frontiers | Frontiers in Materials

TYPE Original Research
PUBLISHED 10 August 2023
DOI 10.3389/fmats.2023.1222456

OPEN ACCESS

EDITED BY
Lei Chen,
University of Michigan–Dearborn, United States

REVIEWED BY
Brian Decost,
National Institute of Standards and Technology (NIST), United States
Pavlo Maruschak,
Ternopil Ivan Pului National Technical University, Ukraine

*CORRESPONDENCE
Ali Riza Durmaz,
ali.riza.durmaz@iwm.fraunhofer.de

RECEIVED 14 May 2023
ACCEPTED 31 July 2023
PUBLISHED 10 August 2023

CITATION
Durmaz AR, Potu ST, Romich D, Möller JJ and Nützel R (2023), Microstructure quality control of steels using deep learning.
Front. Mater. 10:1222456.
doi: 10.3389/fmats.2023.1222456

Microstructure quality control of steels using deep learning

Ali Riza Durmaz[1,2]*, Sai Teja Potu[1,3], Daniel Romich[4], Johannes J. Möller[4] and Ralf Nützel[4]

[1]Group of Meso and Micromechanics, Fraunhofer Institute for Mechanics of Materials IWM, Freiburg im Breisgau, Germany, [2]Chair for Micro and Materials Mechanics, University of Freiburg, Freiburg im Breisgau, Germany, [3]Institute of Mechanics and Computational Mechanics, Leibniz University Hannover, Hannover, Germany, [4]Materials Technology, Schaeffler Technologies AG & Co. KG, Schweinfurt, Germany

In quality control, microstructures are investigated rigorously to ensure structural integrity, exclude the presence of critical volume defects, and validate the formation of the target microstructure. For quenched, hierarchically-structured steels, the morphology of the bainitic and martensitic microstructures are of major concern to guarantee the reliability of the material under service conditions. Therefore, industries conduct small sample-size inspections of materials cross-sections through metallographers to validate the needle morphology of such microstructures. We demonstrate round-robin test results revealing that this visual grading is afflicted by pronounced subjectivity despite the thorough training of personnel. Instead, we propose a deep learning image classification approach that distinguishes steels based on their microstructure type and classifies their needle length alluding to the ISO 643 grain size assessment standard. This classification approach facilitates the reliable, objective, and automated classification of hierarchically structured steels. Specifically, an accuracy of 96% and roughly 91% is attained for the distinction of martensite/bainite subtypes and needle length, respectively. This is achieved on an image dataset that contains significant variance and labeling noise as it is acquired over more than 10 years from multiple plants, alloys, etchant applications, and light optical microscopes by many metallographers (raters). Interpretability analysis gives insights into the decision-making of these models and allows for estimating their generalization capability.

KEYWORDS

quality control, microstructure, grain size, steel, martensite, bainite, deep learning

1 Introduction

Materials in many applications are exposed to complicated loading conditions. Along the value chain, components and materials therein are exposed to process scatter at all stages. This establishes the demand for quality control. The presence of major structural defects in components can be excluded through a variety of non-destructive testing methods which exploit ultrasonic or magnetic sensing principles for instance. Most sensor principles that are applicable in-line, however, retrieve integral information of large volumes as they exhibit neither an adequate spatial resolution nor signal sensitivity to measure subtle microstructural heterogeneity. Moreover, most sensors provide compounded information on residual stresses, defect density, grain size, chemical composition, and more.

Therefore, when particular microstructural aspects such as grain size distribution are of interest, most industries fall back on destructive sectioning and direct imaging methodologies, which are being performed on small sample sizes. A typical example of this is hierarchically-structured steel microstructures, such as martensite or bainite, for which the quantification of feature sizes in the primary microstructure is only possible by imaging. Trained metallographers prepare metallographic cross-sections of components through consecutive cutting, polishing, and etching and then image them using light optical microscopy. The resulting micrographs, depending on the component's intended application and loading paths, can be inspected with respect to different microstructural aspects. In bearing steels in which often plate martensite or bainite is present, the contrasted cross-section shows acicular, i.e., needle-shaped, structures (Bepari, 2017). In this case, the so-called *needle length* affects the material's resistance to rolling contact fatigue (Shur et al., 2005) and is thus of pronounced interest. These microstructures are the consequence of the partitioning of the high-temperature austenite phase to a martensitic or bainitic microstructure. Some examples of martensitic steels with varying cooling rates culminating in different morphologies and phase compositions are depicted in Figure 1. The microstructures are composed of martensite or bainite ('M/B') as the primary microstructure as well as dispersed minor phases such as retained austenite (bright constituents annotated with 'RA') and carbide particles annotated with 'C'. While both minor phases occur bright, the carbide particles are either circular or elliptic and slightly smaller than the irregular-shaped retained austenite constituents. Depending on the exact treatment conditions both minor phases are differently distributed.

In this work, two classification problems for hierarchically-structured steel microstructures are tackled. The first is the classification of the needle length as depicted in Figure 1 and listed in Table 1. This classification is inspired by the micrographic grain size categorization standard according to ISO 643 which is usually applied for equiaxed and unimodally distributed microstructures. Henceforth, we refer to this task as 'grain size', 'needle morphology', or 'needle length' classification. The classification is applied to bearing steels that underwent different hardening heat treatments. Specifically, through-hardening steels such as 100Cr6, 100CrMnSi6-4 in bainitic (B) and martensitic (G) states as well as quenched and tempered martensitic C56E2 steel (M) are taken into consideration. A distinction between these three subtypes is also of interest as these entail different material properties. This poses the second classification objective. In the following, we refer to the joint information of microstructure subtype and needle length, e.g., 'G7', as *structure code*. By attempting the microstructure subtype distinction, we investigate whether nuanced differences between bainitic and martensitic steel variants (Hillert, 1995) can be identified by computer vision approaches. In contrast to efforts by Müller et al. (2020); Gola et al. (2018); Zhu et al. (2022), we attempt a macroscopic distinction on large field-of-view light optical micrographs without discerning substructures at a prior austenite grain level. Since the bainite and martensite subtype labels in the work at hand are provided from the quantitative temperature-time profile during cooling, this classification problem resembles the one presented by Bulgarevich et al. (2019). The classification of needle length following ISO 643 has not been reported in the literature to the best of our knowledge. Published work at the intersection between DL and metallography, to date, entailed image datasets acquired under comparatively controlled and repeatable conditions, see DeCost et al. (2019); Durmaz et al. (2021). In this work, on the other hand, the difficulty lies in the many degrees of freedom present in the process chain causing a profound data variance.

The dataset utilized in this work exhibits a pronounced variance as it covers multiple alloys, heat treatments, polishing protocols, storage times, etchant concentrations, etching durations, and image acquisition parameters. All images are acquired in industrial testing laboratories over a time span of more than 10 years, and their needle length labels (6–10) are assigned by numerous metallographers. A single micrograph is typically inspected by a single metallographer. As the distinction of ISO 643 grain sizes 9 and 10 does not have any application relevance for these bearing steels, both classes in this work are aggregated to the classes 9–10 for each microstructure subtype. Further, since image instances with coarse acicular needles are very scarce in industrial processing, and thus data availability is low, they will be excluded from the classification task, see Table 1. Segregation in the material, especially of carbon, chromium, and nickel can influence the local formation of the hardened microstructure and cause fluctuations in the needle morphology. Depending on the heterogeneity of the microstructure, different criteria have been applied to rate the grain size, see Table 1 and Section 5.1.

In an attempt to render both image classification tasks automated and objective, we apply different deep learning (DL) methodologies. To be precise, two approaches are employed which are illustrated in Figures 2A, B. The first represents a single multi-class classification model which confronts both classification challenges simultaneously, i.e., direct prediction of the structure code. In contrast, a second methodology is presented in which one model categorizes the micrographs based on the microstructure subtype they are showing, and another model distinguishes the image in terms of its needle morphology. Depending on the first model's subtype prediction, a dedicated needle morphology classifier model is selected to perform the needle length classification. Comparison of both approaches allows investigation of whether decomposing the problem and addressing the partial tasks with specialized models can be beneficial.

In order to evaluate the inter-rater reliability of both image classification tasks, a round-robin test is designed and presented here. Thereby, the subjectivity of the subtype and needle morphology classification can be assessed. Further, the round-robin results act as a baseline for the computational assessment.

2 Results

In the following, the overall results are presented for both strategies, the global model and the two-stage (2S) approach. The results are listed in Table 2. Two common deep learning architectures are taken into consideration, ResNet-50 and ResNet-18. Moreover, the 'fw' annotation in the table indicates models trained with frozen weights in the feature extractor portion of the model. As an evaluation metric, accuracy is provided. The accuracy,

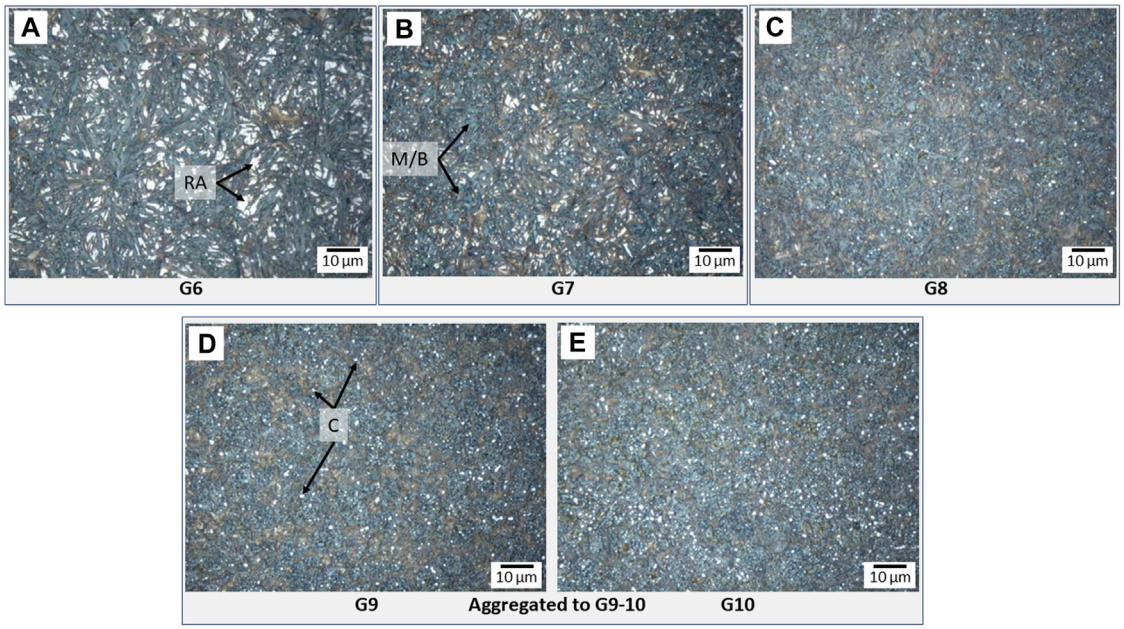

FIGURE 1
Martensitic steels distinguished by their needle morphology. Subfigures **(A–E)** each illustrate a different needle morphology ranging from G6 to G10. The arising martensite/bainite (M/B), retained austenite (RA), and carbide (C) constituents are annotated in the figure. The subfigure captions represent the structure codes associated with the grain size according to the ISO 643 standard. Note that this set of images does not represent the variance of the complete dataset as it only takes a single subtype into consideration. A random set of images showcasing a more realistic representation of the data scatter is depicted in Figure 8.

for the data and models at hand, virtually coincides with the F_1-score. In a repeatability study, where three distinct test-train splits were sampled for the bainite 'B' grain size classification using three-fold cross-validation, the data sampling-induced fluctuations were confirmed to be negligible. Therefore, we report single training performances here.

The combined accuracy for both tasks reaches up to 90.49% using ResNet-18 as an architecture within a two-stage approach. On the dataset at hand, training the ResNet-18 architecture culminates in a better performance than ResNet-50, which in turn outperforms the case in which only the classifier portion of the ResNet-18 was optimized (fw). The two-stage approach marginally yet consistently outperforms the scenario in which a single model performs both classification tasks. It is evident that the merit of the two-stage approach depends on the architectural choice and the training strategy. For instance, the two-stage ResNet-18 (fw) with frozen feature extractor weights outperforms its global counterpart by roughly 4%. In contrast, in both fully tuned examples, the performance improvement through the two-stage approach is less and amounts to 1.2% and 0.3% for the ResNet-50 and ResNet-18, respectively. The model's performances can be further dissected by Table 3 and the confusion matrices provided in Figures 3, 4. In the table, the model performances of each task-specific model contributing to the two-stage model are listed.

From this table, it can be observed that the fully trained ResNet-18 variant improves on or matches the performance of the ResNet-50 variant throughout all grain size classification tasks. In contrast, the ResNet-50 performs slightly better at distinguishing the material subtypes. Moreover, it is apparent that the material subtype classification out of all tasks reaches the highest performance approaching 97% accuracy. Thus, interestingly, the models manage to distinguish the martensitic, bainitic through-hardened, and martensitic through-hardened subtypes very well. This is reflected by the confusion matrix of the ResNet-18 2S model depicted in Figure 4 where it is illustrated that the remaining error cases predominantly arise from confusing the grain size categories (8 and 9–10) associated with small needle features in the B and G class.

In terms of grain size classification, the three models dedicated to the three subtypes show some variation in their performance. On the larger G and B subsets, a performance of 93.9% and 86.4% is attained, respectively. In the confusion matrix in Figure 4, it can be seen that the performance of the bainite grain size model falls short mainly because a notable portion of B instances was predicted as adjacent structure codes. The underlying reason for this will be further explored in the discussion section. Note, that the performance of the grain size models for martensitic grades (M) being seemingly unaffected by the architecture choice could be owed to the small size of the test set measuring only 14 images. One aspect to consider when selecting an optimal model is whether over- or underestimation of needle length is more problematic. In such safety-critical applications, conservative model predictions skewed towards predicting larger grains are preferable. This means that instances below the main diagonal should be minimized. However, when comparing the confusion matrices in Figures 3, 4, it can be observed that the two-stage model is slightly skewed towards predicting smaller grains, i.e., higher structure codes. This is the case despite applying imbalance correction within each subtype-specific grain size model of the two-stage model.

10.3389/fmats.2023.1222456

TABLE 1 Steel grade families and grain sizes considered in this study.

Material subtype	Grain size (ISO 643)	Structure code	Needle morphology class (mean crit.)	Needle morphology class (max crit.)
Martensitic 100Cr6 (G)	6	G6	Coarse acicular to acicular	Coarse Acicular
	7	G7	Acicular	Coarse acicular to acicular
	8	G8	Fine acicular	Acicular
	9 – 10	G9–G10	Fine acicular to structureless	Fine acicular
Martensitic C56E2 (M)	6	M6	Coarse acicular to acicular	Coarse Acicular
	7	M7	Acicular	Coarse acicular to acicular
	8	M8	Fine acicular	Acicular
	9 – 10	M9–M10	Fine acicular to structureless	Fine acicular
Bainitic 100Cr6 (B)	6	B6	Coarse acicular to acicular	Coarse Acicular
	7	B7	Acicular	Coarse acicular to acicular
	8	B8	Fine acicular	Acicular
	9 – 10	B9–B10	Fine acicular to structureless	Fine acicular

The classes corresponding to the rows with gray background color are discarded for model training. There are needle length ranges for typical and sporadically occurring large needles associated with the mean and maximum criterion, respectively. Generally, the needle length ranges for both criteria are disjunct. Depending on the measured lengths and their frequency, a metallographer decides which criterion to apply for the micrograph classification. The exact thresholds can not be provided due to confidentiality.

3 Discussion

3.1 Round robin test and inter-rater reliability assessment

A round-robin test was performed in which 14 metallographers trained on the needle length classification task participated in judging twenty micrographs. Those micrographs were selected from the test set based on the outcome of an intermediate DL model such that some correctly and some incorrectly classified micrographs were picked for each subtype and some grain sizes. In general, the images selected pose rather difficult instances since twelve out of twenty were misclassified by the intermediate deep learning model as well. This is in line with comments by the raters who deemed some images over-etched and thus difficult to classify. We opted for mostly identical classification conditions for the DL models and participants to facilitate comparability. The round-robin test entailed successively rating micrographs with respect to their subtype and grain size. In the case of subtype misclassification (e.g., 'M' instead of 'B'), the participants were presented with the consequential range of structure codes (M5–M10) rather than with the correct spectrum of structure codes (B5–B10) as answer options. Additionally, metallographers were able to provide information on which micrograph regions were involved in their grain size selection—microstructural extrema or an average. Lastly, participants were able to provide free text comments for each image. All functionalities were integrated into a web application for straightforward participation, rating retrieval, and evaluation.

One important aspect to take note of is that the grain size label ('ground truth') for each image is determined by a single person. Therefore, the grain size label is afflicted by subjectivity as well. However, in contrast to the round-robin test participants, the original raters during day-to-day operation are provided with comprehensive background information about the specimen (alloy, heat treatment, storage, and contrasting) and flexibility during assessment (microscopy settings, observation of different specimen regions, and occasionally consulting further colleagues). While the decision about the reference grain size label is better informed than a single round-robin test participant, the reference label arguably does not represent a credible ground truth as it is still founded on the visual distinction of classes which are partly defined by nuanced changes in their image texture (see Figures 1C–E) rather than quantitative measurements. In contrast, the subtype classification is deduced objectively from the heat treatment process parameters. However, the distinction between subtypes is not a task that the raters are commonly facing. Relying solely on images with micron bars rather than additional contextual processing information might incentivize the round-robin test participants to deviate from their daily classification habits and rely more on measuring needle length. Even the awareness of contributing to a round-robin test can result in a non-natural, to some degree disproportionate classification effort.

First, an overview of the round-robin test outcome for all micrograph instances is provided. Most errors were made during grain size assessment rather than subtype distinction. At first glance, this is surprising since the distinction between martensitic and bainitic microstructures is generally considered challenging

10.3389/fmats.2023.1222456

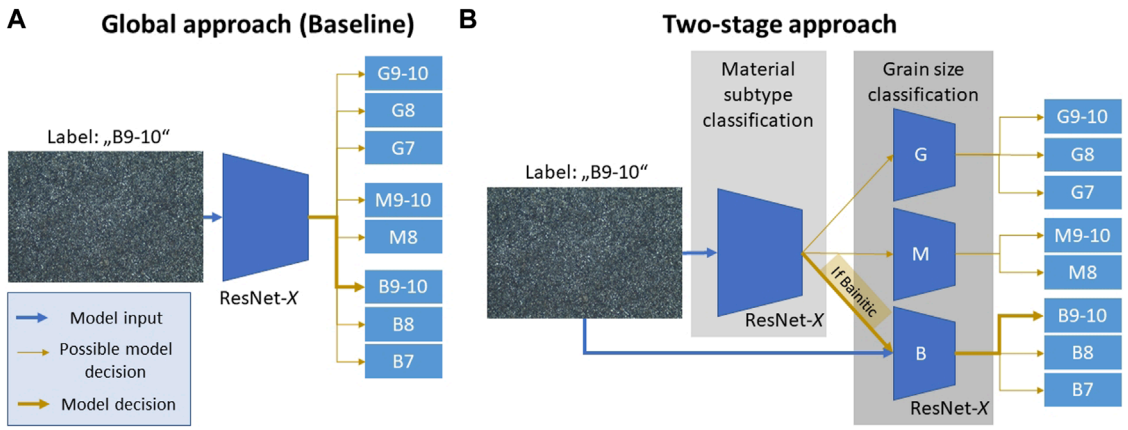

FIGURE 2
A comparison of a single multi-class classification setting **(A)** with the proposed two-stage approach **(B)** of first determining the material subtype and then the needle morphology.

TABLE 2 The overall model prediction accuracies achieved on the complete test set.

Model	Accuracy (%)
ResNet-18 global (fw)	76.08
ResNet-18 2S (fw)	80.06
ResNet-18 global	90.18
ResNet-18 2S	90.49
ResNet-50 global	87.73
ResNet-50 2S	88.98

The abbreviation 'fw' indicates models which were trained with frozen weights of the feature extractor portion. The specifications 'global' and '2S' refer to direct structure code prediction and the two-stage approach depicted in Figure 2, respectively.

and is not a task faced by the participating metallographers on a day-to-day basis. Specifically, out of 280 total predictions, the correct subtype was identified in 179 cases (63.9%). Out of this subset, the correct grain size was determined in 56 instances (31.3%). This corresponds to an overall accuracy of 20%. When aggregating the grain sizes 9–10 and dropping the ones with gray background in Table 1 (by discarding the corresponding instances with those predictions or labels leaving 187 predictions), the accuracy reaches 25.1%. Considering the difficulty of the provided micrographs, the two successive classifications one of which is not commonly tackled during day-to-day operation, the noisy reference labels, and the vaguely defined decision thresholds for the mean or maximum microstructure criterion to be applied, the overall accuracy of 20%–25% is not striking. It was, however, unexpected to us that the subtype classification achieved a relatively high accuracy with 64% under these circumstances, despite the DL models performing well at this task. This means that many metallographers picked up nuanced differences between these subtypes. Note that the 20 images were partitioned into 11, 5, and 4 for the G, B, and M subtypes, respectively. This skew might have simplified the subtype classification task due to the metallographer's awareness that the G subtype is dominating the data (see Supplementary Table S1) and their familiarity with the G material state. Indeed, the metallographers also performed better in the grain size predictions of the more common G and B subtypes and achieved accuracies of 0.21 and 0.23 as opposed to 0.16 for the M minority class. The final global ResNet-18 model evaluated on the round-robin test images, which are rather difficult cases, achieves an overall accuracy of 70.59% surpassing the 20%–25% achieved by human expert classification by a large margin. On a less difficult, more representative set, the classification accuracy of the raters would presumably increase.

Figure 5 shows the distribution of structure code predictions for each image. A majority vote in Figure 5 coincides with the reference label for six image instances. Apart from that, there are eight image instances where a majority vote is adjacent to the reference label. In such cases, it is debatable whether the reference label or the majority vote is more trustworthy. Leaving aside the reference labels, in order to assess the inter-rater reliability, the overall Fleiss' kappa score (Fleiss, 1971) was computed across all 14 participants. It amounts to 0.146, which is fairly low, even in view of the many classes, and thus a sign of elevated subjectivity and mediocre agreement between the participants. Only considering the material subtype distinction results in a higher Fleiss' kappa score of 0.330. When computing the overall Fleiss' kappa after reducing the possible classes, i.e., aggregating predictions and labels of grain size 9–10 and dropping image instances with grain sizes marked gray in Table 1, a value of 0.152 is obtained which is slightly higher than the 0.146. This is not surprising as the score depends on the cardinality of the categorical variables and the aggregation increases the accordance between the raters. Note that among the 14 participants, five deal with microstructures and grain size classification daily, while the others do less regularly (weekly to monthly). This affects the scatter of the round-robin test. When analyzing the overall Fleiss' kappa only for those five participants, the score increases from 0.152 to 0.318. This substantially higher accordance seems reasonable as these metallographers are not only more familiar with differences

10.3389/fmats.2023.1222456

TABLE 3 The prediction accuracies achieved on the test set for the different classification tasks.

Model	Subtype Accuracy (%)	Grain size G Accuracy (%)	Grain size B Accuracy (%)	Grain size M Accuracy (%)
ResNet-18 2S	96.35	93.90	86.36	92.86
ResNet-50 2S	96.96	93.09	81.82	92.86

The values reported here are obtained by the individual elements of the two-stage (2S) models, see Figure 2B.

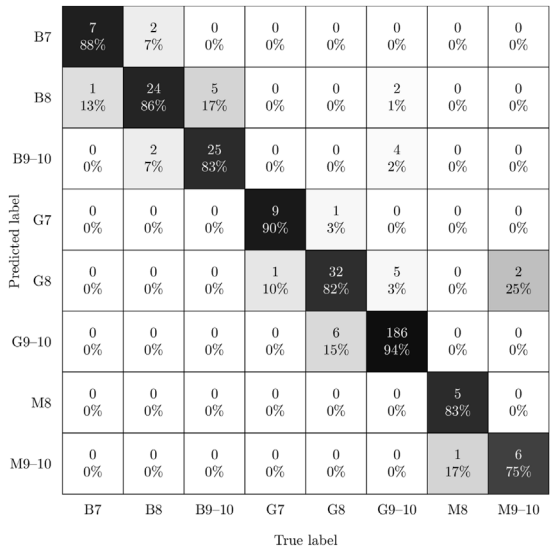

FIGURE 3
The confusion matrix of the global ResNet-18 model with an overall accuracy of 90.18% obtained on the test set.

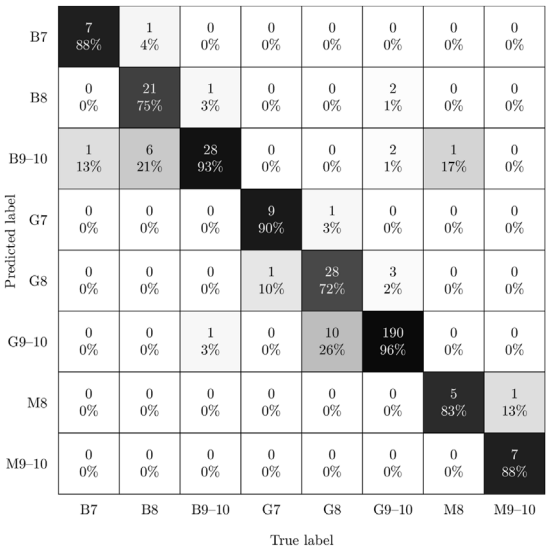

FIGURE 4
The confusion matrix of the two-stage ResNet-18 model with an overall accuracy of 90.49% obtained on the test set.

between the subtypes but also were more thoroughly trained to perform the grain size classification.

3.2 Model performances

As opposed to ratings by humans, the ResNet-18 and ResNet-50 models are deterministic and therefore repeatable as no stochastic elements such as dropout layers are used in their architecture. While at training-time stochastic online data augmentations are applied, at testing-time such operations were not utilized. The DL models supposedly learn an adequate mean representation and accurate decision boundaries from the noisy labels of many raters. An overall accuracy slightly exceeding 90% is very satisfactory. This applies especially in view of the large data variance and the noisy, subjectivity-afflicted reference labels which limit the attainable performance on the test set. Training the complete ResNet models rather than relying on ImageNet weight initialization in the feature extractor allows for a better model specialization towards the downstream tasks which differ significantly from ImageNet classification. Thus, there is a pronounced performance boost for the completely optimized models as opposed to the ones where only the classification head was optimized, see Table 2. The merit of the two-stage model was shown to be larger (4% improvement) in the frozen weight scenario. This can be ascribed to the fact that tuning weights of a single fully-connected classification head in the global approach is presumably inappropriate for the present set of diverse tasks and classes. On the contrary, the performance of the fully trained models is likely most of all limited by annotation noise, thus overshadowing the influence of the two-stage approach.

The two-stage approach is motivated by the presumption that the morphology classification requires an emphasis on different features than distinguishing between martensite and bainite subtypes. It is anticipated that the latter relies on the distribution of retained austenite and carbide phases, while needle morphology classification relies on length scales in the hierarchical substructure. Since the subtypes exhibit distinct microstructure length scales, decomposing the morphology classification further into three subtype-specialized grain size models was deemed promising. However, the results show that the two-stage model only marginally outperforms the global model and it is arguable whether it justifies the additional effort of training four distinct models. The two-stage approach facilitates picking the appropriate architectures for both tasks, depending on the representation power needed. In our case, a comparatively more expressive ResNet-50, and ResNet-18 seem to be beneficial for the subtype and grain size classification stages, respectively, as suggested by Table 3. When exploring how the decision-finding process in the two-stage process differs from the global model, it can be observed

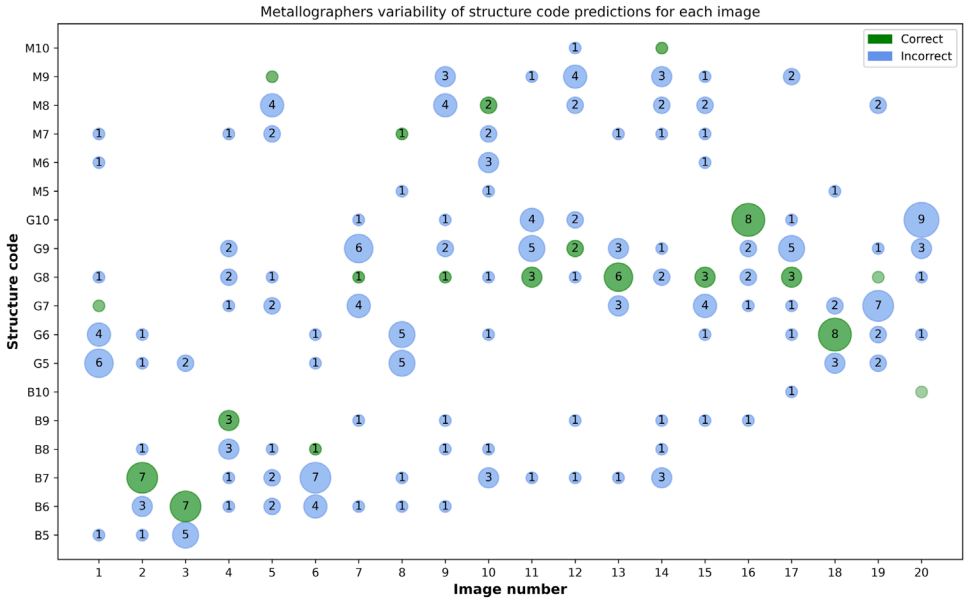

FIGURE 5
A plot in which the circles indicate correct (green) and incorrect (blue) structure code predictions of the round-robin test participants for each image. The size of the circle and the contained number indicate the number of predictions for that specific class. For images where no grading is consistent with the reference label, an empty green circle is plotted to indicate the reference label.

that the two-stage model, as expected, decomposes both decisions and relevant regions. Three Grad-CAM activation maps provided in Supplementary Figure S2 visualize this behavior. The same G class micrograph is passed to the global model and both stages of the two-stage model and Grad-CAM maps are extracted from the last layer of each ResNet-18 model. It seems that the global model considers regions of retained austenite, which are virtually exclusive to the G class, to perform the subtype distinction and simultaneously a few prominent needles to infer the grain size. In contrast, in the sequential approach, high activations occur solely at regions with high retained austenite concentration in the subtype stage and only at distinct needles at the grain size stage. It can be observed that the subtype model effectively takes more retained austenite regions into account to form its decision than the global model. This stronger activation might be caused by comparatively more specialized convolution filters and the increased feature density could potentially help to increase the model's confidence.

Whether and how tasks should be partitioned depends on the similarity of the tasks. For the challenge faced here, multi-task learning (MTL) is a very interesting and related paradigm. It utilizes hard or soft model parameter sharing together combined with distinct, task-optimized classification heads to address multiple tasks and optimize multiple objectives with a single model (Ruder, 2017). MTL can be especially beneficial if the tasks are related and rely on similar image features (Caruana, 1997). This is presumably the case for the grain size distinction tasks across the three different subtypes. In fact, in the G and B subtypes, the needle length thresholds defining the grain sizes are virtually identical and deviate slightly for the M subtype. This could incentivize using a single jointly optimized backbone with three classification heads rather than three distinct subtype-specific grain size models prospectively. Moreover, in the current approach, only the data subsets of the respective subtype are used to train each of the three grain size classification models. This represents a strong restriction in terms of data quantity, especially for the M grain size model. Training the largest portion of the architecture with joint datasets, e.g., in the MTL setting, might alleviate this problem. The data subsets for the three grain size classification tasks are expected to have distinct labeling noise patterns. In such cases, parameter sharing in the architecture's backbone can result in learning more general representations where the data subset-dependent labeling noise is ignored (Ruder, 2017). MTL might incentivize the model to learn the relevance of needle length due to the additional evidence provided by the supervisory signal of different subtypes, despite their distinct needle morphology. Based on the earlier discussion on the decision-finding process in the two-stage model, it is rather unlikely that the subtype and grain size classification tasks rely on similar enough features to justify extensive parameter sharing. Nowadays, self-optimizing MTL models have been developed which optimize their architecture to train specific layer's parameters jointly across different tasks whenever it is beneficial to the overall performance, see Misra et al. (2016); Ruder et al. (2017).

3.2.1 Material subtype classification

When examining the images, some distinctive features between the different subtypes and commonalities between instances of the same subtype can be identified. Specifically, the martensitic instances (M class) do not exhibit any retained austenite and carbide constituents (neither carbides nor retained austenite) and the overall needle structure appears comparatively coarse. In contrast, in the bainitic through-hardened (B class) instances, carbides always occur while virtually none have retained austenite. Moreover, the B class is

characterized by a fairly heterogeneous needle length distribution. In some cases, a bimodal needle length distribution with two distinct regions occurs which is exclusive to the B class. The martensitic through-hardened material (G class) is typically more homogeneous than B and has an overall finer needle morphology. It can contain both, carbides and retained austenite, depending on the grain size. In finer structured material (8–10) carbides occur, while in coarser martensite (6–7) retained austenite is typical, see Figure 1.

It is interesting to explore whether the subtype model picks up on the same discriminative image features to achieve the accuracy of 97% or whether it relies on other, maybe more subtle, aspects, such as slight differences in needle morphology induced by the varying quenching process. In order to investigate this, a few typical and atypical images of all subtypes are passed to the ResNet-18 subtype model, and the resulting Grad-CAM maps are illustrated in Figure 6.

In Figures 6A, B a typical bainite micrograph and its Grad-CAM heat map are shown. From the Grad-CAM map, it is evident that the subtype model concentrates on the coarser needle structures embedded within finer regions and on the clustered carbide regions. Both are typical characteristics of the B class. As opposed to this, in the M class micrograph and Grad-CAM heat map depicted in Figures 6C, D, the attention of the model is spread more uniformly across the image. This is plausible as the C56E2 martensitic material seemingly exhibits a single-phase and relatively uniform microstructure with generally coarser grains. When assessing a through-hardened martensite (G class) micrograph in Figures 6E, F, it can be seen that the model focuses on retained austenite regions that are exclusive to the G class. An image pair in Supplementary Figures S3A, S3B shows a misclassification where the B class micrograph was confused for a martensitic through-hardened (G class) grade. This misclassification is probably owed to the fact that the microstructure is exceptionally homogeneous for a bainitic micrograph with no visible variation in needle size. Therefore, it indeed seems the predictions are only incorrect when

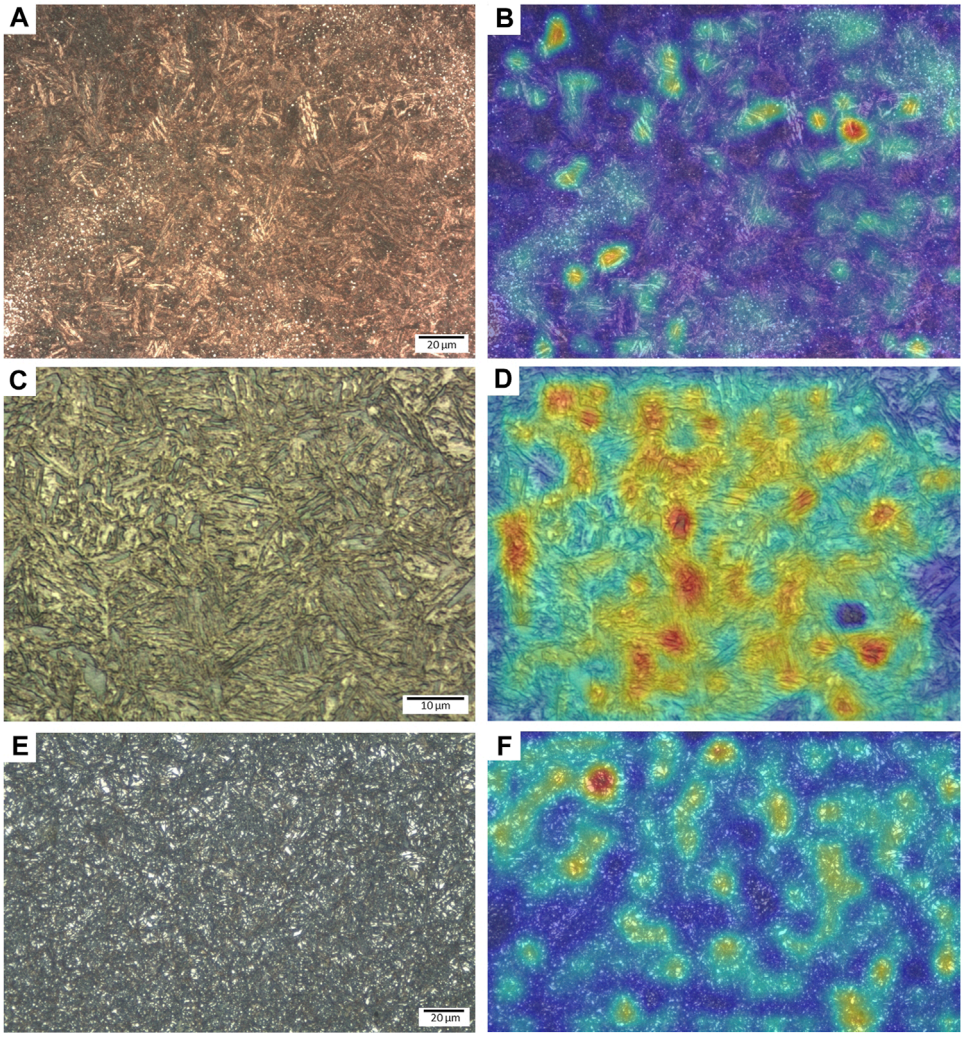

FIGURE 6
Steel micrographs showing different hierarchical microstructures **(A,C,E)** are input into the subtype classification model to extract the micrograph's Grad-CAM maps at the model's last layer **(B,D,F)**. The Grad-CAM maps use 'jet' colormaps. Thus, red, orange, and yellow colors highlight regions with increased activation.

the micrographs deviate from the trends introduced at the beginning of this section. While the features associated with those trends are discriminative for the range of heat treatments and alloys presented here, there are naturally other bainitic and martensitic grades that do not comply with the trends outlined in the first paragraph. This raises questions regarding model generalization to arbitrary bainitic/martensitic steels, e.g., in which retained austenite/martensite-austenite islands or pearlite constituents occur in a bainitic matrix.

3.2.2 Needle morphology classification

The largest portion of the remaining errors can be ascribed to the grain size classification and annotation noise therein. Especially, the distinction between the finest grain sizes, G8/G9–10 and B8/B9–10, seems to cause misclassifications. In the bainitic materials, these misclassifications can be primarily attributed to heterogeneous microstructures. Figure 7 shows two bainite micrographs that have been passed to the bainite grain size classification model along with the resulting Grad-CAM maps. Notably, the model takes the coarser regions into account to perform a correct prediction, see Figure 7A. This is a typical model behavior as the model learns to apply the maximum criterion (see Table 1). In a few cases, however, such as the one illustrated in Figures 7C, D, when the coarser regions occupy a small portion of the micrograph, the model tends to consider the finer regions to perform the prediction. While utilizing some kind of area threshold is the desired behavior, and it is promising that a lower area proportion of coarse regions leads to the model tending to the finer grain size prediction, the reference label was selected based on the coarse regions in this case. Instead of the eventually applied padding pre-processing approach, initially, we attempted to tile the raw images to fixed tile sizes and aggregate the tile's predictions. However, due to the heterogeneity of some micrographs in terms of needle length (see Figure 7), the tiling introduced further labeling noise and thus reduced the performance.

3.3 Implications for quality control of microstructures

While there are significant fluctuations in metallographic cross-section preparation and some variations in the alloying, there is also a notable dataset bias since the data is drawn from repeatable production processes. Specifically, the heat treatments culminate only in a small portion of possible martensitic and bainitic microstructural states. Indeed, this represents a simplification as the few distinct heat treatments result in fairly obvious discriminative features in the microstructure, see Section 3.2.1. Through interpretability analysis, models were shown to rely on these straightforward distinctive regions rather than learning nuanced correlations in the needle morphology. Thus, further data with more diverse martensite and bainite instances might be necessary to achieve better model generalizability by incentivizing the model to depend on a set of discriminative features for subtype

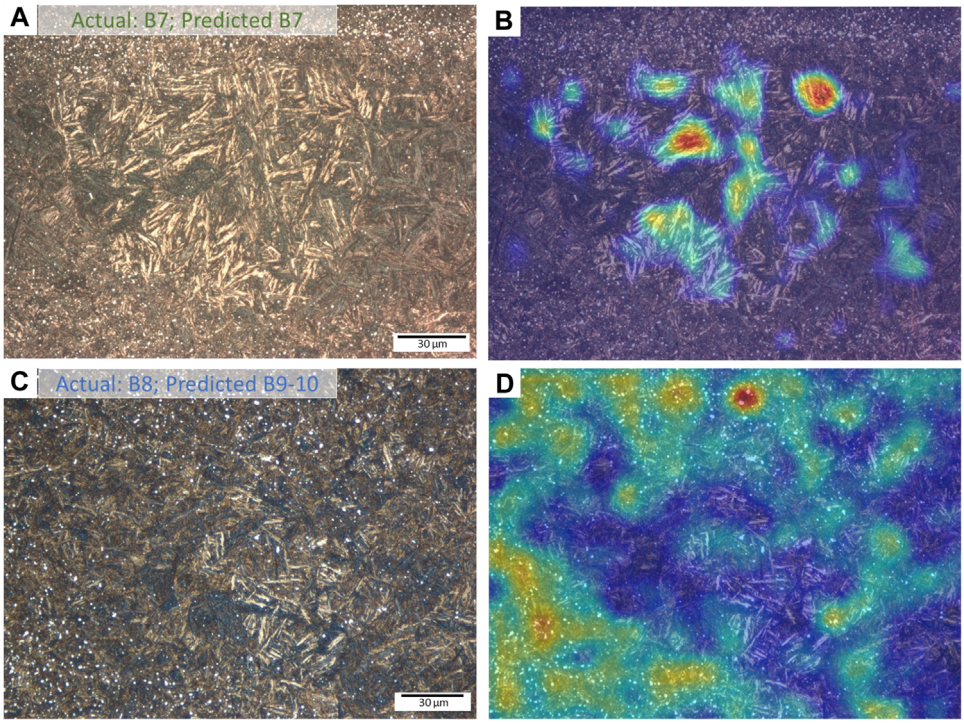

FIGURE 7
Bainitic steel micrographs showing heterogeneous microstructure **(A,C)** are input into the bainite grain size model to extract the micrograph's Grad-CAM maps at the model's last layer **(B,D)**. The Grad-CAM maps use 'jet' colormaps. Thus, red, orange, and yellow colors highlight regions with increased activation.

distinction. When the dataset becomes vaster in terms of heat treatments and alloying, the distribution of retained austenite and carbide phases might not be a sufficient discriminative feature anymore. In such a case, adding a needle length objective as an additional task in an MTL setting could promote learning more nuanced differences in needle morphology between bainitic and martensitic states (Ruder, 2017).

Besides a high model accuracy, we consider three additional points particularly important when it comes to deploying such a model in a productive environment.

1. Images that look similar to hardened bainitic or martensitic microstructures such as equivalent micrographs with some contained detrimental phases or a fully pearlitic micrograph should either be rejected from being classified by the model or a warning should be issued, that the image probably is not within the model's training distribution. This could help to prevent false classifications and raise confidence in the prediction's correctness as deep learning models otherwise were shown to exhibit mediocre generalization to out-of-distribution (OOD) samples (Torralba and Efros, 2011). In literature, depending on the degree of deviation and the available labels, the task of identifying abnormal images is distinguished into near/far OOD, anomaly, or novelty detection, see Ruff et al. (2021); Mirzaei et al. (2022). Typically, these tasks operate under the condition that a large quantity of OOD samples is unavailable which rules out training a binary OOD classifier. This condition is fulfilled in the present case as outliers are extremely rare in production. Subclass labels such as steel subtypes typically render the task of detecting abnormal images easier Vaze et al. (2021) easier and are available in the present case. Also, there is a necessity of detecting micrographs with subtle and local microstructural differences. For these reasons, the task could be framed as near-OOD detection. Data-driven models in literature address OOD detection by measuring distances between image embeddings or by using reconstruction-based approaches Lee et al. (2017); Liang et al. (2017); Ruff et al. (2021). While far-OOD detection has been successfully tackled for a range of domains, near-OOD detection still represents a major challenge, especially when nuanced local changes in unstructured images are concerned (such as the emergence of inclusions or some pearlite in microstructures). Prospectively, incorporating such a near OOD detection could improve the system's robustness against production outliers, arbitrary microstructures, or metallographic artifacts. Images classified as OOD can then be inspected manually by metallographers.
2. In order to continuously improve the model or to review individual model ratings by metallographers and data scientists, software solutions should facilitate providing process data. Dubious or interesting micrograph cases can then be collected to improve the model in the future. Aside from this, micrographs classified as in-distribution can be used for on-the-fly optimization within a semi-supervised learning framework. The process data should comprise saliency maps or other interpretability techniques. Supplementing information on the decision-making process is essential to establish trust in the proposed data-driven methodologies, especially when quality control is concerned.
3. Nowadays, microscope systems are often connected to imaging software that assists users to adjust, acquire, and store images. Usually, additional functionality such as measuring or annotating image features is provided, which supports microstructural analysis and reporting. Integration of DL models into such software solutions will increase model

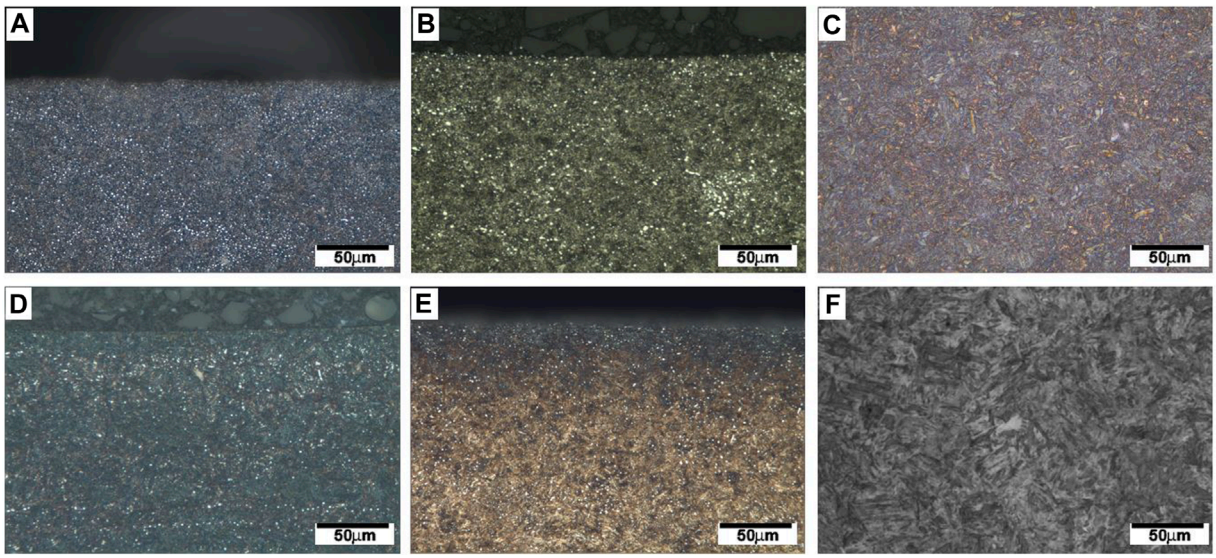

FIGURE 8
A set of images depicting some martensitic and bainitic microstructures which underwent different hardening treatments with varying primary microstructure and distinct distributions of carbide and retained austenite constituents. The images are underlying the fluctuations described in Section 5.1. Some images **(A,B,D,E)** cover regions outside the specimen at the top image border, while others **(C,F)** contain only region of interest.

acceptance. However, the deployment of trained models relies on microscopy software vendors providing appropriate software interfaces. This is a feature that only a few microscopy platforms offer yet will be indispensable going forward.

Treating the grain size estimation as a classification task might be beneficial to obtain a simple measure for repeated quality assessment throughout the supply chain. However, in this case, the categorization introduces a significant subjectivity. An alternative to the grain size assessment after ISO 643 could be to extract the needle length *distribution* and derive physical metrics with more relevance towards target properties, such as probability distributions for rolling contact fatigue resistance. The micrographs at the magnifications required to judge the needle morphology cover a too small area to be representative of the whole material, especially for the more heterogeneous and coarser microstructures. Thus the current procedure requires the metallographer to identify the most critical region on the cross-section before image acquisition and needle assessment. This in itself introduces a subjectivity that is not considered in the round-robin test presented here.

4 Conclusion

A deep learning model was presented that categorizes hierarchically structured, quenched steels first with respect to their microstructure type and then their needle length. The model achieves satisfactory accuracy on both tasks and learns a mean representation of the data which is labeled by many metallographers, effectively reducing the impact of the significant labeling noise. Despite the thorough training of the metallographers, the task of manually assigning an ISO 643 grain size was demonstrated to be subjective in a round-robin test. Here, an objective and deterministic deep learning model can provide a remedy, especially if quality control throughout supply chains is concerned. Together with the bainite/martensite subtype classification, this permits identifying defective processes and thorough quality control. The model's attention map is investigated by the presented interpretability analysis and consequences with respect to its generalization ability are discussed. For the deployment of the model, a robust out-of-distribution sample detection would be helpful since the model, at its current stage, is expected to not generalize across all possible martensitic and bainitic states. To achieve a more extensive generalization, data with fewer dataset biases, i.e., more diversity in terms of alloying and heat treatments than production data, should be supplemented.

5 Materials and methods

5.1 Dataset generation and statistics

Although the basic process of metallographic microstructure assessment, namely, preparing, etching, imaging, and evaluating, is standardized, it is not feasible to render the whole procedure entirely repeatable without automation—especially in a production environment. Achieving this would require, amongst others, a reproducible storage period after polishing, fresh etchant for every specimen, etching times aligned to the millisecond, and the same brightness/contrast settings in the microscopes. Thus, even though all images were taken on upright metallographic microscopes after etching with Nital (2%–3% alcoholic nitric acid solution) using the same imaging software, the image dataset features a significant variance.

Predominantly, these variations result from using the individual magnifications, illumination, white balance, and luminosity settings that the different microscope/camera systems (ten different microscopes in this case) exhibit, as well as different etching times and etchant qualities applied by the respective users. Another major source of data scatter is that the images were captured at two plants that process different alloys. Nonetheless, these image dataset variations are still well within the natural limitations of the metallographic microstructure analysis process.

Last but not least, despite extensive training, every metallographer has a unique way of judging the images. After all, visual perception is very subjective, see Anderson (2011); Panagiotaropoulos et al. (2014). Metallographers with decades of daily experience might not only comprehend the image textures differently than metallographers who rarely perform such tasks but also categorize them using a different approach. While very experienced metallographers will judge the needle coarseness alone by visual perception, inexperienced ones tend to apply a more quantitative approach in measuring needle lengths to correlate them with the grain size. Additionally, there are two possible rating systems applicable to judge the coarseness of the microstructure, as given in Table 1. The first one is the mean criterion, where an image is rated according to the overall visual perception of the microstructure. This approach is pursued mainly if the depicted microstructure is homogeneous and fine-grained. In the case of a more inhomogeneous needle length distribution, and if the coarse-grained portion occupies a significant amount of the micrograph, it is possible to rate an image according to the maximum criterion. In that case, the coarse-grained regions in an image dictate the grain size label. This is motivated by the fact that ensembles of larger grains determine fatigue properties. When observing Figure 1, it is apparent that the finer needle morphologies exhibit more similarity in their image texture due to their narrower decision boundaries.

In total, the dataset contains 1,641 micrographs. A randomly sampled set of as-received images which is supposed to point out the contained variance is depicted in Figure 8. A summary with respect to the class distribution of the dataset is provided in Supplementary Table S1. There it is evident that the dataset is skewed towards the martensitic through-hardened G class (75.6% of the overall data instances) and in particular to higher structure codes G9–10 (60.0%). This is owed to the fact that this material is the predominant outcome of production. Aside from this, the bainitic through-hardened material ranks second (20.1%). In contrast, the martensitic class M has a small share of the dataset (4.3%). All three subtype datasets were split into train, validation, and test subsets using proportions of 64:16:20 in a stratified manner, i.e., ensuring that each subset contains a virtually identical distribution of structure codes.

There are further sources of data variance that are not considered in the data splits or later during imbalance correction. For instance, there is a pronounced imbalance in terms of alloys as depicted in Supplementary Figure S4. Roughly 68.7% of the

data represents 100Cr6 alloy. The remaining data is composed of 100CrMnSi6-4 (15.3%) and otherwise distributed across twelve further 100Cr6 variants and the C56E2 alloy (M class). Some of the micro alloyed variants show up in the dataset with less than 10 instances and are hence clustered into a miscellaneous category in Supplementary Figure S4. Aside from this, the two magnifications with ×500 and ×1,000 are applied depending on the appropriate field of view and spatial resolution for the microstructure at hand. Overall, the ×500 images outweigh the higher magnification as they constitute approximately 82.4% of the data. The magnification of ×1,000 is most frequently employed for structure codes 9–10. Image resizing strategies to match pixel sizes across both magnifications were tested but ultimately disregarded as they turned out to be detrimental to the model performance. Apparently, the models manage to learn image distortions or noise patterns introduced through the distinct optical paths and are thus not adversely affected by distinct magnifications being present in the dataset. Also, the fluctuation in terms of specimen preparation and contrasting conditions (etching) was not traced and are not given special attention during data preparation. While the utilized microscope setup for each micrograph is documented, no means of data adjustments, e.g., correction of optical distortion, were employed.

5.2 Data pre-processing

In Figure 8, some images were shown to feature regions outside the region of interest. Namely, the images typically contain micron bar annotations and often extend over the specimen borders. Therefore, either some defocused background or metallographic embedding resin regions are included which along with the micron bar can lead to spurious correlations. Since this can lead to the models learning non-causal relations, regions outside the region of interest were cropped in advance. In cases of an incoherent resin-sample-interface, so-called bleeding occurred occasionally where etchants or solvents creep out of the slit at the interface resulting in a visually altered surface region of the sample. This can be observed for instance in Supplementary Figure S1 within the green box annotation. Such regions were also removed in advance.

These modifications led to varying image resolutions. In order to cope with this, different strategies were tested in a preliminary ablation study. These strategies included tiling the images to a fixed size and aggregating the tile's predictions 1), resizing all images to the mean resolution of the dataset 2), and mirror padding/cropping to the most frequent resolution 3). As the latter turned out to perform best empirically, all images were transformed to 1994× 1994 resolution through mirror padding or minimal cropping. Similarly, to correct for the subtype and structure code class imbalances, see Supplementary Table S1, different imbalance correction methods were tested. These entailed instance-weighted cross-entropy and oversampling of the minority classes to balance out the data provided to the model. As in this case, oversampling the minority class performed slightly favorably, we carried on with it.

Online data augmentations were applied during training before passing the color images to the models. The augmentation pipeline is composed of random rotations by arbitrary angles, random horizontal/vertical flips, slight random contrast adjustments, and Gaussian blurring. Subsequently, the images were normalized to the ImageNet mean and standard deviation. For the augmentations, the Albumentations package was used (Buslaev et al., 2020).

5.3 Model training and interpretability analysis

All models were trained using ResNet-18 or ResNet-50 architectures. The models were initialized with ImageNet weights from torchvision (Marcel and Rodriguez, 2010) and then either the full model was optimized with the same learning rate or the feature extractor portion of the classification network was frozen. In terms of learning rates, initially, 1E-4 or 1E-5 was used depending on the exact model. These learning rates were then modified by a StepLR scheduler. The scheduler adjusted the learning rate every 30 epochs. Each learning rate adjustment was achieved by multiplying the current learning rate by a decay factor $\gamma = 0.1$.

Cross-entropy was selected as the objective function. All submodels of the two-stage model were optimized individually and not trained end-to-end. When it comes to the two-stage approach, training the subtype model relied on the entire dataset, and training the individual structure code classifiers relied on the relevant data subsets. The models were trained for varying numbers of epochs until convergence, yet no overfitting was observed in the training and validation loss curves. The model performing best on the validation set was then used for the results reported here.

In order to explore the tendencies of a model and understand the reasons for specific failure cases, model interpretability techniques can be helpful. In this work, we utilized a technique called Grad-CAM (Selvaraju et al., 2016). We applied it to explore the activation of the final convolution layer of the employed ResNet architectures. The technique provides heat maps where regions of pronounced activation in a specific image are highlighted. These heat maps are constructed by a weighted combination of all feature maps of that layer. The weights for each feature map correspond to the backpropagated gradients on which a global average pooling operation over width and height dimensions is applied.

Data availability statement

The datasets presented in this article are not readily available because the dataset utilized in the round-robin test for this study can be found in "http://dx.doi.org/10.24406/fordatis/258." The data utilized in training and testing cannot be shared as it is subject to confidentiality obligations and part of an on going study. Requests to access the datasets should be directed to ali.riza.durmaz@iwm.fraunhofer.de.

Author contributions

Conceptualization, funding acquisition, supervision—AD, JM, and RN; data curation—SP, DR, and RN; formal analysis, investigation, methodology—AD and SP; project administration, resources—AD and RN; software, validation—SP;

visualization—AD and SP; writing (original draft)—AD, SP, and RN; writing–review and editing—AD, JM, RN, and SP. All authors contributed to the article and approved the submitted version.

Funding

The work was tackled in a joint industry project between Fraunhofer IWM and Schaeffler Technologies AG and Co. KG for which the latter party provided the funding.

Acknowledgments

We want to express our gratitude to all metallographers of Schaeffler Technologies AG and Co. KG who created the dataset and to all who participated in the round-robin test.

Conflict of interest

Authors DR, JM, and RN were employed by Schaeffler Technologies & Co. KG.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmats.2023.1222456/full#supplementary-material

References

Anderson, B. L. (2011). Visual perception of materials and surfaces. *Curr. Biol.* 21, R978–R983. doi:10.1016/j.cub.2011.11.022

Bepari, M. (2017). Surface and heat treatment processes. *Compr. Mater. Finish.* 8. doi:10.1016/B978-0-12-803581-8.09187-6

Bulgarevich, D. S., Tsukamoto, S., Kasuya, T., Demura, M., and Watanabe, M. (2019). Automatic steel labeling on certain microstructural constituents with image processing and machine learning tools. *Sci. Technol. Adv. Mater.* 20, 532–542. doi:10.1080/14686996.2019.1610668

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., and Kalinin, A. A. (2020). Albumentations: fast and flexible image augmentations. *Information* 11, 125. doi:10.3390/info11020125

Caruana, R. (1997). Multitask learning. *Mach. Learn.* 28, 41–75. doi:10.1023/a:1007379606734

DeCost, B. L., Lei, B., Francis, T., and Holm, E. A. (2019). High throughput quantitative metallography for complex microstructures using deep learning: a case study in ultrahigh carbon steel. *Microsc. Microanal.* 25, 21–29. doi:10.1017/s1431927618015635

Durmaz, A. R., Müller, M., Lei, B., Thomas, A., Britz, D., Holm, E. A., et al. (2021). A deep learning approach for complex microstructure inference. *Nat. Commun.* 12, 6272. doi:10.1038/s41467-021-26565-5

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76, 378–382. doi:10.1037/h0031619

Gola, J., Britz, D., Staudt, T., Winter, M., Schneider, A. S., Ludovici, M., et al. (2018). Advanced microstructure classification by data mining methods. *Comput. Mater. Sci.* 148, 324–335. doi:10.1016/j.commatsci.2018.03.004

Hillert, M. (1995). The nature of bainite. *ISIJ Int.* 35, 1134–1140. doi:10.2355/isijinternational.35.1134

Lee, K., Lee, H., Lee, K., and Shin, J. (2017). *Training confidence-calibrated classifiers for detecting out-of-distribution samples*. arXiv preprint arXiv:1711.09325.

Liang, S., Li, Y., and Srikant, R. (2017). *Enhancing the reliability of out-of-distribution image detection in neural networks*. arXiv preprint arXiv:1706.02690.

Marcel, S., and Rodriguez, Y. (2010). "Torchvision the machine-vision package of torch," in *Proceedings of the 18th ACM international conference on Multimedia*, 1485–1488.

Mirzaei, H., Salehi, M., Shahabi, S., Gavves, E., Snoek, C. G., Sabokrou, M., et al. (2022). "Fake it until you make it: towards accurate near-distribution novelty detection," in *The eleventh international conference on learning representations*.

Misra, I., Shrivastava, A., Gupta, A., and Hebert, M. (2016). "Cross-stitch networks for multi-task learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3994–4003.

Müller, M., Britz, D., Ulrich, L., Staudt, T., and Mücklich, F. (2020). Classification of bainitic structures using textural parameters and machine learning techniques. *Metals* 10, 630. doi:10.3390/met10050630

Panagiotaropoulos, T. I., Kapoor, V., and Logothetis, N. K. (2014). Subjective visual perception: from local processing to emergent phenomena of brain activity. *Philosophical Trans. R. Soc. B Biol. Sci.* 369, 20130534. doi:10.1098/rstb.2013.0534

Ruder, S. (2017). *An overview of multi-task learning in deep neural networks*. arXiv preprint arXiv:1706.05098.

Ruder, S., Bingel, J., Augenstein, I., and Søgaard, A. (2017). *Sluice networks: Learning what to share between loosely related tasks*. arXiv preprint arXiv:1705.08142 2.

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., et al. (2021). A unifying review of deep and shallow anomaly detection. *Proc. IEEE* 109, 756–795. doi:10.1109/jproc.2021.3052449

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D. (2016). *Grad-cam: Why did you say that?* arXiv preprint arXiv:1611.07450.

Shur, E. A., Bychkova, N. Y., and Trushevsky, S. (2005). Physical metallurgy aspects of rolling contact fatigue of rail steels. *Wear* 258, 1165–1171. doi:10.1016/j.wear.2004.03.027

Torralba, A., and Efros, A. A. (2011). "Unbiased look at dataset bias," in *Cvpr 2011* (IEEE), 1521–1528.

Vaze, S., Han, K., Vedaldi, A., and Zisserman, A. (2021). *Open-set recognition: A good closed-set classifier is all you need?* arXiv preprint arXiv:2110.06207.

Zhu, B., Chen, Z., Hu, F., Dai, X., Wang, L., and Zhang, Y. (2022). Feature extraction and microstructural classification of hot stamping ultra-high strength steel by machine learning. *JOM* 74, 3466–3477. doi:10.1007/s11837-022-05265-5