Computer Modeling in
Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2024.056621

ARTICLE

Steel Surface Defect Recognition in Smart Manufacturing Using Deep Ensemble Transfer Learning-Based Techniques

Tajmal Hussain and Jongwon Seok*

Department of Information and Communication Engineering, Changwon National University,
Changwon, 51140, Republic of Korea

*Corresponding Author: Jongwon Seok. Email: jwseok@changwon.ac.kr

Received: 26 July 2024 Accepted: 21 October 2024 Published: 17 December 2024

ABSTRACT

Smart manufacturing and Industry 4.0 are transforming traditional manufacturing processes by utilizing innovative technologies such as the artificial intelligence (AI) and internet of things (IoT) to enhance efficiency, reduce costs, and ensure product quality. In light of the recent advancement of Industry 4.0, identifying defects has become important for ensuring the quality of products during the manufacturing process. In this research, we present an ensemble methodology for accurately classifying hot rolled steel surface defects by combining the strengths of four pre-trained convolutional neural network (CNN) architectures: VGG16, VGG19, Xception, and Mobile-Net V2, compensating for their individual weaknesses. We evaluated our methodology on the Xsteel surface defect dataset (XSDD), which comprises seven different classes. The ensemble methodology integrated the predictions of individual models through two methods: model averaging and weighted averaging. Our evaluation showed that the model averaging ensemble achieved an accuracy of 98.89%, a recall of 98.92%, a precision of 99.05%, and an F1-score of 98.97%, while the weighted averaging ensemble reached an accuracy of 99.72%, a recall of 99.74%, a precision of 99.67%, and an F1-score of 99.70%. The proposed weighted averaging ensemble model outperformed the model averaging method and the individual models in detecting defects in terms of accuracy, recall, precision, and F1-score. Comparative analysis with recent studies also showed the superior performance of our methodology.

KEYWORDS

Smart manufacturing; CNN; steel defects; ensemble models

1 Introduction

In the domain of Industry 4.0 and the move towards smart factories, hot rolled steel strips are one of the primary products of the manufacturing industry, extensively used in the production of home appliances, shipbuilding, automobiles, aircraft, manufacturing industrial goods, and so on [1]. Hot rolled steel is mainly composed of elements like carbon, iron, sulfur, manganese, silicon, and phosphorus. The production process of hot rolled steel strips involves heating the steel above its recrystallization temperature, followed by rolling to achieve the desired thickness and shape. The mechanical properties and surface quality of the steel are influenced by both the chemical composition and the production process [2]. Surface quality is a critical measure of competitiveness in the strip steel manufacturing industry. During the production of hot rolled steel strips, surface defects are unavoidable due to various factors, such as equipment fatigue and human error. These defects not only weaken steel's fatigue resistance but also affect appearance. Despite efforts to improve the manufacturing process, these imperfections cannot be entirely eliminated [1,3,4]. It can result in customer rejection of the product, leading to significant financial losses for the manufacturing industry [5]. Therefore, accurately identifying surface defects in hot-rolled strips is crucial to maintaining product quality and ensuring the integrity of the steel surface. As a result, the development and deployment of effective systems for surface defect detection have become essential in the manufacturing process.

Traditionally, humans inspect surface defects by visually examining the items with their eyes, which is a time-consuming and insufficient method. However, automatic steel surface inspection systems have now become the standard in the industry, offering an alternative to manual visual inspections [5]. As industrial automation has advanced in steel production, defect detection systems for steel surfaces have become common in the manufacturing industry. This has significantly boosted the productivity of steel strip manufacturing. Today, numerous companies (Sipar, Matra, EES, Codnex, Parsytec, Siemens-VAI, etc.) produce these automatic inspection systems for steel surfaces [6].

Recently, numerous efforts have been made to automatically detect and classify surface defects on steel strips, providing guidelines and suggesting suitable directions for future research [1,5]. Existing techniques utilize individual, models to classify steel surface defects. The performance of these models completely depends on the specific model employed. On the other hand, an ensemble model, which combines several base learners, can achieve higher accuracy. By merging different models, the ensemble model takes advantage of the strengths of each model while minimizing their weaknesses [7]. Moreover, an ensemble model can reduce the problem of overfitting due to the diversity of its baseline models [8]. For instance, if one model tends to overfit, the ensemble can adjust by giving less weight to its predictions.

The ensemble method integrates multiple base learners to form a more powerful learner because each base learner has a diverse architecture, enabling them to recognize different patterns within the same data. When one model incorrectly learns a pattern, another model might classify it correctly, providing diverse perspectives on the same data. Consequently, merging predictions from various learners can lead to improved accuracy and better overall predictions compared to individual learners. This strategy has shown good results when compared to previous methods.

The primary contributions of this study are:

- A weighted averaging ensemble model is proposed for the classification of steel surface defects by using an optimal combination of weights that are assigned to four base learners, namely VGG16, VGG19, Xception, and Mobile-Net V2.
- The proposed methodology is evaluated on the XSDD steel surface defect dataset, which has been expanded through offline augmentation from 1360 to 3630 sample images.
- An evaluation was conducted on different weight combinations to identify the optimal weights that would improve performance metrics.
- The proposed ensemble model results are compared to the individual model and other recent research in order to show the performance of the proposed methodology in comparison to other available methodologies.

The remaining structure of the article is as follows: Section 2 provides a review of literature in the field of steel surface defect detection and recognition, followed by an outline of the methodology in

Section 3. Section 4 presents the results and discussion. The paper concludes with Section 5, which summarizes the findings and proposes areas for future investigation.

2 Literature Review

2.1 Machine Learning-Based Methods

Researchers have conducted several studies to identify steel surface defects and have proposed various methodologies in the field of machine learning (ML). Song et al. introduced the adjacent evaluation completed local binary patterns (LBP) descriptor with a support vector machine (SVM) classifier, which enhances robustness against noise and intra-class variations by utilizing an adjacent evaluation window to refine the threshold scheme of traditional LBP [3]. Hu et al. suggested that support vector machines employ a Gaussian radial basis kernel, parameterizes via cross-validation, and employ a one-*vs.*-one strategy to classify defects in steel surfaces. They extracted features, including grayscale, geometric, and shape features, by integrating defect images with their corresponding binary representations [9].

Xiao et al. introduced a classifier called BYEC that integrates a bayes kernel. They began by introducing a rich set of features aimed at capturing detailed defect information. Using these features, they constructed multiple SVMs with random subsets. They then trained a Bayes classifier and fused its results with the sub-SVMs to create a classifier [10]. Liu et al. investigated the performance of various feature extraction methods and classifiers for identifying surface defects in cold-rolled steel strips. Their evaluation included methods such as scale invariant feature transform, speeded-up robust features, and local binary patterns, along with classifiers including back propagation networks, support vector machines, and extreme learning machines (ELM). Among these, the hybrid combination of LBP and ELM for classification stands out for its efficiency [11].

Mentouri et al. presented the Dual Cross Pattern (DCP) feature descriptor to enhance defect classification on the NEU defect database. This technique is highly suitable for capturing variations in size and orientation within the images. By utilizing k-nearest neighbors (KNN) and multi-class support vector machines (SVM) with optimized parameters as classifiers, they found that KNN with three neighbors and Euclidean distance provided superior performance. They integrated LDA and PCA methods for data reduction, along with KNN, to improve the classification results [12].

The support vector hyper-spheres with insensitivity to noise classifier introduced by Gong et al. [13] is notable for its incorporation of pinball loss to reduce noise sensitivity and local within-class sample density weighting to enhance classification performance. This paper proposes a strategy that combines the gray-level co-occurrence matrix with the discrete shearlet transform for classifying defects in steel and achieves an accuracy of 96.00% with this hybrid technique [14].

The traditional ML methods have shown effectiveness in classifying steel strip defects; however, some traditional methods' performance depends on human feature extraction and expert knowledge, which often results in low performance. Traditional ML algorithms struggle to accurately handle large amounts of complex data. Moreover, applying these algorithms to new detection tasks often necessitates redesigning them, causing challenges in transferring solutions to similar problems.

2.2 Deep Learning-Based Methods

Deep learning (DL) has become very popular in recent years. As a result, many researchers have shifted from traditional ML methods to DL methods for surface defect identification and classification in the manufacturing industry due to their powerful feature extraction capabilities that do not rely on manual processes or human involvement. DL, particularly through Convolutional Neural Networks (CNNs), has shown significant advancements in the steel manufacturing industry.

Feng et al. [1] introduced a method that uses ResNet-50 with FCA-Net and the convolutional block attention module to classify a steel surface defect dataset, achieving an accuracy of 93.87%. Lin et al. developed a method for classifying steel defects using an augmentation algorithm called Random-CutMix and an enhanced Mobile-Net architecture known as SP-Mobile-Net. SP-Mobile-Net integrates the inverse residual module with a channel shuffle mechanism and a pyramid split attention module, which enhances information flow, boosting the model's representation capability and performance. This approach achieved an accuracy of 95.97% [15]. In another paper, Feng et al. [16] introduced a method that integrates the RepVGG with the spatial attention method to classify the XSDD steel surface defect dataset into seven categories, with an accuracy of 95.11%.

Zheng et al. [17] proposed a feature extraction approach that combines legendre multi-wavelet transform and auto-encoder (AE) networks to detect steel defects. Their method employs finite element approximation theory to make use of Legendre wavelet bases, which can handle a wide range of regularities and moments, ensuring the correct representation of complicated defect geometries without information loss. The approach comprises collecting defect features from the LWT frequency domain using statistical and texture parameters, then dimensionality reduction and feature selection via an AE network. To examine generalization, classifiers such as SVM and backpropagation neural networks were used, achieving an accuracy of 95.37% on the XSDD dataset.

Wen et al. introduced a unique multiscale, multi-attention CNN for surface defect detection. First, they created a multiscale CNN that uses features from multiple layers to detect defects of varying sizes. Second, they developed a new multi-attention module that generates compact attention maps to enable the model to focus on minor defects. Finally, they applied the multiscale and multi-attention-based technique to fine-grained SDD tasks, achieving 99.59% accuracy on the XSDD dataset [18]. Hao et al. first suggested a unique WGAN model for generating new surface defect images from random noise, increasing the sample size from 1360 to 3773. These generated images are then used to train classification systems. Next, they suggested a Multi-SE-ResNet34 model with an attention mechanism for defect detection, achieving an accuracy of 99.20% [19].

DL-based steel defect classification methods outperformed traditional ML methods. However, existing methods utilize individual models to classify steel surface defects. The performance of these models completely depends on the specific model employed. In contrast, an ensemble model, which combines several base learners, can achieve higher accuracy. Moreover, an ensemble model can more effectively address the overfitting issue and reduce variance by using multiple parameters from different models. So, in this study, we proposed a weighted averaging ensemble model that combines four different base-learners and gives predictions.

3 Materials and Proposed Methodology

This section provides an overview of the materials and methodology used in this paper and describes the data gathering and augmentation techniques, the creation of an ensemble model, the experimental setting, and the performance evaluation measures used in this paper. Fig. 1 shows the methodology of the system.

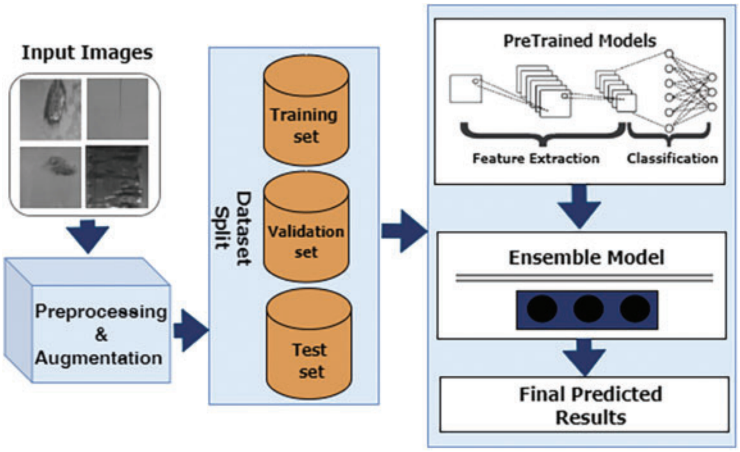
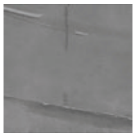
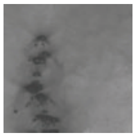
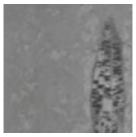


Figure 1: The methodology of system

3.1 Input Dataset

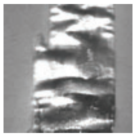
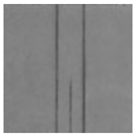
The XSDD dataset of hot-rolled steel strip designed for the classification and detection of steel surface defects in academic research [16]. This dataset comprises 1360 images of various steel defects, categorized into several types: 203 images of finishing roll printing, 122 images of iron sheet ash, 63 images of plate system oxide scale, 203 images of temperature system oxide scale, 397 images of red iron sheets, 238 images of slag inclusions, and 134 images of surface scratches. To meet model requirements, all images have been resized to 224 × 224 × 3 pixels and normalized to pixel values between 0 and 1. Table 1 provides a detailed description of each defect type in the dataset.

Table 1: Types of surface defects in XSDD dataset

Defect types	Images	Description of defects types
Finishing roll printing		This defect occurs due to slippage in the work roll and the support roll, which results in dot-shaped and short strip-shaped damage on the surface of the work roll.
Iron sheet ash		This defect occurs when metal dust, water, oil, and other contaminants accumulate on rolling mill equipment over time. It appears as comet-shaped metal particles with visible black oil residue.
Oxide scale plate system		This defect arises from several factors during high-temperature, high-speed rolling, including roller table dead rolls. It typically appears in a fixed location and resembles scratches.

(Continued)

Table 1 (continued)

Defect types	Images	Description of defects types
Oxide scale temperature		This defect may form due to incorrect use of stand water and excessive temperature during rough rolling, its appearance is loose or sandy.
Red iron		This defect primarily results from high silicon content in the steel and elevated slab heating temperatures. It typically appears reddish-brown and shapes as dots, strips, or flakes spread across the entire strip.
Slag inclusion		Inclusion defects, often found during the slab continuous casting process, appear as visible black non-metallic substances that distinctly contrast with the surrounding metal.
Surface scratches		This defect forms due to projections in the hot rolling area, passive rolls, dead rolls, and surface friction on the steel strip. It appears as straight lines and grooves on the steel strip surface.

Data Augmentation

The process of artificially growing a dataset in order to improve a model's capacity for generalization and reduce overfitting is known as data augmentation [20]. In this study, offline data augmentation is applied as a preprocessing step to balance and increase the size of dataset. The augmentation process is performed by python PIL library, which includes horizontal and vertical flips, as well as rotations of 5°, 10°, 15°, and 20°. The selection of data augmentation techniques (flipping and rotation) is based on the optimal combination of augmentation techniques for various surface defect datasets [21,22]. These techniques improve the model's robustness. The XSDD dataset was increased from 1360 to 3630 images. Fig. 2 shows some examples of the augmented images. Fig. 3 provides details of the XSDD dataset before and after augmentation.

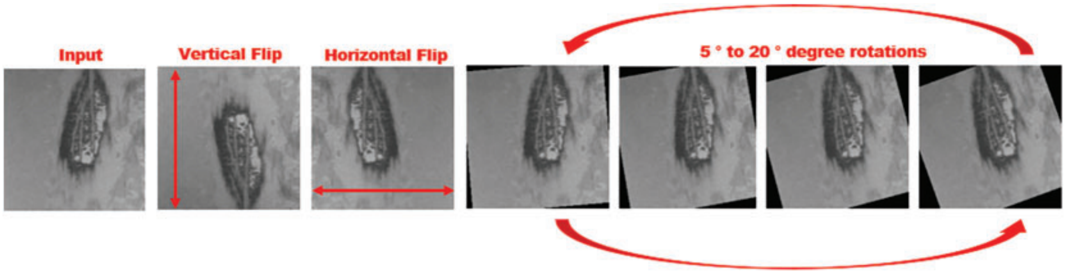


Figure 2: Sample of offline augmented images

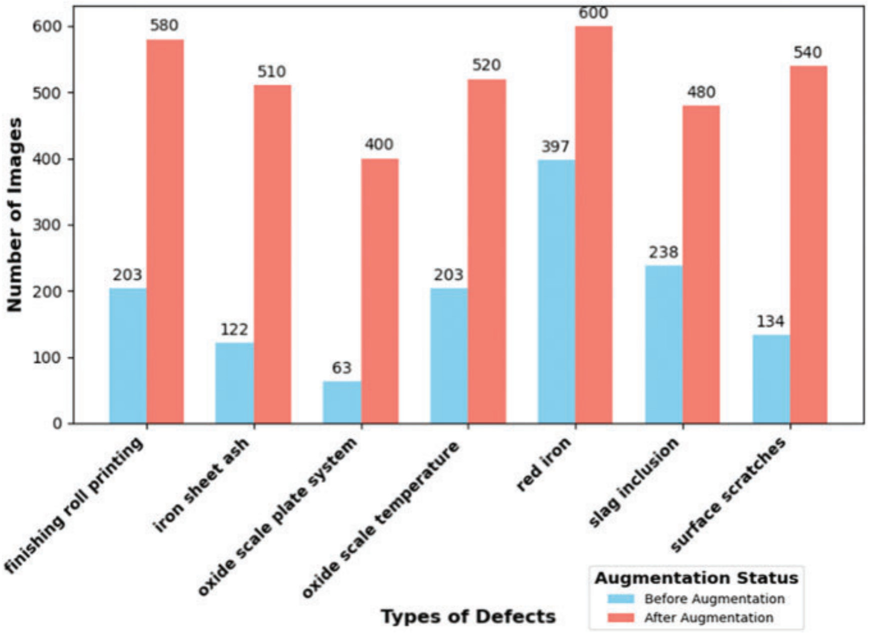


Figure 3: Number of images before and after augmentation

3.2 Methodology

In this research, we proposed a novel ensemble model designed to enhance the classification accuracy of steel surface defects. The following subsections detail the selection process, training strategy of transfer learning-based base learners, and the development of the proposed deep ensemble model.

3.2.1 Selection and Training of Transfer Learning-Based Base Learner

Creating an efficient DL model is a difficult task when your dataset has a limited number of samples. In this situation, transfer learning is a good choice. Transfer learning uses knowledge from one domain to solve tasks in another related domain. This technique benefits from the use of pre-learned feature maps, which avoid the requirement to train a model from start on a large dataset. In such cases, transfer learning assists in developing a computationally efficient model in less time [23,24]. The selection of a base-learner (BL) may differ from one problem to the next, but the goal is to select the best-suited base learner for the particular problem. In this research, we selected four pre-trained transfer learning-based Convolutional Neural Network (CNN) architectures: VGG16 [5], VGG19 [25,26], Xception [27], and Mobile-Net V2 [15,28], due to their diverse architectures and higher accuracy in classifying steel surface defects. These diverse architectures have varying abilities to generalize the given distribution.

All BLs were first trained on the ImageNet dataset [29]. For transfer learning purposes, the first layers are not trained because they simply learn the basic features of the dataset. To learn the unique features of the dataset, only the top layers are trained. Therefore, the initial layers in these CNNs were kept unchanged, and a new classifier was added to the top of every model. A series of layers were added on top of the pre-trained CNN architecture to build a classifier. These layers included a global average pooling (GAP) layer, a fully connected layer with 512 neurons using ReLU activation, and a drop-out layer with a rate of 0.3 to prevent overfitting. After that, another fully connected layer with 256 neurons and another drop-out layer were used. Finally, a 7-node dense output layer with SoftMax activation was added for classification.

The GAP layer [30] calculates the average value of each feature map, producing a fixed-length vector for each image, which is then passed to the next layers. The fully connected layers perform linear transformations on the input data, with the ReLU activation function introducing non-linearity. To reduce overfitting, the dropout layer randomly deactivates some neurons. Lastly, the SoftMax function is applied to the output dense layer to produce the final output. Fig. 4 shows the architecture of transfer learning based pre-trained base-learners.

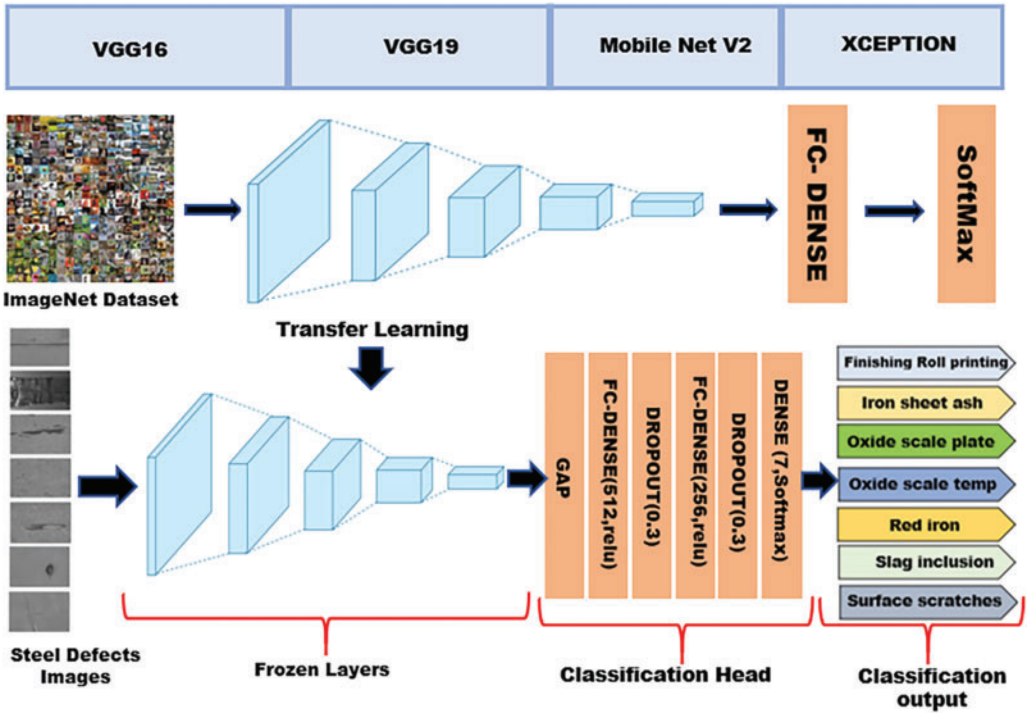


Figure 4: The architecture of transfer learning based pre-trained base-learners

3.2.2 Deep Ensemble Model

This research develops an ensemble learning-based system to enhance the classification accuracy of steel surface defects. In ensemble learning, several BLs are trained on the same dataset, and their results are combined to improve accuracy and reduce the variance of the overall model [8,31]. The aim of ensemble learning is to enhance the final performance of a model by combining the processes of several predictors. As shown in the Fig. 5, we created ensemble models in this study by combining predictions from four different pre-trained BLs: VGG16 (BL1), VGG19 (BL2), Xception (BL3), and Mobile-Net V2 (BL4). These base learners were fine-tuned using various hyperparameters, including the optimizer and learning rate. They were then trained on the training set to learn patterns in the data. The four BLs are combined using two ensemble techniques, averaging ensemble (AE) and weighted averaging ensemble (WAE), to improve performance and reduce variance.

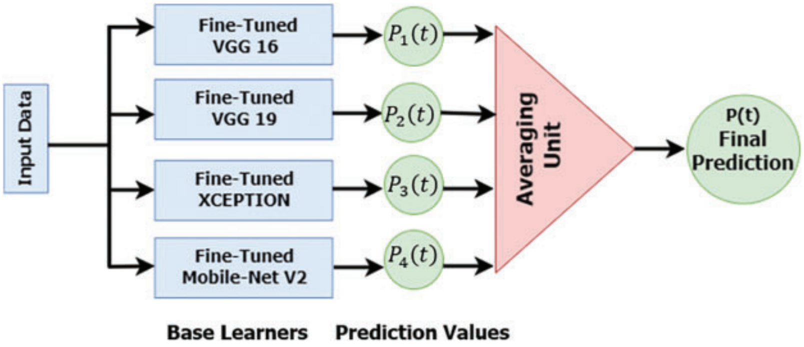


Figure 5: Schematic of averaging and weighted averaging ensemble techniques

Averaging Ensemble (AE)

The averaging ensemble [31,32] technique produces the ensemble model's final prediction by averaging the outputs from the BLs. Averaging these learners reduces the variance among them, which helps enhance generalization performance, given the high variance and low bias characteristic of DL architectures. This can be achieved by either directly averaging the base learners' outputs or by using the SoftMax function to average the classes' predicted probability. The final SoftMax outputs from each learner were combined by averaging, as displayed in Eq. (1).

$$Averaging\ ensemble\ Prediction\ P(t) = \sum\nolimits_{i=1}^{N} p_i(t) \tag{1}$$

where N is the number of base-learner (in this case, N = 4) and p_i is the probability for base-learner i at time t.

Proposed Weighted Averaging Ensemble (WAE)

The proposed weighted averaging ensemble (WAE) model is developed by combining the contributions of four different base learners (BLs) to the final prediction, with weights assigned based on their performance. This method ensures that the final prediction is significantly influenced by the more accurate BLs, thereby enhancing the overall effectiveness of the model. Highly-performing BLs will obtain higher weights compared to lower-performing ones [33]. In the proposed WAE model, each model effects the final output based on its strengths and weaknesses. The formula used to combine the predictions of several BLs in a weighted averaging ensemble is given in Eq. (2).

$$Weighted\ Averaging\ ensemble\ Prediction\ P(t) = \sum\nolimits_{i=1}^{N} w_i p_i(t) \tag{2}$$

where w_i denotes the weight of each base-learner. Fig. 6 shows the system diagram of the proposed WAE model.

3.3 Experimental Configuration, Performance Metrics and Indicators

The XSDD dataset consists of a total of 1360 images, which were increased to 3630 images using augmentation techniques to improve the diversity of the dataset and enhance the performance of the models. The input data was resized to 224 pixels in height and 224 pixels in width to meet the model's requirements. After resizing, the dataset was divided into three sections: 70% allocated for training, 20% for validation, and 10% for testing.

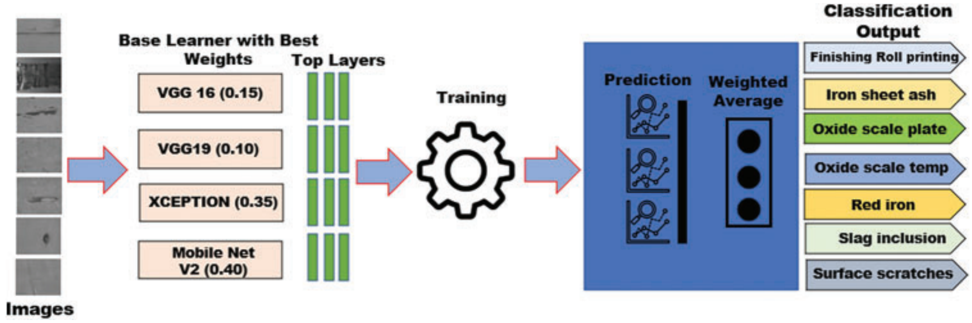


Figure 6: The system of proposed WAE model

For all experiments, we utilized an AMD Ryzen 7 2700X Eight-Core (3.70 GHz) Processor along with a Nvidia GeForce RTX 2080Ti. The code was implemented in a Jupyter Notebook environment using the Anaconda platform with TensorFlow 2.10.0 and Python 3.9.18. The system components and specifications of the setup are shown in Table 2.

Table 2: System components and specification

System components	Specifications
Operating system	Windows 10
CPU	AMD Ryzen 7 2700X Eight-Core (3.70 GHz) Processor
GPU	Nvidia GeForce RTX 2080 Ti
RAM	32 GB
Software	TensorFlow 2.10.0 and Python 3.9.18

In our experiments, we set the batch size to 32 and adjusted the learning rate between 0.001 to 0.0001 to ensure training stability. Optimization was performed using the Adam optimizer for the base learners and the Adamax optimizer for the ensemble models due to its stability during training for complex models. The categorical cross-entropy loss (CCE) was used as the loss function. We trained the model over 50 epochs. This combination of hyperparameters yielded the best results for steel defect classification. Details of the model configuration and hyperparameters are displayed in Table 3.

Table 3: Model configuration and hyper parameters

Options	VGG16	VGG19	Xception	Mobile-Net V2	AE	WAE
Optimizer	Adam	Adam	Adam	Adam	Adamax	Adamax
Loss function	CCE	CCE	CCE	CCE	CCE	CCE
Batch size	32	32	32	32	32	32
Epochs	50	50	50	50	50	50
Learning rate	0.0001	0.0001	0.0001	0.0001	0.001	0.001

3.4 Performance Metrics

To assess the model's performance and classification accuracy, we employed four key performance metrics: accuracy, macro recall (sensitivity), precision, and F1-score. These measures were labelled as follows: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

- True Positive (TP): defect correctly detected in steel surface.
- True Negative (TN): no defect present, correctly identified.
- False Positive (FP): no defect present, incorrectly identified as defective.
- False Negative (FN): defect present, missed during detection in steel.

Eqs. (3)–(6) define the metrics accuracy, macro precision, macro recall, and macro F1-score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Macro\ Recall = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i} \tag{4}$$

$$Macro\ Precision = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i} \tag{5}$$

$$Macro\ F1 - score = \frac{1}{N}\sum_{i=1}^{N} 2 \times \frac{Precission_{(i)} * Recall_{(i)}}{Precission_{(i)} + Recall_{(i)}} \tag{6}$$

These equations measure the overall performance of the classification model to classify defects in steel surfaces, and N is the total number of classes of defects.

4 Results and Discussion

This study evaluates various DL-based techniques to find the most effective and accurate model for identifying steel surface defects. The effectiveness of these techniques is assessed using transfer learning-based learners and ensemble learning techniques. In model averaging, each model is given the equal weight, while in weighted averaging, the model with the best performance is assigned a higher weight than the lower-performing models. To select the best suitable weights for the weighted averaging ensemble model, we evaluated the weighted averaging ensemble model with different combinations of weights and chose the optimal combination of weights for the final model out of four different weight combinations, as displayed in Fig. 7. The weighted averaging ensemble model provided better results with weights BL1: 0.15, BL2: 0.10, BL3: 0.35, and BL4: 0.40 as compared to other combinations of weights during validation and testing as shown in Table 4.

Table 4 provides an in-depth analysis of the accuracy of individual learners, the AE model, and the proposed WAE model. Both the AE and WAE models were evaluated on validation and test data, achieving accuracies of 99.17% and 99.86% on the validation data, respectively, and 98.89% and 99.72% on the test data, respectively. The WAE model produced higher accuracy compared to the other models.

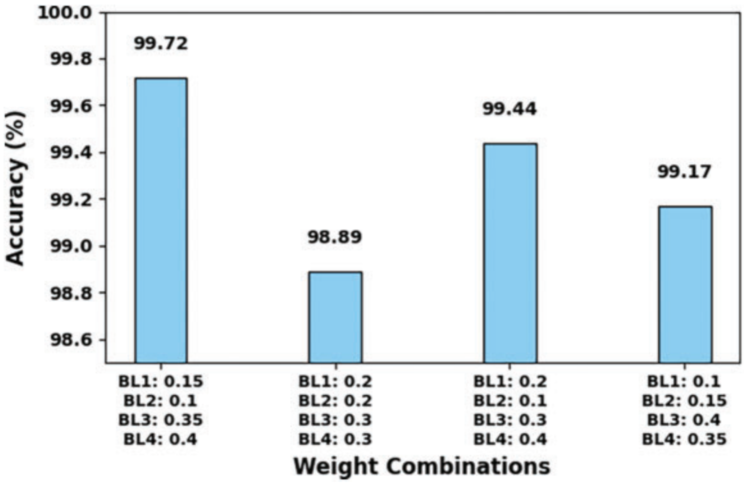


Figure 7: Accuracy of WAE model for different combination of weights

Table 4: Analysis of the accuracy in depth

Models	Train ACC	Val ACC	Test ACC
VGG16 (BL1)	96.26%	96.41%	93.66%
VGG19 (BL2)	92.40%	91.59%	90.63%
Xception (BL3)	99.40	98.62%	98.34%
Mobile-Net V2 (BL4)	99.88%	99.03%	98.62%
AE	99.40%	99.17%	98.89%
WAE (BL1: 0.15, BL2: 0.10, BL3: 0.35, BL4: 0.40)	100%	99.86%	99.72%

Table 5 provides a detailed comparative analysis of various DL models based on performance metrics. The VGG16 model performs with an accuracy of 93.66%, a recall of 93.54%, a precision of 94.20%, and an F1-score of 93.69% on the test data. This indicates a well-balanced capability for correctly classifying instances. In comparison, the VGG19 model shows slightly lower performance with an accuracy of 90.63%, a recall of 90.22%, a precision of 91.42%, and an F1-score of 90.59% on the test data. The Xception model performs robustly on the test data, with an accuracy of 98.34%, a recall of 98.19%, a precision of 98.53%, and an F1-score of 98.33%. Mobile-Net V2 shows outstanding performance on the test data, achieving an accuracy and recall of 98.62%, a precision of 98.56%, and an F1-score of 98.57%. These results show that Mobile-Net V2 performs well compared to all other base learners in the classification of steel surface defects.

Furthermore, the AE model, which integrates the predictions of multiple models, shows enhanced performance on the test data with an accuracy of 98.89%, a recall of 98.92%, a precision of 99.05%, and an F1-score of 98.97%. The WAE model achieves the highest metrics on the test data, with an accuracy of 99.72%, a recall of 99.74%, a precision of 99.67%, and an F1-score of 99.70%. These results show the advantages of ensemble learning techniques in improving model accuracy.

Table 5: Comparison of results of base learners with proposed WAE model

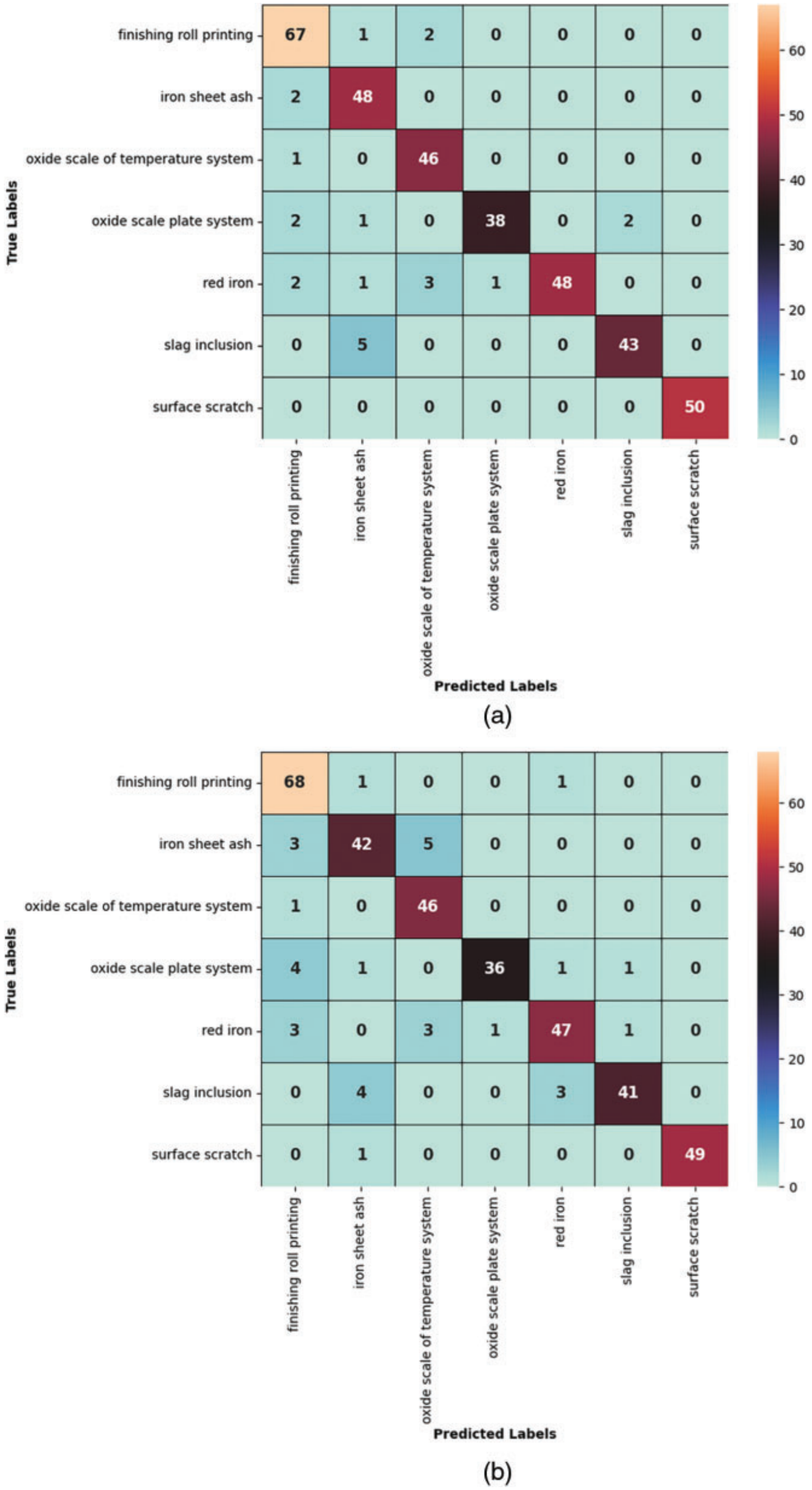
Models	Accuracy	Recall	Precision	F1-score
VGG16	93.66%	93.54%	94.20%	93.69%
VGG19	90.63%	90.22%	91.42%	90.59%
Xception	98.34%	98.19%	98.53%	98.33%
Mobile-Net V2	98.62%	98.62%	98.56%	98.57%
AE	98.89%	98.92%	99.05%	98.97%
Proposed WAE	99.72%	99.74%	99.67%	99.70%

Fig. 8 shows the confusion matrix (CM) for base-learners and ensemble model on test data. Fig. 8a shows the CM for VGG16, highlights its ability to distinguish steel defects. Finishing roll printing had three misclassifications, one as iron sheet ash and two as oxide scale of temperature system. Iron sheet was misclassified twice, as finishing roll printing. Oxide scale of temperature system had one misclassification as finishing roll printing. Oxide scale plate system showed five misclassifications, two as finishing roll printing, two as slag inclusion, and one as iron sheet ash. Red iron had seven misclassifications, including one as oxide scale plate system, one as iron sheet ash, three as oxide scale of temperature system, and two as finishing roll printing. Slag inclusion was misclassified five times, all as iron sheet ash. Surface scratch showed no misclassifications.

CM for VGG19 is shown in Fig. 8b, finishing roll printing had two misclassifications, one as iron sheet and one as red iron. Iron sheet ash had eight misclassifications, five as oxide scale of the temperature system and three as finishing roll printing. Oxide scale of the temperature system had one misclassification as finishing roll printing. Oxide scale plate system showed seven misclassifications, four as finishing roll printing, one as iron sheet ash, one as red iron, and one as slag inclusion. Red iron had eight misclassifications, including one as oxide scale plate system and one as slag inclusion, three as oxide scale of the temperature system, and three as finishing roll printing. Slag inclusion was misclassified seven times, four as iron sheet ash and three as red iron. Surface scratch showed only one misclassification as iron sheet ash.

Fig. 8c shows the CM for Xception, iron sheet ash had one misclassification, as finishing roll printing. Oxide scale plate system showed three misclassifications, one as oxide scale of the temperature system, finishing roll printing, and one as slag inclusion. Red iron had two misclassifications, including one as finishing roll printing, one as oxide scale of the temperature system. All other defects types had no misclassification. CM for Mobile-Net V2 is shown in Fig. 8d, iron sheet ash had three misclassification, two as red iron, one as an oxide scale plate system. Red iron experienced two misclassifications, including one as oxide scale of plate system and one as surface scratch. All other defects types had no misclassification. Mobile-Net V2 shows superior results as compare to all other base-learners.

Fig. 8e shows the CM for AE, red iron had three misclassifications, including one as surface scratch and two as finishing roll printing. Slag inclusion had one misclassification, as red iron. All other defects types had no misclassification. Fig. 8f shows the CM for WAE, red iron had one misclassification, as oxide scale plate system. All other defects types correctly classified and there is no misclassification. The misclassification may have occurred because the visual patterns of these two types of defects are quite similar, which likely caused the confusion. Overall, the proposed WAE model showed outperforming results compared to all others models. This comparison shows the power of ensemble learning, where the strengths of individual learners are combined.



(a)

(b)

Figure 8: (Continued)

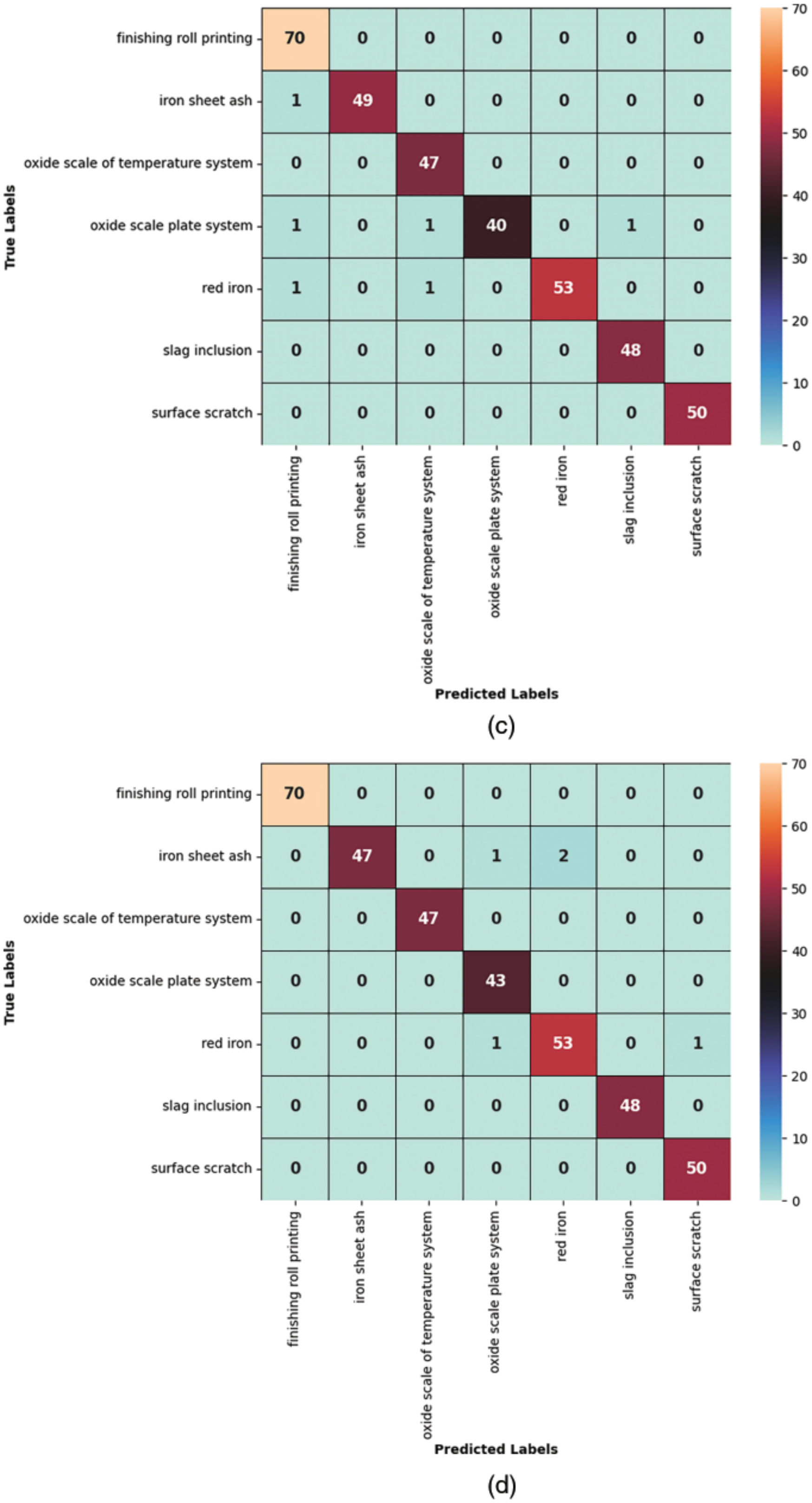
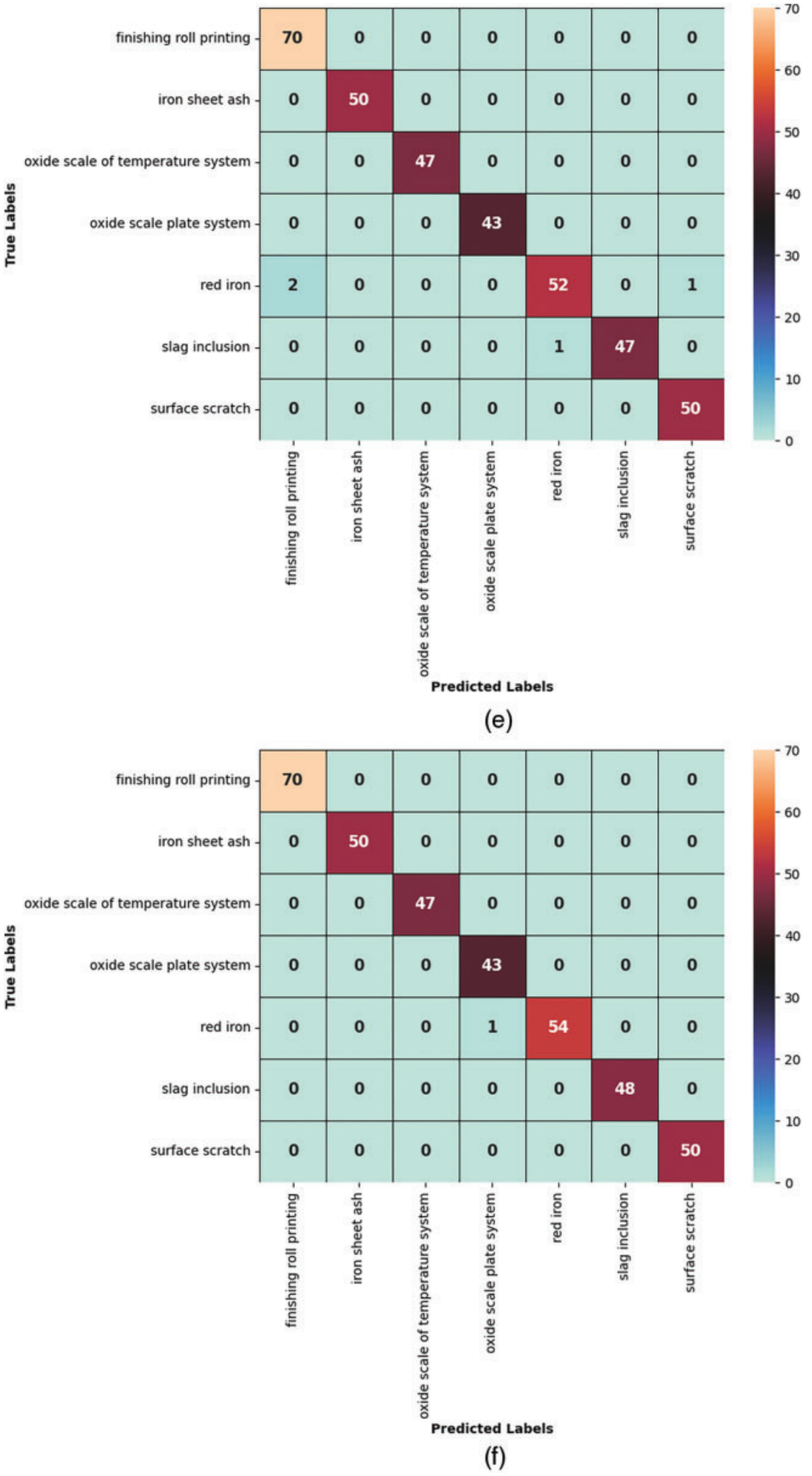


Figure 8: (Continued)



(e)

True Labels \ Predicted Labels	finishing roll printing	iron sheet ash	oxide scale of temperature system	oxide scale plate system	red iron	slag inclusion	surface scratch
finishing roll printing	70	0	0	0	0	0	0
iron sheet ash	0	50	0	0	0	0	0
oxide scale of temperature system	0	0	47	0	0	0	0
oxide scale plate system	0	0	0	43	0	0	0
red iron	2	0	0	0	52	0	1
slag inclusion	0	0	0	0	1	47	0
surface scratch	0	0	0	0	0	0	50

(f)

True Labels \ Predicted Labels	finishing roll printing	iron sheet ash	oxide scale of temperature system	oxide scale plate system	red iron	slag inclusion	surface scratch
finishing roll printing	70	0	0	0	0	0	0
iron sheet ash	0	50	0	0	0	0	0
oxide scale of temperature system	0	0	47	0	0	0	0
oxide scale plate system	0	0	0	43	0	0	0
red iron	0	0	0	1	54	0	0
slag inclusion	0	0	0	0	0	48	0
surface scratch	0	0	0	0	0	0	50

Figure 8: Confusion matrix of models: (a) VGG16, (b) VGG19, (c) Xception, (d) Mobile-Net V2, (e) AE, (f) Proposed WAE

Table 6 presents a comparison of the proposed WAE model with previous research in terms of accuracy. The RepVGG algorithm, combined with the spatial attention mechanism [16], achieved 95.10% accuracy. ResNet50, when combined with the convolutional block attention module and frequency channel attention network [1], reached 93.87% accuracy. The improved Mobile-Net [15] attained 95.97% accuracy. The Multi-SE-ResNet34 [19] achieved 99.20% accuracy by integrating an attention mechanism into the ResNet34 architecture. The MSMA-SDD achieved [18] 99.59% accuracy by combining a multiscale CNN with a multi-attention module. The novel feature extraction legendre multi-wavelet transform and Auto-Encoder network [17] reached 95.37% accuracy with SVM. The spiking vision transformer model combined with leaky integration and fire [34] achieved 97.19% accuracy. The proposed WAE model surpassed all other methodologies with the highest accuracy of 99.72%, shows the power of combining the strengths of individual learners through weighted averaging.

Table 6: Comparison of proposed WAE with previous research

Publication	Years	Methodology	Accuracy
Feng et al. [16]	2021	RepVGG + SA	95.10%
Feng et al. [1]	2021	ResNet50 + CBAM + FcaNet	93.87%
Lin et al. [15]	2022	Improved Mobile-Net	95.97%
Hao et al. [19]	2022	Multi-SE-ResNet34	99.20%
Wen et al. [18]	2023	MSMA-SDD	99.59%
Zheng et al. [17]	2024	LWT-AE	95.37%
Gong et al. [34]	2024	s-ViT + LIF	97.19%
Proposed	–	**WAE**	**99.72%**

5 Conclusion

In this paper, we introduced a weighted averaging ensemble method for classifying hot-rolled steel surface defects. We assessed the ensemble model's performance with both the traditional model averaging ensemble and the proposed weighted averaging ensemble technique. The model averaging treats every models equally and achieved an accuracy of 99.17% on validation data and 98.89% on test data. In contrast, the weighted averaging ensemble assigns higher weights to base-learners with higher accuracy, resulting in an improved accuracy of 99.86% on validation data and 99.72% on test data. Our findings indicate that weighted averaging ensemble significantly outperformed both individual base-learners and the traditional model averaging ensemble technique in terms of accuracy. Specifically, we observed accuracy improvements of 0.83% for the averaging ensemble, 1.1% for Mobile-Net V2, 1.38% for Xception, 9.38% for VGG19, and 6.06% for VGG16 on the test dataset. As a result, the proposed ensemble model achieved exceptionally high performance and significantly improved the classification process. This proposed methodology will help manufacturers in the steel industry to achieve more accurate results.

Although the experimental results show the effectiveness of proposed WAE model, it was evaluated on a specific dataset of hot-rolled steel surface defects, which may limit its generalizability to other types of surface defect datasets. Additionally, the model's performance could be influenced by the choice of base learners, so exploring different combinations of base learners could be beneficial.

In the future, a new architecture will be developed with the other base-learners combination to improve steel surface recognition processes and, the model will be tested with different datasets.

Acknowledgement: The authors want to say thank you to the National Research Foundation of Korea (NRF) and the Ministry of Education for their financial support.

Funding Statement: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2022R1I1A3063493).

Author Contributions: Study conception and design: Tajmal Hussain and Jongwon Seok; data collection: Tajmal Hussain and Jongwon Seok; analysis and interpretation of results: Tajmal Hussain; draft manuscript preparation: Tajmal Hussain and Jongwon Seok. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data are available from the corresponding author upon reasonable request.

Ethics Approval: None.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Feng X, Gao X, Luo L. A ResNet50-based method for classifying surface defects in hot-rolled strip steel. Mathematics. 2021;9(19):2359. doi:10.3390/math9192359.
2. Xu Z-W, Liu X-M, Zhang K. Mechanical properties prediction for hot rolled alloy steel using convolutional neural network. IEEE Access. 2019;7(1):47068–78. doi:10.1109/ACCESS.2019.2909586.
3. Song K, Yan Y. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. Appl Surf Sci. 2013;285(1):858–64. doi:10.1016/j.apsusc.2013.09.002.
4. Hao R, Lu B, Cheng Y, Li X, Huang B. A steel surface defect inspection approach towards smart industrial monitoring. J Intell Manuf. 2021;32(1):1833–43. doi:10.1007/s10845-020-01670-2.
5. Ibrahim AAM, Tapamo JR. Transfer learning-based approach using new convolutional neural network classifier for steel surface defects classification. Sci Afr. 2024;23(1):e02066. doi:10.1016/j.sciaf.2024.e02066.
6. Alkapov R, Konyshev A, Vetoshkin N, Valkevich N, Kostenetskiy P. Automatic visible defect detection and classification system prototype development for iron-and-steel works. In: Global Smart Industry Conference (GloSIC), 2018; Chelyabinsk, Russia; p. 1–8.
7. Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: a review. Eng Appl Artif Intell. 2022;115:105151. doi:10.1016/j.engappai.2022.105151.
8. Mohammed A, Kora R. A comprehensive review on ensemble deep learning: opportunities and challenges. J King Saud Univ–Comput Inf Sci. 2023;35(2):757–74. doi:10.1016/j.jksuci.2023.01.014.
9. Hu H, Li Y, Liu M, Liang W. Classification of defects in steel strip surface based on multiclass support vector machine. Multimed Tools Appl. 2014;69(1):199–216. doi:10.1007/s11042-012-1248-0.
10. Xiao M, Jiang M, Li G, Xie L, Yi L. An evolutionary classifier for steel surface defects with small sample set. Eurasip J Image Vide. 2017;2017(1):1–13. doi:10.1186/s13640-017-0197-y.
11. Liu Y, Xu K, Wang D. Online surface defect identification of cold rolled strips based on local binary pattern and extreme learning machine. Metals. 2018;8(3):197. doi:10.3390/met8030197.

12. Mentouri Z, Doghmane H, Moussaoui A, Boudjehem D. Surface flaw classification based on dual cross pattern. In: 2020 1st International Conference on Communications, Control Systems and Signal Processing (CCSSP), 2020; El Oued, Algeria; p. 137–41.
13. Gong R, Chu M, Yang Y, Feng Y. A multi-class classifier based on support vector hyper-spheres for steel plate surface defects. Chemom Intell Lab Syst. 2019;188(1):70–8. doi:10.1016/j.chemolab.2019.03.010.
14. Ashour MW, Khalid F, Abdul Halin A. Surface defects classification of hot-rolled steel strips using multi-directional shearlet features. Arab J Sci Eng. 2019;44(4):2925–32. doi:10.1007/s13369-018-3329-5.
15. Lin L, Wang Y, Zhao S, Liu J, Zhang S, Zhang G. Small samples data augmentation and improved MobileNet for surface defects classification of hot-rolled steel strips. J Electron Imaging. 2022;31(6):063056. doi:10.1117/1.JEI.31.6.063056.
16. Feng X, Gao X, Luo L. X-SDD: a new benchmark for hot rolled steel strip surface defects detection. Symmetry. 2021;13(4):706. doi:10.3390/sym13040706.
17. Zheng X, Liu W, Huang Y. A novel feature extraction method based on legendre multi-wavelet transform and auto-encoder for steel surface defect classification. IEEE Access. 2024;12(2):5092–102. doi:10.1109/ACCESS.2024.3349628.
18. Wen L, Zhang Y, Gao L, Li X, Li M. A new multiscale multiattention convolutional neural network for fine-grained surface defect detection. IEEE Trans Instrum Meas. 2023;72(1):1–11. doi:10.1109/TIM.2023.3271743.
19. Hao Z, Li Z, Ren F, Lv S, Ni H. Strip steel surface defects classification based on generative adversarial network and attention mechanism. Metals. 2022;12(2):311. doi:10.3390/met12020311.
20. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6(1):1–48. doi:10.1186/s40537-019-0197-0.
21. Chen Y, Fu Q, Wang G. Surface defect detection of nonburr cylinder liner based on improved YOLOv4. Mob Inf Syst. 2021;2021(1):1–13. doi:10.1155/2021/9374465.
22. Zhang J, Cosma G, Watkins J. Image enhanced mask R-CNN: a deep learning pipeline with new evaluation measures for wind turbine blade defect detection and classification. J Imaging. 2021;7(3):46. doi:10.3390/jimaging7030046.
23. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. In: Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, 2018; Rhodes, Greece; p. 270–9.
24. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data. 2016;3(1):1–40. doi:10.1186/s40537-016-0043-6.
25. Wan X, Zhang X, Liu L. An improved VGG19 transfer learning strip steel surface defect recognition deep neural network based on few samples and imbalanced datasets. Appl Sci. 2021;11(6):2606. doi:10.3390/app11062606.
26. Hussain T, Hong J, Seok J. A hybrid deep learning and machine learning-based approach to classify defects in hot rolled steel strips for smart manufacturing. Comput Mater Contin. 2024;80(2):2099–119. doi:10.32604/cmc.2024.050884.
27. Feng X, Gao X, Luo L. A method for surface detect classification of hot rolled strip steel based on Xception. In: 33rd Chinese Control and Decision Conference (CCDC), 2021; Kunming, China; p. 1485–9.
28. Jin G, Liu Y, Qin P, Hong R, Xu T, Lu G. An end-to-end steel surface classification approach based on EDCGAN and MobileNet V2. Sensors. 2023;23(4):1953. doi:10.3390/s23041953.
29. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis. 2015;115(1):211–52. doi:10.1007/s11263-015-0816-y.
30. Lin M, Chen Q, Yan S. Network in network. 2014. doi:10.48550/arXiv.1312.4400.
31. Arpit D, Wang H, Zhou Y, Xiong C. Ensemble of averages: Improving model selection and boosting performance in domain generalization. 2022. doi:10.48550/arXiv.2110.10832.

32. Naftaly U, Intrator N, Horn D. Optimal ensemble averaging of neural networks. Netw Comput Neural Syst. 2009;8(3):283–96. doi:10.1088/0954-898X/8/3/004.

33. Iqball T, Wani MA. Weighted ensemble model for image classification. Int J Inf Tecnol. 2023;15(2):557–64. doi:10.1007/s41870-022-01149-8.

34. Gong L, Dong H, Zhang X, Cheng X, Ye F, Guo L, et al. Spiking ViT: spiking neural networks with transformer—attention for steel surface defect classification. J Electron Imaging. 2024;33(3):033001. doi:10.1117/1.JEI.33.3.033001.