NIPPON STEEL TECHNICAL REPORT No. 136 JULY 2026

Technical Report

# Machine Learning-based Microstructural Analysis for Steel Materials

Kazumasa TSUTSUI*
Koji MORIGUCHI
Hidenori TERASAKI
Koutarou HAYASHI
Shigekazu MORITO

## Abstract

***Steel materials exhibit a broad range of mechanical properties, depending on alloy composition and heat treatment. It is therefore important to grasp such information given that these variations originate from the metallic structure in the micron-scale. The microstructure of steel can be characterized by scanning electron microscopy (SEM) and electron backscattering diffraction (EBSD). Since analyses by SEM and EBSD require deep expertise in morphology and crystallography, respectively, this can sometimes lead to individual differences in identification results. We sought to leverage machine learning algorithms to automate and standardize the identification of microstructures in steel, with the objective of eliminating inter-individual variations. In this paper, we present the findings from the application of various machine learning algorithms to EBSD and SEM data collected from low-carbon steel.***

## 1. Introduction

Steel materials are metallic materials used in a wide range of applications, including automobiles, construction, and shipbuilding. To meet these diverse application requirements, steel materials must be designed through the control of alloying elements and the application of various heat treatments, so that their internal microstructures at the micron-scale are appropriately tailored. Understanding the mechanisms by which specific microstructures are formed, as well as their associated physical and chemical characteristics, is critically important for achieving the desired mechanical properties. Therefore, analytical techniques such as scanning electron microscopy (SEM) and electron backscattering diffraction (EBSD) have become indispensable tools.[1, 2]

As an example, high-strength automotive steel sheets derived from low-carbon steel often exhibit a complex microstructure in which martensite, upper bainite, and lower bainite coexist, depending on the alloy composition and heat-treatment conditions.[3] When applying SEM to such materials, the identification of these microstructures and the determination of their spatial distribution are challenging, and assessments may differ even among experienced specialists. Furthermore, while combining EBSD-based crystal orientation analysis can improve microstructure identification accuracy, it requires advanced crystallography knowledge and significant analysis time.[4] Consequently, considerable effort has been directed toward establishing objective and quantitative methods for microstructural evaluation that minimize analyst-dependent variability and allow for rapid and reliable identification of microstructures.[5]

In light of this background, our research group has pursued research with the objective of applying machine learning techniques to automate and accelerate microstructural identification in steel materials, while establishing quantitative and highly reproducible identification technologies.[6–14] Specifically, we sought to develop automated microstructure identification technologies by extracting features from EBSD-based crystallographic orientation analyses for use in machine learning model training, as well as by applying feature extraction and deep learning techniques to SEM analyses that have traditionally depended on visual identification.

The originality of this study lies in its dual approach: (1) incorporating morphological features and crystallographic metrics into classical machine learning models, and (2) employing deep learning models to directly learn features from images, with the aim of achieving highly accurate and versatile microstructure identification.

* Ph.D., Senior Researcher, Mathematical Science & Technology Research Lab., Advanced Technology Research Laboratories
20-1 Shintomi, Futtsu City, Chiba Pref. 293-8511

In the former approach, the discriminative features are predefined, whereas in the latter, the features are automatically learned through training. By leveraging these techniques, not only can the efficiency and quantification of steel microstructure analysis be improved, but broader industrial impacts are also anticipated, including the proposal of novel material design guidelines and the advancement of quality evaluation methodologies. This paper provides an overview of these studies and discusses the obtained results and their significance.

The remainder of this paper is as follows. Chapter 2 introduces the datasets used in this research. Chapter 3 provides an overview of the investigations regarding the aforementioned (1) and (2). Chapter 4 presents the results obtained through this research and discusses them. Chapter 5 summarizes this paper.

## 2. Specimen and Dataset Construction

For this study, low-carbon steel[6, 14] with the composition Fe-0.1C-0.01Si-2.0Mn-0.008P-0.001S (mass%) was selected as the specimen. To obtain different microstructures, a total of eight heat treatments, as shown in **Fig. 1**, were applied. The eight samples obtained under each heat treatment condition were carefully selected to include diverse microstructures such as martensite (MS), upper bainite (UB, BII), and lower bainite (LB, BIII). These microstructures were identified through SEM observation using the morphological identification criteria by Bramfitt and Speer[3] described later, and by evaluating the martensite fraction using the dilation method (**Table 1**). Specifically, this study used these eight sample types as the target categories for machine learning classification. The following sections describe the acquisition methods for the dataset, including SEM and EBSD measurements for each sample.

### 2.1 Construction of the EBSD dataset

For each of the eight samples, crystal orientation measurements using EBSD were performed twelve times without overlapping fields of view. All measurements were conducted under the following standardized conditions: the measurement area was fixed at 12 800 $\mu m^2$. The acceleration voltage was set to 15 kV, and the measurement step size was 0.2 $\mu$m.[6–9] An example of the inverse pole figure (IPF) map obtained from the orientation measurements is shown in **Fig. 2**.

In this study, kernel averaged misorientation (KAM) values and variant boundary density were estimated from the above crystal orientation data as features for the machine learning model. KAM value were derived by averaging the misorientation angles up to the third nearest neighbor within the selected kernel, provided the misorientation angle did not exceed 2°.[4] An overview of variant boundary density is provided below.

During the transformation from austenite ($\gamma$) grains to ferrite ($\alpha$') grains in low-carbon steel, multiple $\alpha$' grains with different crystal orientations are formed. These are called variants. It is known that the orientation relationship among variants of $\alpha$' grains generated from a single $\gamma$ grain satisfies the following Kurdjumov-Sachs (KS) relationship[15, 16]:

$$\{111\}_{\gamma} \| \{110\}_{\alpha'}, \ \langle 110\rangle_{\gamma} \| \langle 111\rangle_{\alpha'}$$

According to the KS relations, 24 distinct transformation orientations, i.e., 24 KS variants, form within a single $\gamma$ grain. Adjacent variants are called variant pairs. The ease of adjacency between variant pairs is known to vary with heat treatment and composition.[17, 18] Specifically, when one of the 24 variants is considered as V1, five orientations exist that share the same close-packed plane (CPP) relationship. Hereafter, following the notation of Morito et al.[4], these variant pairs will be referred to as V1-V2, V1-V3/V5, V1-V4, and V1-V6. The corresponding orientation relationships are shown in **Table 2**. The variant boundary density is defined as the boundary length divided by the area of the observed field of view.[18]

Through the above procedures, we obtained a total of 96 data points, comprising the densities of V1-V2, V1-V3/V5, V1-V4, and V1-V6, along with KAM value, forming a five-dimensional dataset. Due to the small total number of data points, five-fold cross-validation was performed to evaluate the accuracy during training of the machine learning models described later.[7]

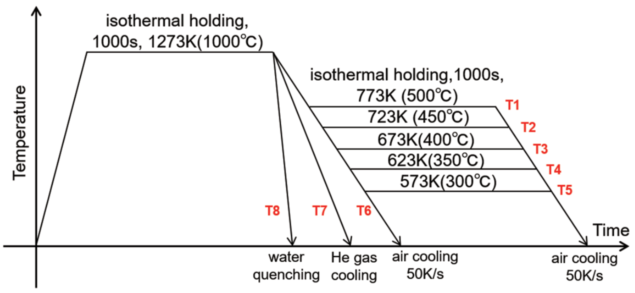


**Fig. 1 Schematic of thermal processing patterns of specimens**
**Source: reproduced from Tsutsui et al. (2025)[14], with permission**

**Table 1 Information for the specimens of martensite fraction and types of microstructures**

| Label | Martensite fraction (%) | Microstructural type |
|---|---|---|
| T1 | 0 | UB |
| T2 | 2.1 | UB and LB (UB rich) |
| T3 | 63.6 | UB and LB (LB rich) |
| T4 | 63.9 | UB and LB (LB rich) |
| T5 | 62.2 | UB and LB (LB rich) |
| T6 | 70 | LB |
| T7 | 97.8 | LB and MS |
| T8 | 100 | MS |

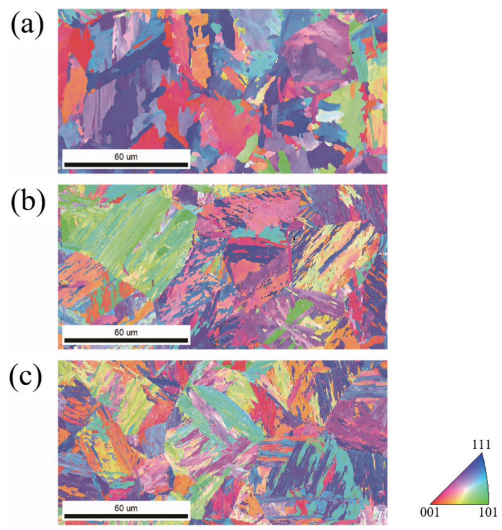


**Fig. 2 Examples of IPF maps of (a) T1 (UB), (b) T6 (LB), and (c) T8 (MS)**

**Table 2 KS variant in same CPP relationship**
**The relationship is selected by $(111)_\gamma//(011)_{\alpha'}$. Refer to the definition of the variant numbers by Morito et al.[4)]**

| Variant no. | Close-packed orientation relationship | Misorientation angle and axis with V1 (theoretical) | Misorientation angle and axis with V1 (experimental) |
|---|---|---|---|
| V1 | $[\bar{1}\ 0\ 1]_\gamma//[\bar{1}\ \bar{1}\ 1]_{\alpha'}$ | – | – |
| V2 | $[\bar{1}\ 0\ 1]_\gamma//[\bar{1}\ 1\ \bar{1}]_{\alpha'}$ | $60.0°/[1\ 1\ \bar{1}]_{\alpha'} = 70.5°/[0\ \bar{1}\ \bar{1}]_{\alpha'}$ | $68°/\langle 0\ 1\ 1\rangle_{\alpha'}$ |
| V3 | $[0\ 1\ \bar{1}]_\gamma//[\bar{1}\ \bar{1}\ 1]_{\alpha'}$ | $60.0°/[0\ 1\ 1]_{\alpha'}$ | $60°/\langle 0\ 1\ 1\rangle_{\alpha'}$ |
| V4 | $[0\ 1\ \bar{1}]_\gamma//[\bar{1}\ 1\ \bar{1}]_{\alpha'}$ | $10.5°/[0\ \bar{1}\ \bar{1}]_{\alpha'}$ | $7°/\langle 0\ 1\ 1\rangle_{\alpha'}$ |
| V5 | $[1\ \bar{1}\ 0]_\gamma//[\bar{1}\ \bar{1}\ 1]_{\alpha'}$ | $60.0°/[0\ \bar{1}\ \bar{1}]_{\alpha'}$ | $60°/\langle 0\ 1\ 1\rangle_{\alpha'}$ |
| V6 | $[1\ \bar{1}\ 0]_\gamma//[\bar{1}\ 1\ \bar{1}]_{\alpha'}$ | $49.5°/[0\ 1\ 1]_{\alpha'}$ | $52°/\langle 0\ 1\ 1\rangle_{\alpha'}$ |

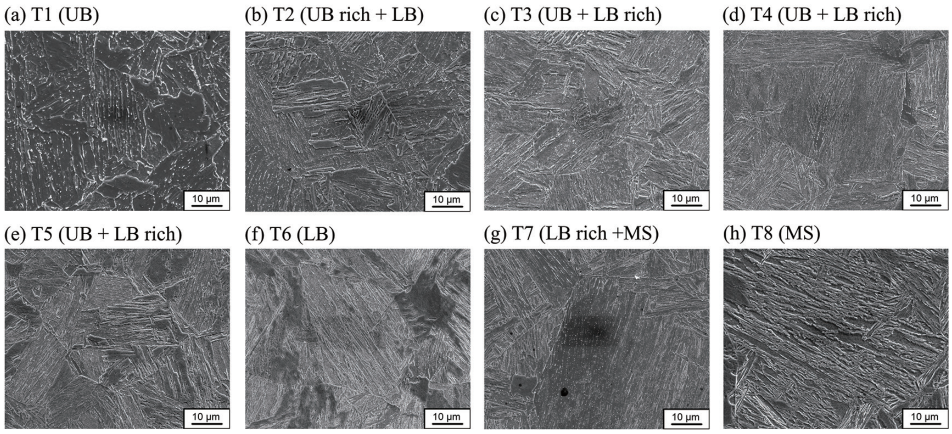


**Fig. 3 Examples of FE-SEM images**
**Source: reproduced from Tsutsui et al. (2025)[14)], with permission**

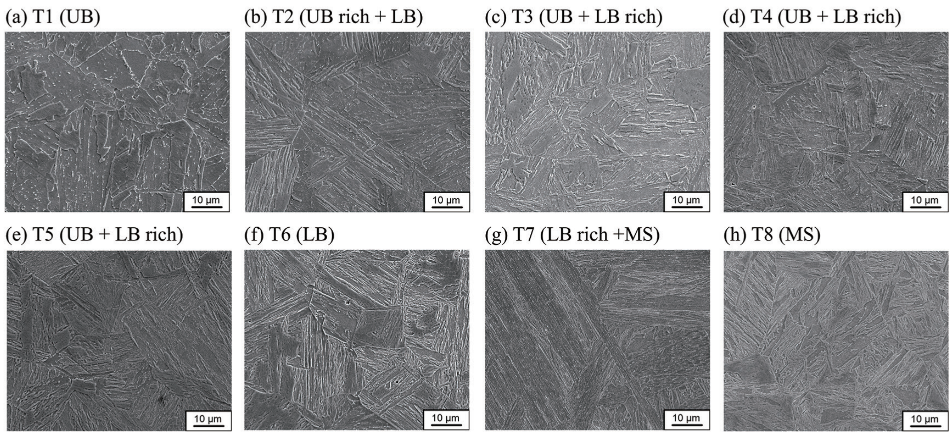


**Fig. 4 Examples of W-SEM images**
**Source: reproduced from Tsutsui et al. (2025)[14)], with permission**

### 2.2 Preparation of SEM image dataset

The SEM imaging conditions are described below. The eight prepared samples were polished using 9 $\mu$m and 3 $\mu$m diamond paste, respectively, followed by etching with 3% nitric acid. Two types of SEM sources—tungsten (W) and field emission (FE)—were used to acquire 1500 × magnification images.[12–14)] Thirty images were acquired for each source and sample (total: 2 × 30 × 8 = 480 images). The acceleration voltage was set to 15 kV for both sources. The spatial resolutions were 30 nm for the W source and 3.0 nm for the FE source. The dimensions of the obtained SEM images were 1280 × 1024 px for the W source and 1280 × 960 px for the FE source. FE-SEM and W-SEM images of each sample are shown in **Figs. 3** and **4**, respectively. Visually, the FE-SEM images capture the surface topography of the material with smoother representation of surface undulations, whereas the W-SEM images appear to depict the surface as comparatively flatter.

As described in Chapter 1, this study adopted two approaches: (1) feature extraction and classification using classical machine learning models, and (2) direct image classification using deep learning models. The preprocessing steps and feature extraction procedures applied to the SEM image dataset are outlined below.

#### 2.2.1 Preprocessing for the SEM dataset

The preprocessing applied to the acquired SEM images is described below. First, contrast-limited adaptive histogram equalization (CLAHE) was applied to adjust contrast between images. Subsequently, the dataset was split into an 80%:20% ratio for model training and testing. Furthermore, for approaches (1) and (2) using SEM images, we aimed to improve accuracy through majority voting. Therefore, we decided to use these images after dividing them into smaller sub-images.

For approach (1), the original image was cropped to 880 × 880 px region centered on its origin, then subdivided into 16 parts as mini-images. Each mini-image size was 220 × 220 px. This procedure was applied uniformly to both training and test data.[11]

Additionally, (2) in the approach using deep learning models, data augmentation was applied to the training data. Specifically, by randomly cropping and rotating images, 400 mini-images at 224 × 224 px were extracted per original image. For the test data, a total of 20 non-overlapping images were extracted at 224 × 224 px, centered on the image. The dimensions of these small images were adjusted to match the input size of the deep learning model we used.[12–14]

#### 2.2.2 Extraction of morphological features

The morphological identification criteria for upper bainite, lower bainite, and martensite proposed by Bramfitt and Speer[3] are as follows. Both upper bainite and lower bainite are characterized by having a lath-like *α*' structure with carbide precipitation. The difference lies in the distribution: upper bainite is finely distributed between ferrite grain boundaries, while lower bainite is finely distributed within grains. Martensite is characterized by the absence of carbide precipitation.

For approach (1) applied to SEM images, selecting appropriate features is necessary to capture the above morphological differences. As shown in Figs. 3 and 4, the morphological features of martensite and bainite, including the flatness of the ferritic structure and the presence of lath boundaries and carbides, are represented in the image texture as differences in brightness, appearing as variations in surface smoothness or randomness. The gray-level co-occurrence matrix (GLCM) method[19, 20] is a feature extraction method capable of evaluating such brightness patterns. An overview of the GLCM method is as follows.

In the GLCM method, a co-occurrence matrix $P(i,j|d,\theta)$ is first constructed. This matrix represents the frequency of occurrence of combinations of a pixel with value $i$ and a pixel with value $j$ located at a distance $d$ and orientation $\theta$ from pixel $i$. By assigning various expectation values to $P(i,j|d,\theta)$, textures within the image can be quantified. Representative metrics include the following homogeneity $H$ and entropy $E$.

$$H \equiv \sum_{i,j} \frac{P(i,j|d,\theta)}{1+|i-j|}, \quad E = -\sum_{i,j} P(i,j|d,\theta)\log P(i,j|d,\theta)$$

These are textural values representing the flatness and randomness of the image, respectively. In this study, following Webel et al.[21, 22], we constructed a co-occurrence matrix by setting $d = 1$ and averaging $\theta$ over the range 0° to 180°, calculating a total of 19 textural values including $H$ and $E$. Details of the calculated textural values are described in our previous work[10, 12].

In addition, as another example of image textural analysis, the local binary pattern (LBP) method has been reported.[23, 24] The LBP method is a feature extraction technique that calculates the brightness difference between each pixel in an image and its neighboring pixels, converts the magnitude of this difference into binary values, and encodes them as decimal numbers. Furthermore, methods exist that utilize the intermediate layer outputs of trained deep learning models for feature extraction,[25] but in this paper, we opted to adopt the GLCM method, which is easier to correlate with morphological indicators.

## 3. Machine Learning and Deep Learning Algorithms

### 3.1 Models used in each approach

This section introduces the algorithms used in this study. As mentioned earlier, our approaches can be broadly divided into two strategies: (1) feature extraction followed by classification using classical machine learning models, and (2) direct image classification using deep learning models. In the following, these are referred to as ML models and DL models, respectively.

In approach (1), EBSD and SEM datasets were used independently to train ML models, enabling an investigation of classification performance and the identification of effective features. For the EBSD dataset, support vector machines (SVMs),[26, 27] which belong to the class of kernel-based methods, were mainly adopted in consideration of the subsequent construction of microstructural diagrams. For the SEM image dataset, random forest (RF)[28] and light gradient boosting machine (LightGBM),[29] both based on decision tree algorithms, were employed with the aim of estimating the GLCM features that contribute to classification. Specifically, the FE-SEM and W-SEM datasets were trained independently to examine whether differences in classification performance and effective features arose depending on the imaging conditions.

For approach (2), ResNet50,[30] which incorporates residual blocks, was used as the DL model. Separate ResNet50 models were trained independently on the FE-SEM and W-SEM datasets. Furthermore, since this approach directly uses images, we decided to estimate the regions within the images contributing to classification using the local interpretable model-agnostic explanations (LIME) method[31]. Related to approach (1), the classification accuracy of ResNet50 was found to degrade significantly when SEM images acquired under imaging conditions different from those used for training (FE or W) were used as test data. To mitigate this issue, imaging condition conversion was explored using cycle-consistent Generative Adversarial Networks (CycleGAN),[32] a variant of GAN.

Among the ML and DL models employed, the LIME method and CycleGAN are described in detail in the following section due to the complexity of their mechanisms. For the remaining models, detailed explanations are omitted, as they are well documented in the original literature,[26–30] as well as in standard textbooks such as those by Bishop[33] and Goodfellow[34]. Detailed training conditions for the models are provided in our previous work cited above.

### 3.2 LIME method

The LIME method is an approach that estimates the components or regions of input data contributing to a prediction by using a local surrogate model, particularly for predictions made by ML and DL models whose inference processes are complex or effectively operate as black boxes. This method is also applicable to DL models that take images as input. The application of the LIME method to DL models is explained below.

First, the input image is divided into small regions as superpixels[35]. Next, multiple masked images are created by randomly altering the presence or absence of each superpixel. By inputting these into a trained deep learning model to obtain prediction values, pairs of prediction values for the masked images, i.e., the neighborhood dataset, are obtained. Furthermore, a local linear model $g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$ is trained to locally approximate the predictions of the original model. Here, $\boldsymbol{x}$ is a vector with dimensions equal to the total number of superpixels, where 0 and 1 are assigned to masked and unmasked regions, respectively. $\boldsymbol{w}^T$ is the weight vector, determined through training the linear model. In the LIME method, the superpixel regions corresponding to locations where the components of $\boldsymbol{w}^T$ have large weights are judged to be regions that contributed to classification in the original ML or DL model. In our subsequent research described below, we segmented SEM images and determined the regions that the DL model likely considered important for classification based on the magnitude of the weights $\boldsymbol{w}$. For a more mathematical background, refer to the original paper[31].

### 3.3 CycleGAN

CycleGAN is one of the style-translation algorithms[36] proposed by Zhu et al.[32] This algorithm is capable of mutually converting images obtained from two different domains. Here, a domain refers to a set of data obtained from specific imaging conditions and subjects. Unlike the similar algorithm pix2pix[37], CycleGAN is characterized by not requiring images with common fields or angles of view between domains. More specifically, for any sample $x \in X$ and $y \in Y$ from two different domains $X$ and $Y$, CycleGAN constructs a pair of translators $X \leftrightarrow Y$ based on the features of each domain.

**Figure 5** shows a schematic diagram of the CycleGAN architecture. CycleGAN contains a total of four DL models internally: $G_{Y\to X}$, $G_{X\to Y}$, $D_X$, and $D_Y$. Among these, $G_{Y\to X}(G_{X\to Y})$ is called the Generator. It takes an image $y \in Y (x \in X)$ sampled from domain $Y(X)$ as input and converts it into an image that aligns with the characteristics of domain $X(Y)$. Meanwhile, $D_X(D_Y)$ is the discriminator that determines whether an input image corresponds to a real image $x(y)$ or a translated image $G_{Y\to X}(y)\,(G_{X\to Y}(x))$. During CycleGAN training, the generators and discriminators are optimized adversarially, with the former attempting to deceive the latter and the latter improving their ability to correctly identify real and generated images.

In this study, the FE-SEM and W-SEM datasets were treated as separate domains ($X\to$FE, $Y\to$W), and a CycleGAN model was trained. The detailed network architecture was based on the model proposed by Nain.[38, 39] In the following, the "fake" SEM images generated by the trained generators $G_{\mathrm{FE}\to\mathrm{W}}$ and $G_{\mathrm{W}\to\mathrm{FE}}$ are referred to as FE2W-SEM and W2FE-SEM, respectively.

## 4. Results and Discussion

This chapter presents our research results and discusses them.

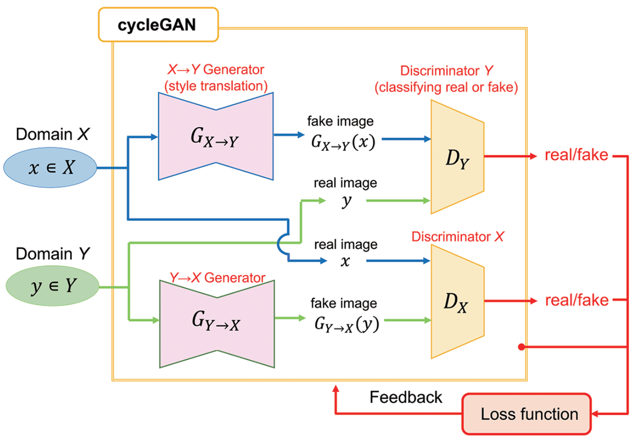


**Fig. 5 Schematic of cycleGAN**

We first provide a summary of the results, followed by detailed descriptions in subsequent sections. Finally, we discuss these results and outline future prospects for steel microstructure identification using machine learning techniques.

### 4.1 Brief summary

This section provides a summary of the research activities conducted by our research group. As shown in **Fig. 6**, multiple ML and DL models were applied to the SEM and EBSD datasets.

In approach (1) shown in Fig. 6, ML models were trained independently using crystallographic features such as variant boundary density and KAM value, and morphological features derived from SEM images.[7–9] When using crystallographic features, we selected those that were informative for classification. By reducing the dimensionality of the dataset from five to two, it was found that the regions occupied by each microstructure could be visualized as a map of microstructures on a two-dimensional plane spanned by the crystallographic features. In this process, the selection of an appropriate ML model was found to be crucial for the visualization of the microstructural map, and SVM based on a linear kernel was found to be effective. The ML model trained on crystallographic features achieved an accuracy of 87.5%.

Furthermore, when training RF and LightGBM models based on decision trees using morphological features extracted from SEM images, and morphological features obtained from the GLCM method, these models achieved a maximum accuracy of 87.5%.[11] Additionally, it was found that models trained solely on SEM images obtained from either the FE or W source exhibited significantly reduced identification accuracy when applied to SEM images from the other source.

Next, in approach (2) shown in Fig. 6, ResNet50 was employed as the classifier, and the classification accuracy reached approximately 94% regardless of the SEM imaging conditions. This performance was found to be superior to that of approach (1). Furthermore, the LIME method was applied to ensure interpretability of the DL model. The results confirmed that the trained ResNet50 model emphasized typical morphologies of martensite and bainite structures.[12, 13]

Moreover, the accuracy of ResNet50 was reduced to approximately 40% for SEM images acquired from different SEM sources than those used for training, consistent with the behavior observed in approach (1). By applying CycleGAN to convert the imaging conditions between FE-SEM and W-SEM, the classification accuracy was improved to approximately 90%.[14]

As mentioned above, for the classification of the SEM image dataset, both approaches (1) and (2) divided each test image into smaller patches (mini-images), which were then used for classification. In this study, two evaluation schemes were adopted: one in which the classification accuracy was evaluated for each individual mini-image independently (non-voting), and another in which the classification results of the mini-images corresponding to a single original image were aggregated by majority voting to evaluate the classification accuracy of the original image (voting)[40]. In practical visual inspection, microstructural identification is also performed by integrating locally observed morphological features; therefore, the voting scheme can be interpreted as mimicking the human decision-making process in microstructure classification. A schematic illustration of these concepts is shown in **Fig. 7**.

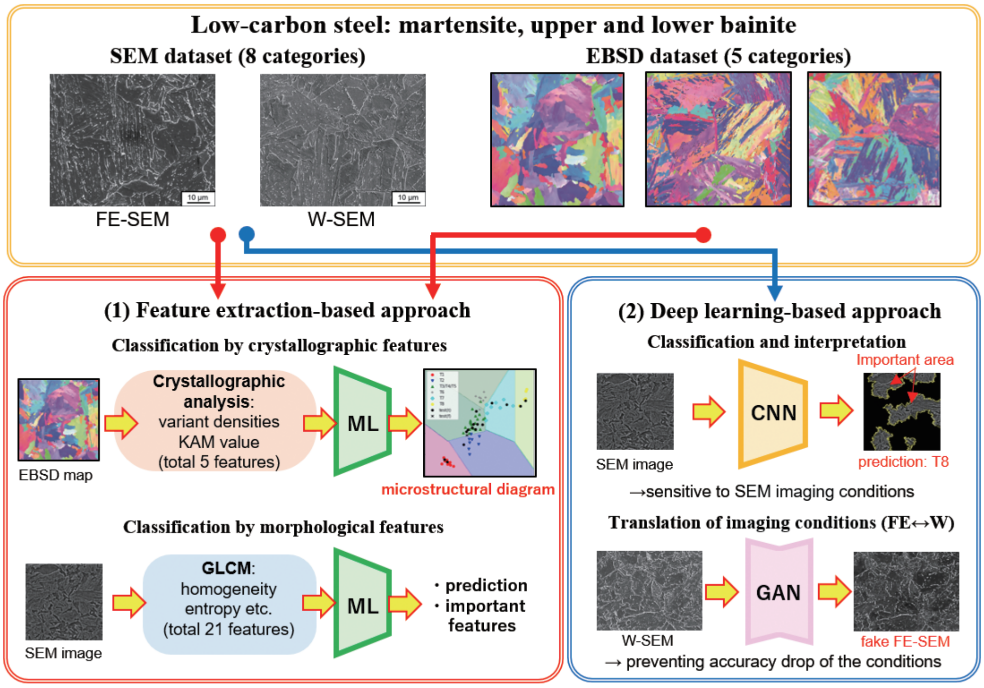


**Fig. 6 Summary of our study on identification for low-carbon steel microstructures**

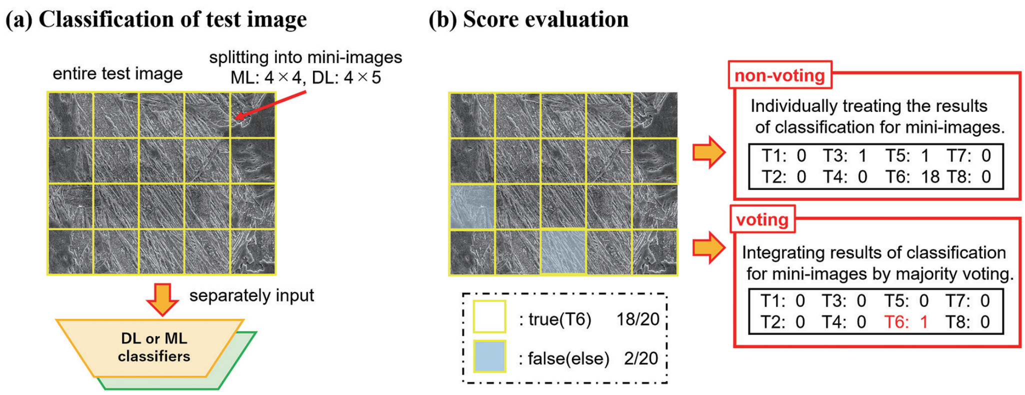


**Fig. 7 Schematic of the classification strategy for SEM dataset in this study**

### 4.2 Application of machine learning using feature extraction

#### 4.2.1 Microstructure classification and microstructural diagram using EBSD data

Here, the results of microstructure classification using the variant boundary densities and the KAM value are presented. Prior to model training, outlier removal and feature selection were performed on a total of 96 five-dimensional data points, consisting of 12 data points from each of the T1–T8 specimens. An overview of this procedure is shown in **Fig. 8**. Specifically, after standardizing the data, outliers were removed using the interquartile range (IQR), defined as the range from the 25th to the 75th percentile.[41] As a result, the total number of data points was reduced to 78. Furthermore, in this analysis, two-dimensional classification was explored by selecting two features from the five available features and searching for combinations in which the data points were widely distributed in the corresponding feature space. Consequently, the variant boundary density corresponding to high-angle grain boundaries with misorientations of approximately 20° (V1–V6) and the KAM value were selected.[6] Furthermore, microstructure classification using the EBSD dataset found it difficult to sufficiently accurately distinguish between T3, T4, and T5, which have equivalent martensite fractions. Therefore, these categories were treated as identical (total of 5 categories) and used for training the ML model. An SVM was employed for the ML model, with a linear kernel selected. The accuracy of the resulting classification model, obtained through five-fold cross-validation, was 87.5%.

The trained SVM predicts the microstructural category by taking the values of the V1–V6 variant boundary density and the KAM value as inputs. By projecting the predicted labels onto a two-dimensional plane, the regions occupied by different microstructures in the feature space can be visualized. The results are shown in **Fig. 9**. Our research group refers to this representation as a microstructural diagram[7–9]. Such diagrams enable newly acquired samples to be positioned relative to the microstructures included in the training dataset. For example, by constructing microstructural diagrams in advance for representative steel grades, it becomes possible to readily assess the type of microstructure obtained in newly designed materials.

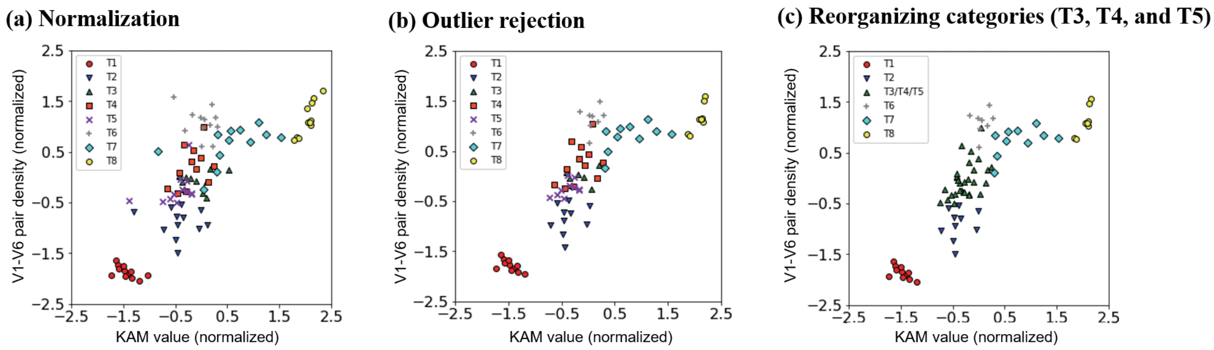


Fig. 8 The scatter plot of V1-V6 density and KAM values procced after (a) normalization, (b) outlier rejection, and (c) treating T3, T4, and T5 as the same category.

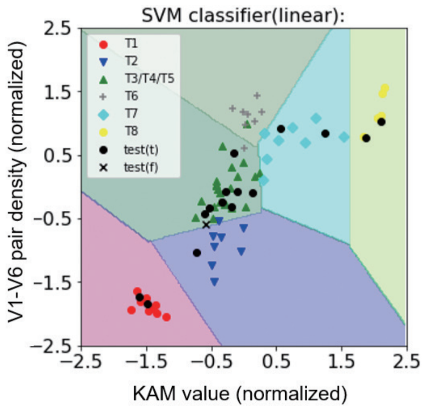


Fig. 9 Microstructural diagram for low-carbon steel
The space of visualization is spanned by V1-V6 pair density and KAM value[8].

#### 4.2.2 Classification of SEM datasets using morphological features

For the FE-SEM and W-SEM image datasets, RF and LightGBM models were trained using a total of 19 features extracted by the GLCM method. In this process, the FE-SEM and W-SEM datasets were treated independently. The classification accuracies of the models on the test data are shown in **Fig. 10**. First, when the imaging conditions of the training and test data were identical, both RF and LightGBM were confirmed to operate appropriately. In terms of performance, LightGBM tended to achieve higher classification accuracy than RF. In addition, comparison between the non-voting and voting schemes revealed that the latter consistently outperformed the former by more than 10%, demonstrating the effectiveness of the voting approach. The highest classification accuracy was obtained by the LightGBM model trained on W-SEM images, achieving 87.5% accuracy under the voting scheme.[11] For reference, a similar classification was attempted for the SEM dataset examined in this study using the LBP method[24], but the accuracy was approximately 80%, which was inferior to that obtained using the GLCM-based features.

When the imaging conditions of the training and test data differed, a significant degradation in classification accuracy was observed, regardless of the model or the evaluation scheme employed. In addition, the accuracy obtained with the voting scheme was found to be lower than that of the non-voting scheme under these conditions. This indicates that most of the patch-level classification results associated with a single original image were incorrect, and that aggregating these misclassifications by voting further exacerbated the overall performance degradation.

Finally, an investigation into the features emphasized by RF and LightGBM during classification revealed that homogeneity was consistently selected as an important feature under all training conditions[11]. Homogeneity is a texture descriptor that responds to surface flatness in images. Within the present dataset, the primary contributors to image flatness are the size of the $\alpha$' laths and carbides. This finding is consistent with the morphological criteria for visual classification discussed in Subsection 2.2.2, in which the morphology of lath-like shaped $\alpha$' grains and the distribution of carbides were identified as key factors. Thus, through the extraction of GLCM features and the evaluation of their importance, it was demonstrated that indicators analogous to those used in visual inspection are also critical for classification based on machine learning.

### 4.3 Microstructure classification using deep learning models

#### 4.3.1 Performance of ResNet50

**Figure 11** presents the classification accuracy of ResNet50 trained on the SEM image dataset[12, 13]. ResNet50-FE and ResNet50-W represent the models trained on FE-SEM and W-SEM images, respectively. When the imaging conditions of the training and test datasets match, ResNet50 achieves accuracies above 94% for both non-voting and voting schemes, with a modest improvement observed when voting is applied.

When the imaging conditions of the training and test data differed, the classification accuracy decreased, consistent with the results described in Subsection 4.2.2. In particular, intermediate representations obtained from layers near the output of these ResNet50 models were clustered using the t-distributed stochastic neighbor embedding (t-SNE) method[42]. When the imaging conditions of the training and test data were identical, the plots of the test images were localized according to their respective classification classes. In contrast, when the imaging conditions differed between the training and test datasets, the corresponding plots were intermixed irrespective of the classification classes[13]. These results clearly demonstrate the importance of maintaining dataset quality when developing automated classification techniques based on machine learning.

#### 4.3.2 LIME-based interpretation of ResNet50 classification

As described above, ResNet50 was confirmed to achieve higher classification accuracy than feature-based methods; however, it is difficult to directly determine which regions of the image are emphasized during the classification process. Therefore, the LIME method was applied to the trained ResNet50-W model. The results

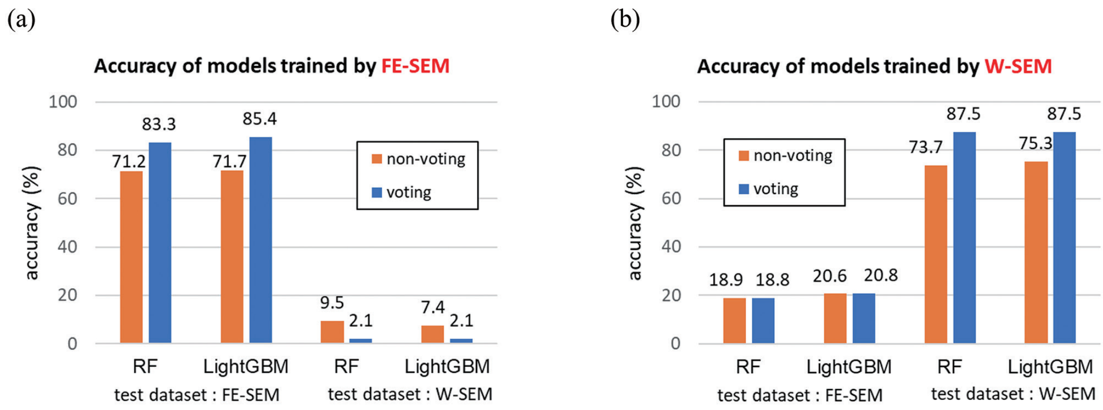


**Fig. 10 Accuracies of RF and LightGBM trained by GLCM features**
**(a) accuracies from models trained by the FE-SEM dataset and (b) those from models trained by the W-SEM dataset**

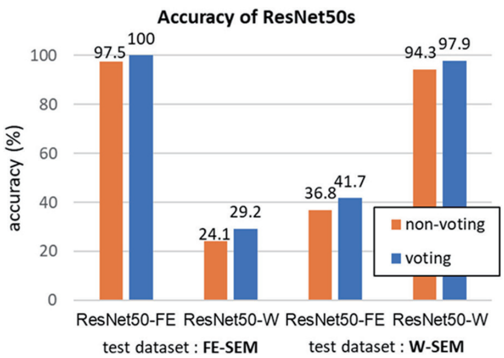


**Fig. 11 Accuracies of SEM dataset by ResNet50s**
**The notations of ResNet50-FE (ResNet50-W) mean the model trained by FE-SEM (W-SEM) images.**

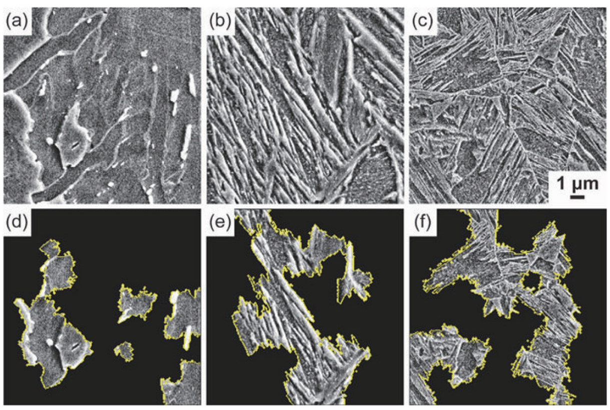


**Fig. 12 Input SEM images and LIME output images of T1 (UB), T6 (LB), and T8 (MS)**
**(a), (b), and (c) ((d), (e), and (f)) correspond to the input (output) images of T1, T6, and T8, respectively.**
**Source: reproduced from Maemura et al. (2020)[12], with permission**

obtained for W-SEM images of UB, LB, and MS are shown in **Fig. 12**. For UB, coarse α' grains and grain-boundary carbides were identified as important regions for classification, whereas for LB, lath-shaped grains and intragranular carbides were emphasized. In the case of MS, fine lath structures and lath initiation sites (packet boundaries) were highlighted. These results demonstrate that ResNet50 focuses on image regions that are consistent with conventional criteria[3] used in visual microstructure identification.

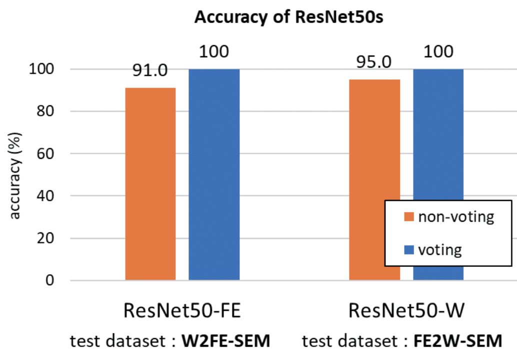


**Fig. 13 Accuracies of ResNet50 models for the translated images by CycleGAN**

### 4.3.3 Translation of imaging conditions using CycleGAN

As described above, differences in imaging conditions between the training and test datasets were found to cause significant degradation in the classification accuracy of both ML and DL models. In this subsection, we first demonstrate the extent to which the classification accuracies of ResNet50-FE and ResNet50-W can be improved by translating the imaging conditions between FE-SEM and W-SEM using CycleGAN, and subsequently provide a visual assessment of the quality of the translated images.[14]

**Figure 13** shows the classification accuracy of ResNet50 before and after the application of CycleGAN. As can be seen from the figure, the classification accuracy of the ResNet50 models, which had dropped to approximately 40% due to mismatched imaging conditions between the training and test data, recovered to over 90%, comparable to the performance achieved when the imaging conditions were identical. These results indicate that image translation using CycleGAN is an effective means of mitigating the degradation in classification accuracy caused by variations in imaging conditions.

Next, to visually evaluate the quality of the style translation achieved by CycleGAN, FE-SEM and W-SEM images before and after translation are shown in **Fig. 14**. The SEM images presented in this figure were acquired such that the fields of view overlap between the FE-SEM and W-SEM images, and it should be noted that these images are not included in the SEM image dataset introduced in Section 2.2. As can be seen from the figure, the distinct high-contrast texture characteristic of FE-SEM images and the comparatively flat texture characteristic of W-SEM images are successfully translated between the two domains.

On the other hand, the translated images were reconstructed by stitching together mini-images cropped at a resolution of 224 × 224 px after translation, and brightness variations can be observed among individual mini-images. This behavior reflects the variation in image brightness present in the original dataset. This issue could potentially be mitigated by increasing the input image size or by reducing brightness variations within the dataset through appropriate preprocessing.

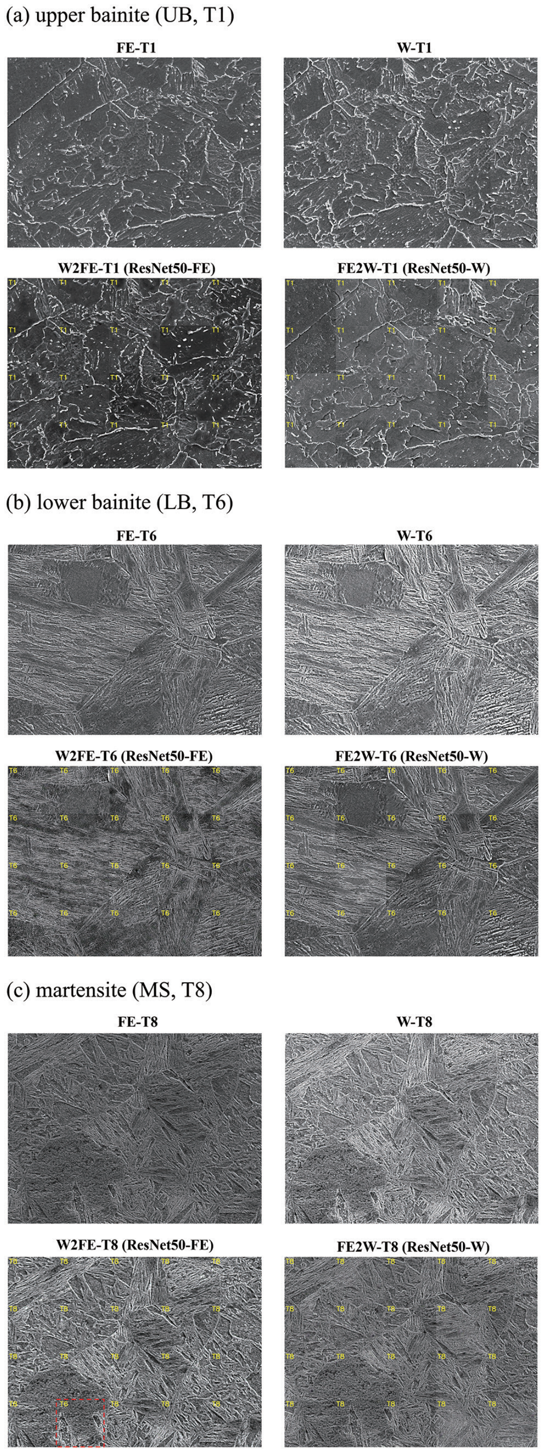


**Fig. 14 FE-SEM and W-SEM images obtained from the same field of view of (a) UB (T1), (b) LB (T6), and (c) MS (T8)**
**Two pictures at upper side in (a), (b), and (c) correspond to the original FE-SEM and W-SEM images, and those of the lower side correspond to the translated images (left: W2FE, right: FE2W generators). Mini-images with a red dashed line represent incorrect prediction by the ResNet50 trained on the corresponding domain.**
**Source: reproduced from Tsutsui et al. (2025)[14], with permission**

### 4.4 Discussion of results and future prospects

In this study, datasets were constructed through EBSD and SEM measurements for eight low-carbon steel samples subjected to various heat treatments, and machine learning techniques were applied to these datasets to obtain the following results.

(1) In the feature-based approach using ML models, it was demonstrated that microstructural diagrams can be constructed by employing an SVM model trained on the variant boundary densities and KAM value derived from EBSD analysis. Furthermore, for the SEM dataset, it was confirmed that a maximum classification accuracy of 87.5% can be achieved by combining GLCM-based features with decision tree–based ML models.

(2) In the image classification approach using DL models, a higher classification accuracy exceeding 90% was achieved compared with approach (1). In addition, visualization of the basis for DL model predictions using the LIME method revealed that the model focuses on image regions that are consistent with conventional criteria for visual identification of steel microstructures.

Furthermore, for both approaches (1) and (2) applied to the SEM dataset, it was found that discrepancies in imaging conditions between the training and test data significantly degraded the classification accuracy. By applying CycleGAN-based style translation, this performance degradation was shown to be effectively mitigated. Based on these results, future technological perspectives are discussed below.

In this study, we applied machine learning techniques to datasets with a relatively limited number of classification categories and confirmed high classification accuracy. These findings suggest that practical performance can be achieved by constructing suitable datasets and training ML or DL models tailored to specific, localized tasks. However, in real manufacturing settings, machine learning models are required to handle datasets that involve substantially greater diversity in acquisition conditions, data modalities, target objects, and application scenarios.

Although constructing mathematical models capable of handling such diverse datasets is not straightforward, one promising approach at present is to leverage large foundation models, particularly large language models (LLMs)[43, 44]. LLMs constitute a class of large-scale models capable of tasks such as text generation, summarization, and question answering, with ChatGPT[45] and DeepSeek[46] being representative examples. These models are all based on architectures incorporating self-attention mechanisms. In particular, ChatGPT possesses image generation capabilities by transmitting protocols generated from input text to external diffusion models. Meanwhile, DeepSeek achieves high accuracy and rapid response by maintaining multiple networks specialized for domain-specific tasks, referred to as expert networks. Such an architecture is known as a Mixture of Experts (MoE)[47].

Indeed, methods for training LLMs tailored to use within specific companies or communities have been proposed.[48, 49] One such approach is adapter injection, in which small-scale networks (adapters) are incorporated into an LLM and only these components are trained.[50] In addition, attempts have been made to construct MoE architectures composed of experts with advanced knowledge in specific domains and to integrate them into existing LLMs.[51] In fact, the application of LLMs to the field of materials science, including steel materials, is already underway[52, 53], and further progress in this area is highly anticipated. In this context, the results of the present study can be interpreted as opening a pathway toward the construction of expert networks and adapters responsible for the identification and interpretation of steel materials, which represent the final products in the steel manufacturing industry.

Finally, it should be emphasized that the high-accuracy machine learning models developed in this study are fundamentally supported by measurement technologies for constructing appropriate datasets, as well as by the accumulation of crystallographic and morphological knowledge. Similarly, the future perspectives involving LLMs discussed above cannot be realized without a foundation for acquiring high-quality measurement data. Further advancement in this research field therefore requires complementary progress in materials science–based understanding of steel materials and information science–based algorithms, including machine learning.

## 5. Conclusions

In this paper, we provided an overview of our research group's efforts to apply machine learning techniques to the analysis of microstructures in low-carbon steels. Datasets were constructed using EBSD and SEM measurements for eight types of steels containing martensite and bainite, and two approaches were employed: (1) the construction of ML models based on feature extraction, and (2) the construction of DL models based on SEM images. As a result, the following findings were obtained.

(i) It was demonstrated that microstructural diagrams can be constructed by employing ML models trained on the variant boundary densities and KAM value calculated from EBSD crystallographic orientation maps.

(ii) By extracting microstructure-related features derived from SEM images using the GLCM method and training ML models, a maximum classification accuracy of 87.5% was achieved. Furthermore, the feature that contributed most significantly to classification was homogeneity, which corresponds to surface flatness. This feature responds to the size of $\alpha$' grains and carbides, and is consistent with conventional criteria used in visual microstructure identification.

(iii) In the application of DL models to SEM images, classification accuracies exceeding 90% were achieved. Moreover, by applying the LIME method, it was revealed that the DL models focus on regions such as carbides and lath initiation sites, which are also emphasized in visual inspection.

(iv) In both (ii) and (iii), it was found that discrepancies between the training and test datasets led to a significant degradation in the performance of trained models. As a remedy, it was demonstrated that GAN-based style translation can effectively restore the classification accuracy.

These findings indicate that practical classification performance can be ensured by constructing appropriate datasets and training ML or DL models specialized for localized tasks. Building upon these results, further developments—such as the construction of LLMs specialized for steel materials by incorporating these models—are expected in future research.

**Acknowledgments**

In conducting this research, we received valuable suggestions and support from Mr. Yuya Nakao of NS Solutions Corporation. We express our sincere gratitude to him.

**References**

1) Jorge, J.C.F. et al.: J. Mater. Res. Technol. 10, 471–501 (2021)
2) Tasan, C.C. et al.: Annu. Rev. Mater. Res. 45, 391–431 (2015)
3) Bramfitt, B.L., Speer, J.G.: Metall. Trans. A. 21, 817–829 (1990)
4) Morito, S. et al.: Acta Mater. 54 (19), 5323–5331 (2006)
5) Tsutsui, K. et al.: ISIJ Int. 64 (11), 1619–1640 (2024)
6) Terasaki, H. et al.: Mater. Charact. 129, 305–312 (2017)
7) Tsutsui, K. et al.: Comput. Mat. Sci. 159, 403–411 (2019)
8) Tsutsui, K. et al.: CAMP-ISIJ. 32 (2), 618 (2019)
9) Terasaki, H. et al.: Journal of the Japan Welding Society. 88 (7), 536–539 (2019)
10) Tsutsui, K., Moriguchi, K.: Calphad. 74, 102303 (2021)
11) Tsutsui, K. et al.: Mater. Today Comn. 25, 101514 (2019)
12) Maemura, T. et al.: Mater. Trans. 61 (8), 1584–1592 (2020)
13) Tsutsui, K. et al.: Mater. Today Comn. 32, 104062 (2022)
14) Tsutsui, K. et al.: Mater. Trans. 66 (8), 941–952 (2025)
15) Gourgues, A.-F. et al.: Mater. Sci. Technol. 16 (1), 26–40 (2000)
16) Terasaki, H. et al.: Metall. Trans. A. 48 (12), 5761–5765 (2017)
17) Takayama, N. et al.: Acta Mater. 60 (5), 2387–2396 (2012)
18) Morito, S. et al.: Mater. Today: Proc. 2, S913-S916 (2015), International Conference on Martensitic Transformations, ICOMAT- 2014, Bilbao, Spain, (2014)
19) Haralick, R.M.: Proc. IEEE. 67 (5), 786–804 (1979)
20) Haralick, R.M. et al.: IEEE Trans. Syst. Man Cybern. SMC-3, 6, 610–621 (1973)
21) Webel, J. et al.: Mater. Charact. 144, 584–596 (2018)
22) Gola, J. et al.: Comput. Mater. Sci. 160, 186–196 (2019)
23) Ojala, T. et al.: Proceedings of 12th International Conference on Pattern Recognition, Jerusalem, Israel, vol.1, 582–585 (1994)
24) Ojala, T. et al.: IEEE Transactions on Pattern Analysis and Machine Intelligence. 24 (7), 971–987 (2002)
25) Decost, B.L. et al.: Acta Mater. 133, 30–40 (2017)
26) Vapnik, V., Lerner, A.: Autom. Remote Control. 24, 774–780 (1963)
27) Boser, B.E. et al.: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92, ACM, New York, USA, 144–152 (1992)
28) Breiman, L.: Mach. Learn. 45 (1), 5–32 (2001)
29) Ke, G. et al.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems. 3149–3157 (2017)
30) He, K. et al.: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778 (2016)
31) Ribeiro, M.T. et al.: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, United States, 1135–1144 (2016)
32) Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: arXiv:1703.10593 (2017)
33) Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, 2006
34) Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning, Adaptive Computation and Machine Learning series. MIT Press, 2016
35) Neubert, P., Protzel, P.: Compact Watershed and Preemptive SLIC: On Improving Trade-offs of Superpixel Segmentation Algorithms. ICPR 2014, 996–1001 (2014)
36) Lee, J.-W., Goo, N.H., Park, W.B., Pyo, M., Sohn, K.-S.: Eng. Rep. 3, e12274 (2021)
37) Isola, P. et al.: arXiv:1611.07004 (2016)
38) Nain, A.K.: https://keras.io/examples/generative/cyclegan/ (cited 2025 08-31)
39) Nain, A.K.: https://github.com/keras-team/keras-io/blob/master/examples/generative/cyclegan.py (cited 2025-08-31)
40) Zhang, B. et al.: Rapid Prototyp. J. 25 (2019)
41) Upton, G., Cook, I.: Understanding Statistics. OUP Oxford, 1996
42) Van der Maaten, L., Hinton, G.: J. Mach. Learn. Res. 9, 2579–2605 (2008)
43) Bommasani, R. et al.: arXiv:2108.07258 (2018)
44) Vaswani, A. et al.: Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 30, (2017) https://arxiv.org/

abs/1706.03762 (cited 2025 08-31)
45) Radford, A. et al.: Improving language understanding by generative pre-training. OpenAI Technical Report. (2018) https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (cited 2025 08-31)
46) Liu, A. et al.: arXiv:2405.04434 (2024)
47) Dai, D. et al.: arXiv:2401.06066 (2024)
48) Leng, J. et al.: Adv. Eng. Informatics. 65, A, 103179 (2025)
49) Hegiste, V. et al.: arXiv:2306.17645 (2023)
50) Hu, E. J. et al.: arXiv:2106.09685 (2021)
51) Souza, F. et al.: arXiv:2211.00558 (2022)
52) Tian, S. et al.: Acta Mater. 285, 120663 (2025)
53) Jiang, X. et al.: Npj Comput. Mat. 11, article no.79 (2025)

Kazumasa TSUTSUI
Ph.D., Senior Researcher
Mathematical Science & Technology Research Lab.
Advanced Technology Research Laboratories
20-1 Shintomi, Futtsu City, Chiba Pref. 293-8511

Koutarou HAYASHI
Ph.D., Manager, Head of Section
Materials Characterization Research Lab.
Advanced Technology Research Laboratories

Koji MORIGUCHI
Ph.D., Executive General Manager
Industry & Resource Circulation Technology Div.
Sustainability Solution Unit
Nippon Steel Technology Co., Ltd.

Shigekazu MORITO
Ph.D., Professor
Department of Materials for Energy
Shimane University

Hidenori TERASAKI
Ph.D., Professor
Faculty of Advanced Science and Technology
Kumamoto University