS	EN	RS	KLS	LC	MS	EN	RS
Accuracy	32	35	35	41	35	35	50	38	41	44
Precision	32	35	35	41	32	35	50	38	41	44
Recall	32	35	35	41	35	35	50	38	41	44
F1-score	32	35	35	41	35	35	50	38	41	44

Table 4. *Performance score [%] achieved using labeled data that account for 44 % of the dataset*

Metric	GAPC					GMPC				
	KLS	LC	MS	EN	RS	KLS	LC	MS	EN	RS
Accuracy	97.8	94.8	95.5	93.3	93.8	93.8	88.0	91.8	80.6	90.2
Precision	97.9	94.8	95.8	93.3	94.0	93.8	88.8	92.4	83.4	91.0
Recall	97.8	94.8	95.5	93.3	93.8	93.8	88.0	91.8	80.6	90.2
F1-score	97.8	94.8	95.5	93.3	93.8	93.8	87.8	91.8	80.2	90.0

Table 5. *The accuracy and amount of labeled data on the NEU-CLS dataset of proposed model and other deep-learning based approaches*

Methods	Accuracy [%]	Training data [% of dataset]	Validation data [% of dataset]	Labeled data [% of dataset]
ETE [49]	95.8	60	-	60
SDC-SN-ELF + MRF [13]	98.1	80	-	80
Supervised learning method mentioned in section 3.2	96.4	60	20	80
Ours	97.8	24	20	44

4.4 Visualization of Defect Area Identification

Class activation graph (CAM) [50] and [51] is used to improve the interpretability of the proposed network model. The adopted model is obtained from the previous experiment, the network structure is GAPC-based neural network, and the training method is active learning method based on KLS. As shown in Fig. 9, the significant features that play a decisive role in the prediction of the network model are visualized in the heatmap. The warmer the color of the area in the heatmap, the more that area contributes to the model prediction. By superimposing the heatmap with

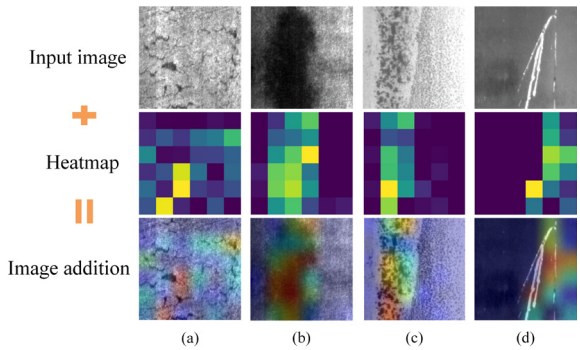

Fig. 9. *Defect area identification; a) crazing, b) patch, c) pitted surface, d) scratch*

the input image, it is found that the network focuses especially on the discriminative parts in the input images, which also proves the effectiveness of the proposed network model and learning method.

5 EXPERIMENTAL VERIFICATION

5.1 Dataset Preparation

As one of the raw materials commonly used in mechanical manufacturing, steel plate needs to be processed before it can be put into use. For example, in the manufacture of a linear guide plane, the steel plate usually needs to be milled. The defects on the surface of the steel plate after milling may have an impact on the positioning accuracy and service life of the linear guide plane. However, unlike hot-rolled steel plates, the probability of defects on the steel plates after processing is relatively small. Therefore, the number of defect samples is usually limited and belongs to the dataset with a small amount of data. This is the current situation of sample shortage in industry.

Therefore, to verify the applicability of the proposed method in industry, the steel plate milling experiment is carried out, and the surface defect dataset of processed steel plate is obtained. The parameters of milling processing are shown in Table 6. The defect images are collected on the image acquisition platform (VMC220) (Fig. 10). The defects on the surface of the steel plate after milling are pitted surface, scratch, and patch (Fig. 11). The number of collected images of pitted surface, scratch and patch defects are 200, 200 and 100, respectively. Due to the small number of original images, the data augmentation method of geometric transformation is used to amplify the image data, resulting in a total of 900 images, 300 for each type of defect. The resolution of each image is 200×200 pixels.

Table 6. *Experimental configurations*

Configurations	Parameters
Machine tool	VMC-C30
Workpiece	Steel S45C Steel S15C
Milling cutter	Kennametal 40A03RS45SE14EG
Spindle speed [r/min]	2000 2500 3000
Cutting depth [mm]	0.2 0.15 0.1
Feed per tooth [mm]	0.005 0.01 0.015

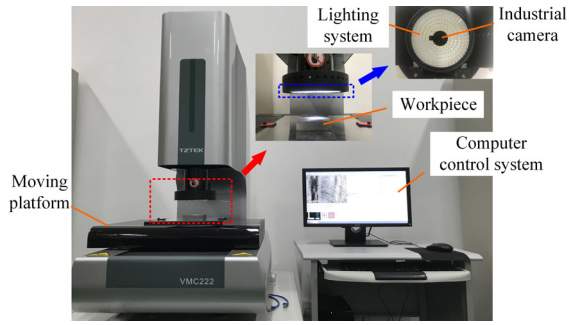

Fig. 10. *Image acquisition platform*

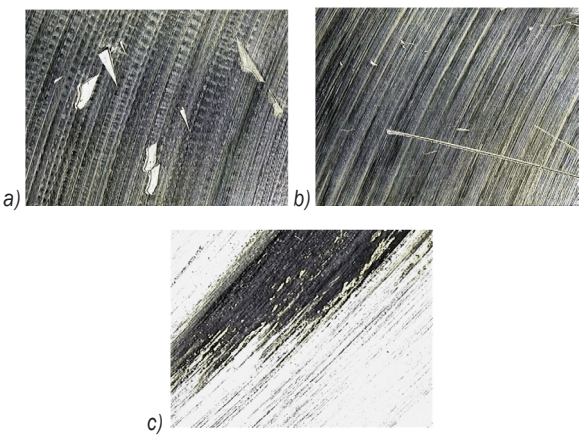

Fig. 11. *Images in milling steel surface defect dataset; a) pitted surface, b) scratch, and c) patch*

5.2 Experimental Setup

The main setups of the experiment have been shown in section 3.2. In this experiment, 60 % of the dataset is used for training, 20 % for validation and 20 % for testing. Then 10 % training data (18 images per class) are sampled at random as initial labeled data pool D^{L}. The rest of the training data make up the unlabeled data pool D^{U}.

5.3 Results and Discussion

Fig. 12 shows that the proposed GAPC-based network model still performs stably on the dataset with good regularization effect. As shown in Table 7 and Fig.13, when the amount of label data used by the model is 50 % of the dataset, the proposed KLS sampling method can achieve 92.3 % classification accuracy, while the maximum accuracy of other traditional methods is only about 80.4 %.

Hence, the experimental results indicate that the proposed method can still achieve more than 90 % classification accuracy when the number of samples is small, which reflects its application possibility in industry.

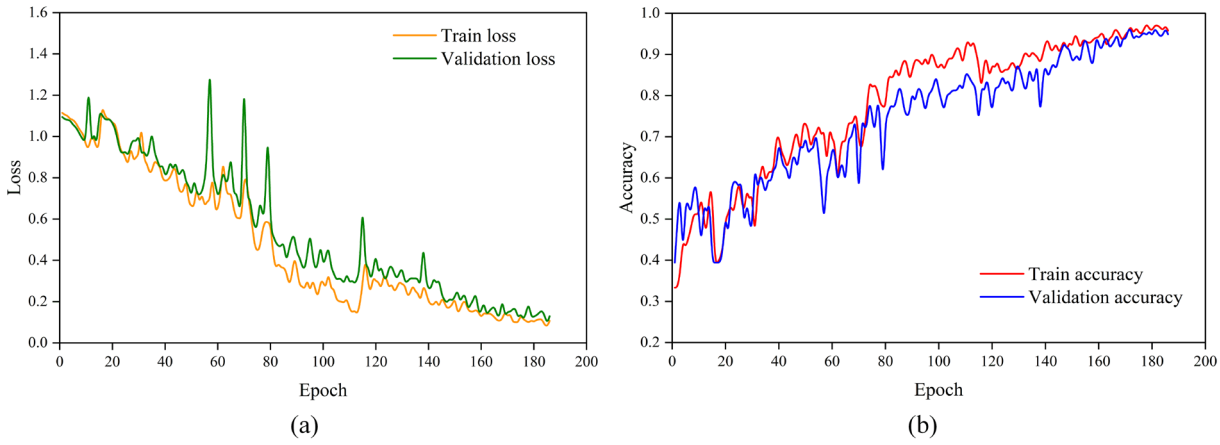

Fig. 12. *a) The loss, and b) accuracy of proposed method over time*

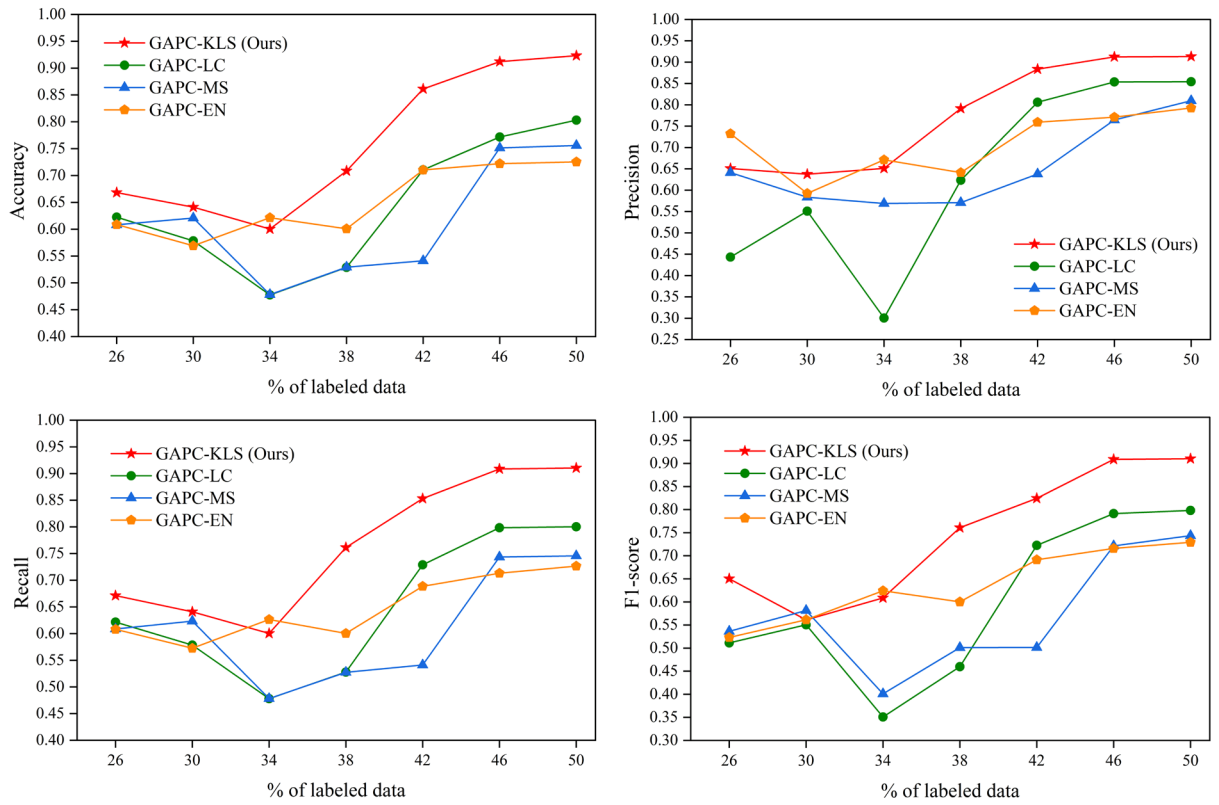

Fig. 13. *The performance of different sample strategies with GAPC*

Table 7. *Comparison results of different sample strategies with GAPC*

Metric	Ours [%]	LC [%]	MS [%]	EN [%]
Accuracy	92.3	80.4	75.0	72.5
Precision	91.5	85.2	81.2	79.0
Recall	91.0	80.0	74.6	72.5
F1-score	91.3	79.8	74.3	72.5

6 CONCLUSIONS

To reduce the labeling cost of steel plate surface defect classification in industrial production, a lightweight CNN model with strong regularization ability is designed, and an efficient deep active learning method is proposed by combining it with the KLS strategy. The specific conclusions are as follows:

1. The GPC-based classifier can greatly reduce the training time while maintaining the same performance as the traditional classifier in steel plate surface defect classification.
2. A GPC-based lightweight convolutional neural network model is proposed. The result indicates that the performance of the GAPC-based network model is more stable than that of the GMPC-based network model.
3. The labeling cost can be significantly reduced by using the KLS strategy as the uncertainty sampling method. Comparative analysis shows that the GAPC-KLS model only needs 44 % labeled data to achieve 97.8 % classification accuracy, and its performance is optimal. Meanwhile, this model can still achieve 92.3 % classification accuracy with 50 % labeled data on the milling steel surface defect dataset. Therefore, the proposed method can achieve classification accuracy (≥92 %) with limited labeled data (≤50 % of the dataset to be labeled) on both NEU-CLS and milling steel surface defect datasets.

To further improve the classification efficiency, the subsequent research will focus on the optimization of the convolutional base to reduce the training time and improve the training efficiency while ensuring the quality of feature extraction. The proposed method can provide a reference for steel plate production enterprises to reduce the cost of surface defect image annotation. In addition, this method may provide a new idea for efficient classification of other surface defects in industry.

7 ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (52175254); Postgraduate Scientific Research Innovation Project of Hunan Province (CX20220603, CX20230550).

8 REFERENCES

[1] Zhu, M., Lu, X., Li, H., Cao, H., Wu, F. (2023). Applicability analysis of nickel steel plate friction coefficient model based on fractal theory. *Coatings*, vol. 13, 1096, **DOI:10.3390/coatings13061096**.

[2] Mohtaram, Y.F., Kahnamouei, J.T., Shariati, M., Behjat, B. (2012). Experimental and numerical investigation of buckling in rectangular steel plates with groove-shaped cutouts. *Journal of Zhejiang University SCIENCE A*, vol. 13, p. 469-480, **DOI:10.1631/jzus.A1100226**.

[3] Wang, Y., Shen, X.L., Wu, K., Huang, M.Q. (2022). Corrosion grade recognition for weathering steel plate based on a convolutional neural network. *Measurement Science and Technology*, vol. 33, 095014, **DOI:10.1088/1361-6501/ac7034**.

[4] Dung, C.V., Sekiya, H., Hirano, S., Okatani, T., Miki, C. (2019). A vision-based method for crack detection in gusset plate welded joints of steel bridges using deep convolutional neural networks. *Automation in Construction*, vol. 102, p. 217-229, **DOI:10.1016/j.autcon.2019.02.013**.

[5] Jiménez-Peña, C., Goulas, C., Preußner, J., Debruyne, D. (2020). Failure mechanisms of mechanically and thermally produced holes in high-strength low-alloy steel plates subjected to fatigue loading. *Metals*, vol. 10, no. 3, 318, **DOI:10.3390/met10030318**.

[6] Park, C.Y., Kim, J.W., Kim, B., Lee, J. (2020). Prediction for manufacturing factors in a steel plate rolling smart factory using data clustering-based machine learning. *IEEE Access*, vol. 8, p. 60890-60905, **DOI:10.1109/ACCESS.2020.2983188**.

[7] Yoshioka, S., Fujii, A., Tohara, M., Gotoh, Y. (2021). Proposed inspection method for opposite-side defect in steel plate using synthetic magnetic field with high and low excitation frequencies. *Sensors and Materials*, vol. 33, no. 7, p. 2511-2520, **DOI:10.18494/SAM.2021.3380**.

[8] Wang, G., Xiao, Q., Gao, Z.H., Li, W.H., Jia, L., Liang, C., Yu, X. (2022). Multifrequency AC magnetic flux leakage testing for the detection of surface and backside defects in thick steel plates. *IEEE Magnetics Letters*, vol. 13, 8102105, **DOI:10.1109/LMAG.2022.3142717**.

[9] Zheng, X., Zheng, S., Kong, Y. G., Chen, J. (2021). Recent advances in surface defect inspection of industrial products using deep learning techniques. *The International Journal of Advanced Manufacturing Technology*, vol. 113, p. 35-58, **DOI:10.1007/s00170-021-06592-8**.

[10] Bhatt, P.M., Malhan, R.K., Rajendran, P., Shah, B.C., Gupta, S.K. (2021). Image-based surface defect detection using deep learning: a review. *Journal of Computing and Information Science in Engineering*, vol. 21, no. 4, 040801, **DOI:10.1115/1.4049535**.

[11] Chen, Y.J., Ding, Y.Y., Zhao F., Zhang, E.H., Wu, Z.N., Shao, L. (2021). Surface defect detection methods for industrial products: a review. *Applied Sciences*, vol. 11, no. 16, 7657, **DOI:10.3390/app11167657**.

[12] Lecun, Y., Bottou, L., Bengio, Y., Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, vol. 86, no. 11, p. 2278-2324, **DOI10.1109/5.726791**.

[13] Zhou, S.Y., Chen, Y.P., Zhang, D.L., Xie, J.M., Zhou, Y.F. (2017). Classification of surface defects on steel sheet using convolutional neural networks. *Materials and Technologies*, vol. 51, no. 1, p. 123-131, **DOI:10.17222/mit.2015.335**.

[14] He, X., Wang, T.Q., Wu, K.X., Liu, H.H. (2021). Automatic defects detection and classification of low carbon steel WAAM products using improved remanence/magneto-optical imaging and cost-sensitive convolutional neural network. *Measurement*, vol. 173, 108633, **DOI:10.1016/j.measurement.2020.108633**.

[15] Pan, S.J.L., Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, p. 1345-1359, **DOI:10.1109/TKDE.2009.191**.

[16] Dhamala, J., Bajracharya, P., Arevalo, H.J., Sapp, J.L., Horácek, B. M., Wu, K.C., Trayanova, N.A., Wang, L.W. (2020). Embedding high-dimensional Bayesian optimization via generative modeling: parameter personalization of cardiac electro-physiological models. *Medical Image Analysis*, vol. 62, 101670, **DOI:10.1016/j.media.2020.101670**.

[17] Wang, Q.X., Yang, R.H., Wu, C.J., Liu, Y. (2021). An effective defect detection method based on improved generative adversarial networks (iGAN) for machined surfaces. *Journal of Manufacturing Processes*, vol. 65, p. 373-381, **DOI:10.1016/j.jmapro.2021.03.053**.

[18] Lee, D.H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop: Challenges in Representation Learning*, p. 1-7.

[19] Odena, A. (2016). *Semi-supervised learning with generative adversarial networks*. arXiv:1606.01583, **DOI:10.48550/arXiv.1606.01583**.

[20] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X. (2016). *Improved techniques for training GANs*. arXiv:1606.03498, **DOI:10.48550/arXiv.1606.03498**.

[21] Weigl, E., Heidl, W., Lughofer, E., Radauer, T., Eitzinger, C. (2016). On improving performance of surface inspection systems by online active learning and flexible classifier updates. *Machine Vision and Applications*, vol. 27, p. 103-127, **DOI:10.1007/s00138-015-0731-9**.

[22] Fu, G.Z., Sun, P.Z., Zhu, W.B., Yang, J.X., Cao, Y.L., Yang, M.Y., Cao, Y.P. (2019). A deep-learning-based approach for fast and robust steel surface defects classification. *Optics and Lasers in Engineering*, vol. 121, p. 397-405, **DOI:10.1016/j.optlaseng.2019.05.005**.

[23] Yang, Y.T., Yang, R.Z., Pan, L.H., Ma, J.X., Zhu, Y.S., Diao, T., Zhang, L. (2020). A lightweight deep learning algorithm for inspection of laser welding defects on safety vent of power battery. *Computers in Industry*, vol. 123, 103306, **DOI:10.1016/j.compind.2020.103306**.

[24] Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K. (2016). *Squeeze Net: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size*, arXiv:1602.07360, **DOI:10.48550/arXiv.1602.07360**.

[25] He, D., Xu, K., Wang, D.D. (2019). Design of multi-scale receptive field convolutional neural network for surface inspection of hot rolled steels. *Image and Vision Computing*, vol. 89, p. 12-20, **DOI:10.1016/j.imavis.2019.06.008**.

[26] Song, K.C., Yan, Y.H. (2013). A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Applied Surface Science*, vol. 285, p. 858-864, **DOI:10.1016/j.apsusc.2013.09.002**.

[27] Kingma, D.P., Welling, M. (2013). *Auto-encoding variational Bayes*. arXiv:1312.6114, **DOI:10.48550/arXiv.1312.6114**.

[28] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). *Generative adversarial networks*, arXiv:1406.2661, **DOI:10.48550/arXiv.1406.2661**.

[29] Yun, J.P., Shin, W.C., Koo, G., Kim, M.S., Lee, C., Lee, S.J. (2020). Automated defect inspection system for metal surfaces based on deep learning and data augmentation. *Journal of Manufacturing Systems*, vol. 55, p. 317-324, **DOI:10.1016/j.jmsy.2020.03.009**.

[30] Sohn, K., Lee, H., Yan, X.C. (2015). Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems*, vol. 28, p. 3483-3491.

[31] Tang, W.Q., Yang, Q., Xiong, K.X., Yan, W.J. (2020). Deep learning based automatic defect identification of photovoltaic module using electroluminescence images. *Solar Energy*, vol. 201, p. 453-460, **DOI:10.1016/j.solener.2020.03.049**.

[32] Tran, T., Do, T., Reid, I., Carneiro, G. (2019). Bayesian generative active deep learning. *International Conference on Machine Learning*, p. 6295-6304.

[33] Gao, Y.P., Gao, L., Li, X.Y., Yan, X.G. (2020). A semi-supervised convolutional neural network-based method for steel surface defect recognition. *Robotics and Computer Integrated Manufacturing*, vol. 61, 101825, **DOI:10.1016/j.rcim.2019.101825**.

[34] He, D., Xu, K., Zhou, P., Zhou, D.D. (2019). Surface defect classification of steels with a new semi-supervised learning method. *Optics and Lasers in Engineering*, vol. 117, p. 40-48, **DOI:10.1016/j.optlaseng.2019.01.011**.

[35] He, Y., Song, K.C., Dong, H.W., Yan, Y.H. (2019). Semi-supervised defect classification of steel surface based on multi-training and generative adversarial network. *Optics and Lasers in Engineering*, vol. 122, p. 294-302, **DOI:10.1016/j.optlaseng.2019.06.020**.

[36] Masci, J., Meier, U., Ciresan¸ D., Schmidhuber, J. (2011). Stacked convolutional auto-encoders for hierarchical feature extraction. *International Conference on Artificial Neural Networks*, p. 52-59, **DOI:10.1007/978-3-642-21735-7_7**.

[37] Yang, L., Zhang, Y.Z., Chen, J.X., Zhang, S.Y., Chen, D.Z. (2017). Suggestive annotation: A deep active learning framework for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Computer Science*, p. 399-407, **DOI:10.1007/978-3-319-66179-7_46**.

[38] Huang, Z.J., Li, F.M., Luan, X.D., Cai, Z.W. (2020). A weakly supervised method for mud detection in ores based on deep active learning. *Mathematical Problems in Engineering*, vol. 2020, no. 1, 3510313, **DOI:10.1155/2020/3510313**.

[39] Lv, X.M., Duan, F.J., Jiang J.J., Fu, X., Gan, L. (2020). Deep active learning for surface defect detection. *Sensors*, vol. 20, no. 6, 1650, **DOI:10.3390/s20061650**.

[40] Wang, K.Z., Zhang, D.Y., Li, Y., Zhang, R.M., Lin, L. (2017). Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, p. 2591-2600, **DOI:10.1109/TCSVT.2016.2589879**.

[41] Luo, C., Yu, L.J., Yan, J.X., Li, Z.W., Ren, P., Bai, X., Yang, E.F., Liu, Y.H. (2021). Autonomous detection of damage to multiple steel surfaces from 360° panoramas using deep neural networks. *Computer-Aided Civil and Infrastructure Engineering*, vol. 36, no. 12, p. 1585-1599, **DOI:10.1111/mice.12686**.

[42] Mao, W.S., Li, L.S., Tao, Y.F., Zhou, W.Y. (2023). Surface defect image classification of lithium battery pole piece based on deep learning. IEICE Transactions on Information and

Systems, vol. E106.D, no. 9, p. 1546-1555, **DOI:10.1587/transinf.2023EDP7058**.

[43] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, p. 1097-1105.

[44] Simonyan, K., Zisserman A. (2015). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, **DOI:10.48550/arXiv.1409.1556**.

[45] Szegedy, C., Liu, W., Jia, Y.Q., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2015). Going deeper with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition*, p. 1-9, **DOI:10.1109/CVPR.2015.7298594**.

[46] He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, p. 770-778, **DOI:10.1109/CVPR.2016.90**.

[47] Ponti, M., Kittler, J., Riva, M., Campos, T.D., Zor, C. (2017). A decision cognizant Kullback-Leibler divergence. *Pattern Recognition*, vol. 61, p. 470-478, **DOI:10.1016/j.patcog.2016.08.018**.

[48] You, K.C., Long, M.S., Wang, J.M., Jordan, M.I. (2019). How does learning rate decay help modern neural networks? arXiv:1908.01878, **DOI:10.48550/arXiv.1908.01878**.

[49] Yi, L., Li, G.Y., Jiang, M.M. (2017). An end-to-end steel strip surface defects recognition system based on convolutional neural networks. *Steel Research International*, vol. 88, no. 2, p. 1600068, **DOI:10.1002/srin.201600068**.

[50] Wang, X.Q., Gu, Y. (2022). Classification of macular abnormalities using a lightweight CNN-SVM framework. *Measurement Science and Technology*, vol. 33, 065702, **DOI:10.1088/1361-6501/ac5876**.

[51] Lee, S.Y., Tama, B.A., Moon, S.J., Lee, S. (2019). Steel surface defect diagnostics using deep convolutional neural network and class activation map. *Applied Sciences*, vol. 9, no. 24, 5449, **DOI:10.3390/app9245449**.