

ARTICLE OPEN



A study of real-world micrograph data quality and machine learning model robustness

Xiaoting Zhong¹, Brian Gallagher², Keenan Eves³, Emily Robertson¹, T. Nathan Mundhenk⁴ and T. Yong-Jin Han¹✉

Machine-learning (ML) techniques hold the potential of enabling efficient quantitative micrograph analysis, but the robustness of ML models with respect to real-world micrograph quality variations has not been carefully evaluated. We collected thousands of scanning electron microscopy (SEM) micrographs for molecular solid materials, in which image pixel intensities vary due to both the microstructure content and microscope instrument conditions. We then built ML models to predict the ultimate compressive strength (UCS) of consolidated molecular solids, by encoding micrographs with different image feature descriptors and training a random forest regressor, and by training an end-to-end deep-learning (DL) model. Results show that instrument-induced pixel intensity signals can affect ML model predictions in a consistently negative way. As a remedy, we explored intensity normalization techniques. It is seen that intensity normalization helps to improve micrograph data quality and ML model robustness, but microscope-induced intensity variations can be difficult to eliminate.

npj Computational Materials (2021)7:161; <https://doi.org/10.1038/s41524-021-00616-3>

INTRODUCTION

Micrographs constitute an important class of scientific data and play a key role in the interpretation of material process–structure–property (PSP) linkage by revealing material microstructures. Microstructures are often diverse and sophisticated. Although automated quantitative analysis has been achieved for some micrograph classes¹, the analysis of complicated micrographs often remains qualitative and relies on the knowledge and experience of individual human experts until recent years. The recent game changer is the adoption of machine-learning (ML) techniques^{2,3}. Researcher have shown that micrograph-based ML models can achieve highly accurate material classification^{4–6}, defect detection^{7,8}, property prediction^{9–11}, and material quality monitoring¹².

High-quality large datasets are essential to the success of ML models. It has been shown that ML models can achieve high performance without learning truly important and generalizable data characteristics if there exists a bias in the training data^{13–15}. However, precise control of micrograph data quality is very difficult, especially in the case of large datasets. The content of a micrograph depends on multiple factors, including the microstructure being captured, the manufacturer of the microscope¹⁶, the user-controlled settings (such as magnification, brightness, contrast), and instantaneous microscope conditions (such as filament aging)¹⁷. We differentiate two classes of image signals: (1) microstructure-induced signals, which come from the microstructure content of the material sample, and (2) microscope-induced signals, which come from instrument conditions and user-controlled settings. Signal variations often present as pixel intensity variations in grayscale micrographs like scanning electron microscope (SEM) images. Some examples are given in Fig. 1, in which microstructure contents are controlled constant within each column, and images within different rows of the same column look different due to different microscope conditions. Microstructure-induced signals are generally reproducible as long as the sample is well preserved, while microscope-induced signals

can be much more difficult to reproduce. Microscope-induced signals are usually inconsistent and nonrandom if the collection was conducted by different people using different microscopes¹⁶. Even if all micrographs are collected by the same person on the same microscope using high-throughput auto collection techniques, the microscope-induced signals can still vary if the experiment takes a prolonged time and instrument drift happens during the collection¹⁸. It is important to note that we define microscope-induced signals to describe signal variations that are not correlated with material property and performance. Strictly speaking, material chemistry (e.g., atomic number) also contributes to micrograph pixel intensities, but holistically they are intrinsic to the material sample and are reproducible.

Microscope-induced intensity variations usually do not prevent human experts from making a correct qualitative interpretation of the microstructure content, but may affect ML model predictions^{6,19,20}. To be clear, an ML model in this context means either a traditional ML model, in which the model is a combination of a featurization step and a prediction (regression) step, or an end-to-end deep-learning (DL) model, in which the featurization step and the prediction step are not explicitly separable. Featurization refers to the process of encoding raw images into compact and informative descriptors that can be processed by the succeeding ML predictor. Image feature descriptors play a key role in ML performances but are not always robust to image quality issues like luminance, scale, translation, and occlusion^{21–24}. For example, Tsutsui et al. reported that the source of SEM, including field emission (FE) and tungsten (W), affects gray-level co-occurrence matrix (GLCM) based texture features, and a classifier trained with SEM images of one source does not classify SEM images of another source accurately²⁵. In the field of medical images, Strzelecki and Materka et al.^{26,27} also reported inhomogeneous brightness and contrast of magnetic resonance images (MRI) affect the calculation of statistical texture descriptors.

¹Materials Science Division, Lawrence Livermore National Laboratory, Livermore, CA, USA. ²Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA, USA. ³Defense Technologies Engineering Division, Lawrence Livermore National Laboratory, Livermore, CA, USA. ⁴Computational Engineering Division, Lawrence Livermore National Laboratory, Livermore, CA, USA. ✉email: han5@llnl.gov

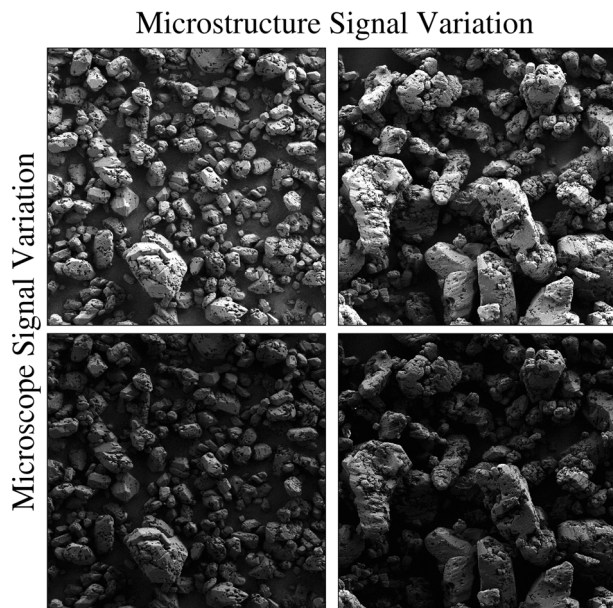


Fig. 1 SEM images showing signal variations. Images show variations from microstructure signals (column), derived from differences in microstructures of the samples and microscope signals (rows), derived from instrument conditions and setting.

Intensity normalization, or more generally image standardization, is one way to reduce microscope-induced signal variations and improve micrograph quality. However, there seems to be no consensus on the intensity normalization step in the micrograph ML community. Image binarization is a simple standardization step that works well for some relatively simple microstructures^{10,12}, but cannot be applied when greyscale or color information is desirable. Many studies involve non-binary micrographs but do not report an intensity normalization step^{4,7–9,28}. Bulgarevich et al. noted that a random forest pixel classifier of optical microscopy (OM) images is sensitive to image contrast and a brightness contrast adjustment step should be performed but provided no details on the standardization procedure and its effect⁶. Local histogram equalization is reported in a feedstock powder classification task¹⁹. Adaptive histogram equalization is reported in a microstructure cluster analysis²⁰ and a steel image classification task²⁵.

A thorough understanding of if and how microscope-induced signals affect ML model performances is a necessary step toward real-world applications of ML-powered micrograph analyses but is not yet available. One probable reason is that many tasks in current micrograph ML studies (like classification and clustering) involve diverse microstructures, in which both the microstructure-induced and microscope-induced signals vary simultaneously, and the latter may be shadowed by the former. Most current micrograph datasets are also small, especially in terms of the number of micrographs per microstructure class, so there is often not enough data for a study of microscope-induced signals specifically^{4–9,19,28,29}. Tsutsui et al.²⁵ recently demonstrated that different SEM sources can confuse an ML classifier, but they investigated only one kind of image feature descriptor and one intensity normalization technique. Also, they did not compare the effect of microscope signals in the context of well-controlled microstructure signals. To the best of our knowledge, different ML models and intensity normalization techniques have not been evaluated on the same micrograph dataset with well-controlled signal variations.

The objective of this research is to understand how microscope-induced signal variations affect ML model performances.

We examine thousands of SEM images, a popular class of micrographs, and present a well-controlled study of SEM image pixel intensity variation and its effect on five different ML models, which cover four popular image feature descriptors and an end-to-end advanced DL model. The image feature descriptors are chosen from different feature classes (including a binarized texture filter, a local key-point-based feature, and two CNN transfer learning features) as different image features can have different robustness to the microscope-induced signal variation problem. Several intensity normalization algorithms are also tested for their effectiveness in removing microscope-induced pixel intensity signal variations.

RESULTS

We first introduce an SEM image dataset and the machine-learning task. This first SEM image dataset was collected via high-throughput automated collection techniques and will be referred to as the original dataset. Then we discuss the quantification of micrograph pixel intensity, the variation of pixel intensities within the original dataset, the ML model hyperparameter choices, and the effect of intensity normalization. Finally, to better evaluate the effect of microscope-induced signals and ML model robustness, we collected a new SEM image dataset with carefully controlled microstructure content and varying microscope settings. Our conclusions are validated with this new dataset.

Original SEM image datasets and the ML task

The details of our SEM image data, and some early ML prediction efforts, have been published³⁰ and are briefly summarized here. The dataset contains 59,690 SEM images, covering 30 sample lots of TATB (2,4,6-triamino-1,3,5-trinitrobenzene) crystals with various microstructures. Each sample lot contains 732–2980 images. The SEM images were collected automatically on a Zeiss Sigma HD VP microscope with Atlas software. Microscope settings were kept constant during the collection of all images. The crystals were consolidated after SEM image collection and microstructure assessment. Uniaxial ultimate compressive strength (UCS) was measured from the consolidated sample for each lot. The inputs of our machine-learning models are the SEM images. The prediction targets are the lot-specific UCSs. To establish some state-of-the-practice baselines for the ML models, several empirically important microstructure statistics (including particle size, porosity, dispersity, facetness, and surface area) were determined for each lot from subjective estimations of two human experts. For more information about the baseline statistics and their performances, please refer to Gallagher et al.³⁰.

Because some image feature descriptors are computationally expensive (see “Methods”), we did not always use the 59,690-image full dataset. The four traditional machine-learning models were evaluated on a 6000-image reduced dataset, which contains 200 randomly sampled images for each of the 30 lots. The DL model was evaluated on the full dataset as the 6000-image reduced dataset is not enough for this large-capacity model to be well trained.

Micrograph intensity characterization

Let the intensity of all pixels in an image be an array $\{I_i\}$, $i = 1, \dots, 1,048,576$, where 1,048,576 is the number of pixels in the image. $0 \leq I_i \leq 1$, 0 is equal to black and 1 is equal to white. Four image pixel intensity metrics, including mean (\bar{I}), median (\tilde{I}), mode, and standard deviation (s_I), of $\{I_i\}$ were calculated to characterize the overall image pixel intensity level from perspectives of brightness and contrast. \bar{I} , \tilde{I} , and mode are designed to capture the overall image brightness level. s_I is intended to capture the image contrast. Four example images

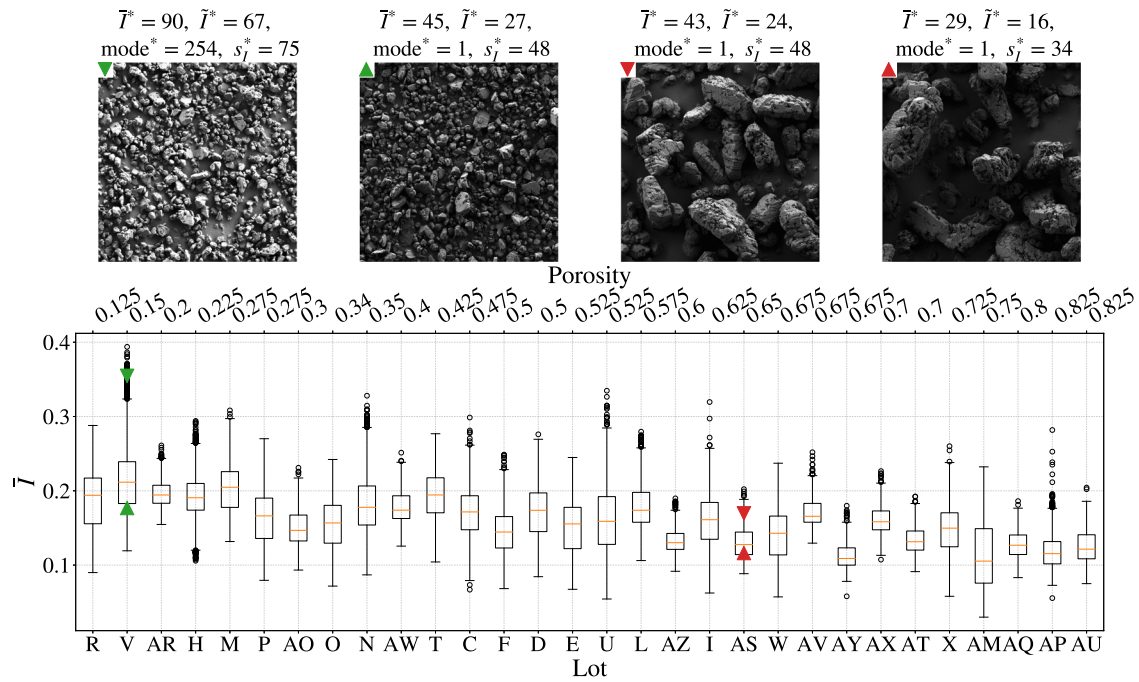


Fig. 2 Example SEM images and box plot of image pixel intensity distributions within each lot. Image intensity measurements are shown above each example image. Asterisk superscript denotes a scaling up factor of 255. For example, $\bar{I}^* = \bar{I} \times 255$, rounded to the nearest integer. \bar{I} of each example image is indicated in the box plot using the corresponding image label (green and red triangles). Box plot central lines indicate medians. The boxes extend from Q1 to Q3 quartiles of the data. Whiskers extend from box edges to the farthest data point within 1.5 interquartile range. Further data points are considered as outliers and are plotted separately.

are drawn from two lots and shown in the upper row of Fig. 2. The scaled brightness metric values (noted with asterisks and range from 0 to 255) are indicated above the images. It can be noticed that both \bar{I} and \bar{I}^* capture the overall image brightness reasonably well while mode does not. s_I is also a proper approximation of the image contrast. Variations of \bar{I} , \bar{I}^* , and s_I are generally similar, and we will only report \bar{I} as the image pixel intensity metric in the remaining of this document.

The distributions of \bar{I} in the 30 lots are shown in the second row of Fig. 2. It can be seen that \bar{I} is never a constant within one lot. One major source of this image intensity variation is the microstructure content. For example, an inclined facet is often brighter than a horizontal facet because of a larger interaction volume³¹. Pores are darker than facets because an electron that falls into a pore has a smaller chance to reach the detector. Note, sample lots in Fig. 2 are ordered according to their porosity level, which refers to the relative volume fraction of voids in feedstock particles³⁰. It can be seen that lots with larger porosity are generally darker than lots with smaller porosity. To quantify this kind of relationship, we calculated the Pearson correlations between lot microstructure characteristics and lot average pixel intensity measurements $\langle \bar{I} \rangle$, $\langle \bar{I}^* \rangle$ and $\langle s_I \rangle$, in which $\langle x \rangle = \frac{\sum_{j=1}^n x_j}{n}$ and n is the total number of images within the sample lot of interest. The (Pearson) correlations are shown in Fig. 3 and the p -values are given in the Supplementary Material. It can be seen that the correlations between $\langle \bar{I} \rangle$ and microstructure characteristics are generally significant, indicating that image intensity is correlated with the microstructure being captured. Microstructure contents are never exactly the same within two images in the original dataset. This is one reason that we always see distributions in Fig. 2.

Another source of pixel intensity variation is the microscope conditions, which were carefully controlled, by ensuring that all images were collected on the same microscope with the same experiment settings, but not eliminated. We noticed that images

collected near the end of one collection batch were often brighter than those collected at the beginning of the same batch. This instrument drift is probably related to the heat accumulation in the microscope filament, which results in an enhanced electron flux near the end of one image collection batch. This instrument drift effect is seen in Fig. 2 as the first and second example images come from the same sample lot and have similar microstructure contents, but the overall image intensities are different. Nevertheless, note the distributions within different lots are generally highly overlapped, which suggests that the microscope setting control was overall successful during the data collection step.

ML models and robustness to intensity variations

We compare five ML models, chosen from different popular model classes, to allow reasonable generality in our conclusions. Four models are built following a traditional ML pipeline, by combining different featurization methods and a predictor (random forest regressor). We briefly outline the featurization methods and hyperparameter search results here. Detailed backgrounds and reasons for choosing these methods can be found in “Methods”. The four featurization methods include a binarized texture featurization (BSIF)³², a local key-point-based featurization (final choice is KAZE₂₀₀₀-VLAD), and two CNN transfer learning-based featurization methods (final choices are C_{3,3}-VLAD and FC₂). Hyperparameters tuned in the local key-point-based featurization include the local key-point detector and descriptor extractor (KAZE³³, ORB³⁴, STAR detector³⁵ + BRIEF descriptor³⁶), the number of key-points on each image (500, 1000, and 2000), the clustering algorithm (KMeans³⁷ and GMM³⁸), the number of clusters (32 and 64), and the final feature generation algorithm (BoW³⁹ and VLAD⁴⁰). The CNN transfer learning-based featurization methods were extracted from an ImageNet⁴¹ pre-trained VGG16 network⁴² and the hyperparameters include which layer to extract, the clustering algorithm, and the number of clusters in the VLAD⁴⁰ step.

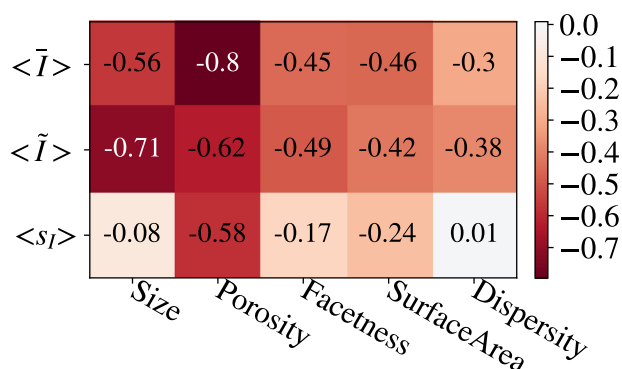


Fig. 3 Microstructure feature correlations with image intensities. Figure shows correlations between lot-average image intensity measurements (rows) and microstructure characteristics (columns), using all images.

The predictor in all four traditional ML models was kept as the same random forest regressor⁴³ (RF) used by Gallagher et al.³⁰ without fine-tuning. These four traditional ML models will be referred to by their featurization choice (BSIF, KAZE₂₀₀₀-VLAD, C_{3,3}-VLAD, and FC₂) in the remaining of this document. Hyperparameters of the four models were evaluated on a randomly sampled 6000-image validation dataset, and model performances are reported as mean absolute percentage error (MAPE) in Table 1. The fifth model is an end-to-end neural network trained from scratch using the full original dataset. Six model architectures were tried and our final choice was the Wide Residual Network (WRN)⁴⁴. Details about model evaluation can be found in “Methods”.

To understand if the microscope-induced intensity signals can affect machine-learning model predictions, we first bin images within the same lot according to their intensities and analyze the corresponding UCS predictions. Specifically, images within the same lot are sorted according to \bar{I} from small (dark) to big (bright), and then cut into ten groups according to deciles of \bar{I} . Each decile group has approximately the same number of images. The average UCS prediction of the lot image decile group, $\langle \hat{y}_{\text{BSIF}} \rangle_{l,d} / y_l$, is calculated for every decile group of each lot. y_l is the ground truth UCS of the lot l . \hat{y} is the UCS prediction of one image. Angle bracket denotes average, and $\langle \hat{y}_{\text{BSIF}} \rangle_{l,d}$ is the average BSIF model prediction for images within the decile level d of lot l . Results for five example lots are shown in Fig. 4, in which the color scheme is determined within each row (lot) and high prediction corresponds to dark color. Performances of the full 30 lots can be found in the Supplementary Material. Other ML models also exhibit similar trends.

The darkest blocks often sit on the most left of Fig. 4, which suggests that the darkest images are often predicted to have the highest UCSs. If image UCS prediction is independent of the image intensity, the color blocks would have been distributed randomly. One possible source for this low-intensity-high-prediction correlation is instrument drift (Fig. 2, top row). Nevertheless, note the gap between UCS predictions of the darkest (\bar{I} decile level = 0) and the brightest (\bar{I} decile level = 9) images is usually small. The small gaps suggest that images within the same lot are generally similar, as intended during the data collection step.

Intensity normalization

Intensity normalization is a common approach to minimize signal variations and improve image quality. We evaluated four image pixel intensity normalization algorithms in this work, including histogram equalization, adaptive histogram equalization⁴⁵, gamma normalization, and Tan-Triggs normalization⁴⁶. Histogram equalization is one of the most common image enhancement

algorithms⁴⁷. Adaptive histogram equalization applies histogram equalization on different local patches of the image, and has been applied in some micrograph analysis tasks^{20,25}. Gamma normalization is an in-house algorithm inspired by the fact that distributions of image pixel intensities often seem to follow a gamma distribution (see Supplementary Material for details). Tan-Triggs normalization⁴⁶ is an algorithm that eliminates image illuminance variations and has demonstrated good performance in face recognition tasks. Effects of intensity normalization algorithms are illustrated in Fig. 5. Implementation details are given in “Methods”.

When we apply an intensity normalization step, we apply it on both the training and the test images to keep the training and test data as similar as possible. An inconsistent intensity normalization routine in training and test images generally leads to bad performance (high MAPE). After the intensity normalization step, new features were calculated from the normalized images and fed to new random forest regressors.

The (Pearson) correlations between $\langle \bar{I} \rangle$ (the average image pixel intensity score within each sample lot) and sample lot human labels are shown in Fig. 6 (1st to 5th column). The absolute critical correlation for a two-tail t test 0.05 significance level is 0.36. The full P values are given in the Supplementary Material. We see that on a lot-average level, the correlations between image pixel intensity ($\langle \bar{I} \rangle$) and microstructure characteristics, especially size and porosity, are generally significant both before and after intensity normalization. This indicates that the intensity normalization methods do a good job of preserving microstructure relevant signals in the image.

The correlation between \bar{I} (the pixel intensity score of an image) and \hat{y} (the predicted UCS) is evaluated using the 6000-image dataset and shown in the 6th to 11th column of Fig. 6. The 6000-image reduced dataset was used in the traditional ML models because some image features were large in memory and can take too long to compute (see “Methods”). As a reference for dataset size impact, BSIF, which is the most computationally efficient feature, was evaluated on both the reduced dataset and the full dataset. The machine-learning models are indicated in the column labels, and models evaluated on the reduced dataset are denoted with a prime symbol. The intensity normalization methods are indicated in the row labels. None stands for the original images without any intensity normalization.

It is interesting to note that $\text{corr}(\bar{I}, \hat{y})$, the correlation between image pixel intensity \bar{I} and image UCS prediction \hat{y} , changed sign from negative to positive after histogram equalization, adaptive histogram equalization, and gamma normalization. Also, the correlation magnitude is generally weak without intensity normalization (None) and becomes enhanced after intensity normalization, especially histogram equalization (Hist_Equal). Though the pairwise correlations do not imply causal relations, we found these observations interesting and give our interpretations in “Discussion”.

Model performances on original datasets

We evaluated the different ML models with the original and intensity normalized images. Results are shown in Fig. 7. It can be seen that intensity normalization does affect ML model performances. The exact effect of an intensity normalization method is different when combined with different ML models, while histogram equalization and gamma normalization generally perform the best. One may find the BSIF column of Fig. 7 interesting because all intensity normalizations seem to hurt model performance. We note that this observation is probably related to data size and learning efficiency. Normalized images may achieve a performance similar to, or even better than, the original images if given more data. See Supplementary Material for more experiments and discussions.

Table 1. Model performances (given in MAPE) for hyperparameter tuning or model search.

(a)		VLAD				BoW			
		KM32	KM64	GMM32	GMM64	KM32	KM64	GMM32	GMM64
ORB	500	0.229	0.235	0.229	0.233	0.211	0.215	0.223	0.220
	1000	0.221	0.228	0.215	0.226	0.206	0.210	0.210	0.204
	2000	0.212	0.221	0.207	0.216	0.204	0.206	0.197	0.206
BRIEF	500	0.250	0.263	0.250	0.257	0.241	0.240	0.246	0.238
	1000	0.243	0.257	0.237	0.250	0.234	0.236	0.234	0.236
	2000	0.226	0.243	0.208	\	0.212	0.208	0.205	\
KAZE	500	0.189	0.202	0.184	0.191	0.210	0.199	0.197	0.187
	1000	0.178	0.188	0.174	\	0.183	0.190	0.179	\
	2000	0.164	0.172	\	\	0.179	0.181	\	\
(b)						KM32		KM64	
C _{2,2} (8)						0.143		0.151	
C _{3,3} (15)						0.134		0.154	
C _{4,3} (22)						0.152		0.156	
C _{5,3} (29)						0.177		0.186	
(c)									
AlexNet	DenseNet169	ResNet50	WRN	SqueezeNet	VGG16				
0.151	0.187	0.136	0.115	0.271	0.240				

Performances of final choices are shown in bold. (a) Local key-point-based featurization. Key-point descriptors and the number of key points in each image are shown in row indices. Local descriptors and the respective clustering parameters (clustering algorithm and number of clusters) are shown in column indices. KM stands for KMeans. GMM stands for Gaussian Mixture Models. Blank entries correspond to experiments for which the clustering step did not converge within 24 h.

(b) CNN-based transfer learning features. Output convolution layers are shown in row indices. C_{ij} stands for the convolution layer *i* in the convolution block *j*. The corresponding layer ID in the PyTorch pre-trained VGG16 is shown in parenthesis. The clustering parameters of VLAD are shown in column indices.

(c) End-to-end DL models trained from scratch.

New SEM image dataset with varying microscope-induced signals

Both the microstructure content and the microscope condition can affect image pixel intensities. The best way to evaluate microscope-induced pixel intensity signals is probably to hold image microstructure constant and conduct experiments with various pixel intensity levels.

We collected new experiment images for six lots with various microscope brightness and contrast settings. AQ was the first collected lot, for which the same sample region was scanned 12 times with a constant microscope setting to check if the electron beam can cause unexpected damage to the sample during repeated scans. After confirming that electron beam damage is negligible to our sample, we collected one scan for lot D, and multiple scans with different microscope settings for lot AO, AT, AX, and AZ. Detailed microscope settings are given in the Supplementary Material. Images in the same-lot scans have the same microstructure content but different pixel intensities due to the microscope settings. Some image examples are shown in the top row of Fig. 8. It can be seen that the microstructure content within each scan is well-controlled, and the region of interest (ROI) is off by a few pixels at most.

Model robustness on the new dataset

Image intensity (\bar{I}) distributions of the raw new scans are given in Fig. 8a, as well as the corresponding old lot distributions. The intensity distributions after histogram equalization are shown in Fig. 8b. From Fig. 5, we expect the intensity normalization step to

alleviate microscope-induced signal variations and make micrographs more similar. This expectation is confirmed in Fig. 8a, b, which shows that image intensity distributions generally become more similar after the intensity normalization step. Nevertheless, the microscope-induced signal variations were not completely eliminated, which can be seen from the fact that intensity distribution differences still exist between different scans of the same lot after the normalization step.

The new images were not shown to the regression models during the training stage. Predictions were made by regression models trained for the corresponding old lot (in other words, trained using the old 29 lots). The BSIF model prediction median of each new scan is shown in Fig. 8c, as well as the prediction median of its corresponding old lot. We see a clear trend that dark images are generally associated with high UCS predictions while bright images usually correspond to small UCS predictions. An intensity normalization step tends to narrow the UCS prediction gap between the brightest and the darkest images but cannot remove microscope-induced brightness differences completely. This trend is observed in all the explored ML models to varying extents. Results of other models are given in the Supplementary Material.

Average performances of the new scans are summarized in Fig. 9. We see that the performance of most models (BSIF, VGG16_C_{3,3}, VGG16_FC₂, WRN) degraded when given the new images with unseen intensities. An intensity normalization step helps to improve the performance degradation in most cases, except for WRN which has the smallest MAPE to start with. However, we note that a WRN trained with the unprocessed original images is sensitive to bright

	Decile Level of \bar{I}									
	0	1	2	3	4	5	6	7	8	9
AR	1.06	1.04	1.05	1.05	1.04	1.04	1.02	1.02	1.03	1.02
F	0.96	0.99	1.02	1.01	1.01	1.01	0.98	0.97	0.95	0.91
AV	0.97	0.96	0.95	0.96	0.95	0.95	0.95	0.95	0.94	0.94
L	0.87	0.86	0.86	0.86	0.86	0.86	0.85	0.86	0.85	0.85
AP	1.21	1.19	1.18	1.17	1.17	1.17	1.16	1.16	1.15	1.13

Fig. 4 Average UCS prediction (BSIF model) of lot image intensity decile groups for five randomly sampled lots. The color scheme is determined within each row (or sample lot) and high prediction corresponds to dark color. It can be seen that the left side is generally darker than the right side, meaning that darker images generally correspond to higher UCS predictions.

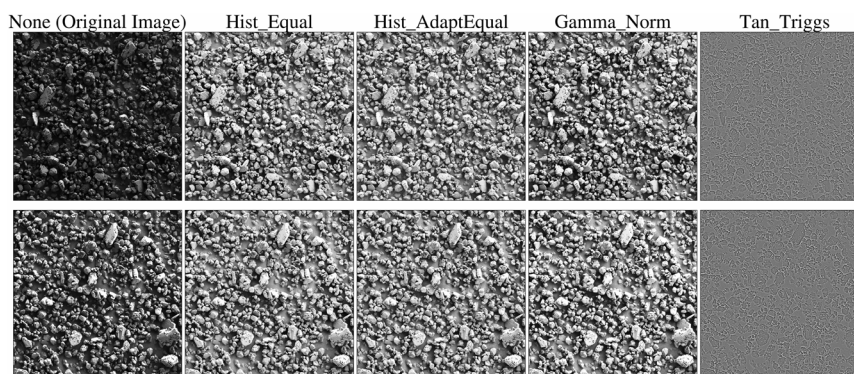


Fig. 5 Effects of intensity normalizations. This figure shows results of applying different intensity normalization techniques.

images. This trend is not obvious in the aggregated results of Fig. 9 but is clear in Supplementary Fig. 6. Though the features we implemented are by no means exhaustive, our results suggest that many ML techniques are not immune to microscope-induced image intensity variations. Different ML models show different robustness to micrograph intensity variations, and an intensity normalization step reduces micrograph intensity variations and improves prediction robustness. Histogram equalization, which shows good performance in the initial dataset (Fig. 7), continues to show good performance in the new scans with unseen brightness.

DISCUSSION

There are generally two main sources of SEM image pixel intensity: the microstructure content and the microscope condition. Variations in the microscope signals are not correlated with material properties thus should not affect ML property predictions. However, we see the opposite with our SEM micrographs and ML models, as suggested in Fig. 4, and then confirmed in Figs. 6 and 8. Figure 6 shows that on the lot averaged level, correlations between image intensity and sample lot microstructure characteristics are generally significant with or without intensity normalization (1st to 5th column of Fig. 6). On the individual image level, the correlation between \bar{I} (image brightness) and \hat{y} (UCS prediction of the image) is generally negative before intensity normalization and becomes positive after intensity normalization, except for Tan_Triggs which is known to behave very differently from the other intensity normalization methods (Fig. 5). The positive $\text{corr}(I, \hat{y})$ seems to contradict our observation that dark images are consistently associated with high UCS predictions (Figs. 4 and 8). This seeming contradiction results from the different contexts in Fig. 6, in which the positive $\text{corr}(I, \hat{y})$ is calculated for all images covering all lots and very different microstructures, and Figs. 4 and 8, in which the microstructure content is either similar or constant. The positive $\text{corr}(I, \hat{y})$ after most intensity normalizations is also consistent with the negative $\text{corr}(\langle \bar{I} \rangle, \text{size})$ seen in the 1st column of Fig. 6, as we expect $\text{corr}(\langle \bar{I} \rangle, \text{size})$ and $\text{corr}(I, \hat{y})$ to have opposite signs according to the Hall–Petch rule⁴⁸ which states that smaller grain sizes should lead to higher mechanical strengths.

Also note that $\text{corr}(I, \hat{y})$ is weak in the original images and generally becomes enhanced after intensity normalization. The stronger correlations align with our intuitions that the image intensity is correlated with its microstructure content, thus should be correlated with material properties like UCS. Figures 4 and 6 provide indirect evidence because both the microstructure-induced and the microscope-induced signals vary. Figure 8 shows that microscope-induced signals alone can have a great impact on ML model performance directly.

We observe a similar trend for all four explored feature descriptors: darker images are generally predicted to have higher UCS values, though different feature descriptors show different robustness to the microscope-induced image intensity signal variations (Figs. 4, 8c and Supplementary Material). An intensity normalization step helps to reduce microscope-induced intensity variations (Figs. 6, 8b, 8c and Supplementary Material). This does not mean that the intensity normalization step will surely improve model performance (Fig. 7), because the standardized images may be more difficult to train (Supplementary Material). The benefit of intensity normalization is much more obvious when the data quality is worse. Note, our original datasets were collected with care and had high data quality (Figs. 2 and 4). When we test models trained on the original datasets with a new dataset, in which image pixel intensities are designed to be more diverse, the intensity normalization step usually helps to improve model performance (Figs. 8, 9 and Supplementary Material).

Among the four intensity normalization methods, histogram equalization performs the best in most cases. We note that adaptive histogram equalization is sometimes considered as a more advanced intensity normalization technique than the vanilla histogram equalization because resulting images of the former can have better contrast⁴⁹. With our results, we see that histogram equalization is usually associated with better performance (Fig. 7) though images enhanced by both techniques generally look good to human eyes (Fig. 5). One possible reason is that the overall intensity within one image is more or less uniform in our SEM micrographs. In this case, the local tile treatment in adaptive histogram equalization has limited extra benefits but may magnify

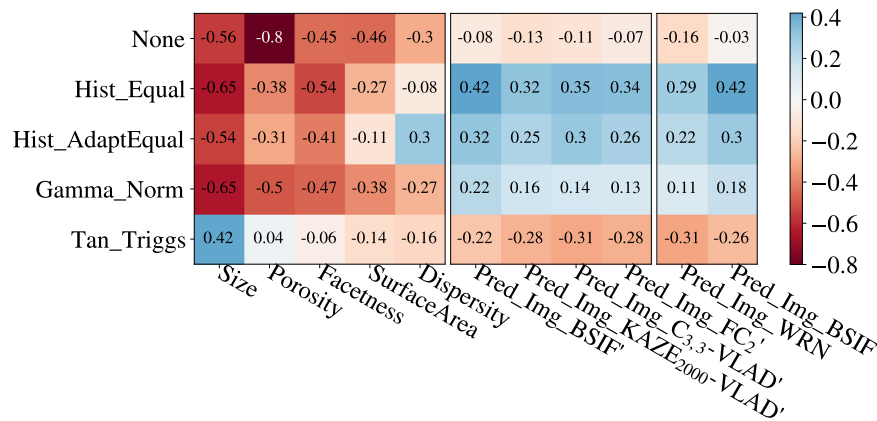


Fig. 6 Correlations between lot averaged intensity ($\langle \bar{I} \rangle$) and lot microstructure characteristics (1st to 5th column), and correlations between image intensity (\bar{I}) and image UCS predictions (6th to 11th column). The first five columns and the last two columns were evaluated on the full image set. The traditional ML models (middle four columns) were evaluated on a 6000-image subset and are denoted with primes.

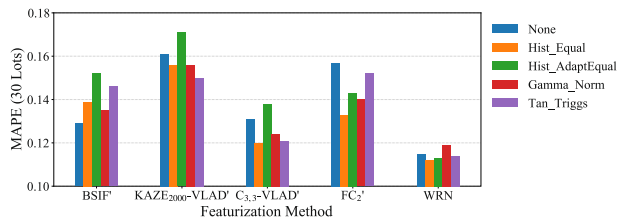


Fig. 7 Average performances of the ML models with different intensity normalization routines. Prime superscript denotes that the model was trained and evaluated on the 6000-image reduced dataset. No prime superscript means the model was trained and evaluated on the full dataset.

local noises⁴⁵ and hurt the performance of ML models. Though such noises can probably be reduced with extra parameter tuning effort, our opinion is that histogram equalization serves as a good default choice for micrograph intensity normalization due to its effectiveness, simplicity, and accessibility.

We would also like to point out some possible limitations of our work. One might conclude from Fig. 8 that bright experiment images work better than dark experiment images, but this is not true for all ML models. For example, WRN is more sensitive to bright images (Supplementary Material). We also note that the low-intensity-high-prediction correlation observed in our experiments is probably not universal but related to the characteristics of our data. Note that our material is in the form of particles, which are not flat and casts shadows on the SEM stub (Figs. 2 and 8). We have noticed that images of large particles often have more dark areas. In the case of large particles, the top particle surface is far from the stub and spaces between the particles are effectively holes. Electrons that interacted with the stub between particles have a small chance to make their way back to the detector, and the stub inevitably ends up relatively dark. This kind of darkness from depth variation is inherent to the imaging technique, thus is reproducible as long as the sample is properly preserved. However, it is not universal in all SEM images. Other signature characteristics of our material, like pores and facets, are also not ubiquitous. The point is that microscope-induced signals can be encoded into image features and affect ML models in an undesired but consistent way, not the exact dark-image-high-prediction trend.

While micrograph quality can affect ML property predictions, we note that setting a too high standard for micrograph quality during the expensive data collection process can do more harm than good, especially if the high standard leads to a reduction in

the available amount of the data. The standard of micrograph quality should vary depending on the difficulty of the material prediction task: while diverse imaging conditions may be acceptable for the classification of obviously different microstructures, more consistent conditions are probably needed for the prediction of subtly different microstructures. The user should always understand the task and the data at hand and examine ML prediction results with care rather than blindly trust them. For better data understanding, some helpful techniques include but are not limited to data visualization^{50,51}, anomaly detection⁵², and prototype (representative data points) and criticism (data points that are not well represented by prototypes) identification^{53,54}. Other interesting and emerging paths include uncertainty quantification^{55,56} and explainable machine learning⁵⁷.

In summary, we showed in this work: (1) quantitative characterizations for microscope-induced signal variations within micrographs, (2) both microstructure-induced and microscope-induced signals can affect ML model performance, (3) neither bright nor dark experiment images are universally optimal because different ML models have different robustness (traditional image features seem to be more sensitive to dark images while an end-to-end WRN is more sensitive to bright images), and (4) an intensity normalization step can help reduce microscope-induced signal variations and improve ML model robustness, and histogram equalization generally performs the best.

METHODS

Feature implementation details

Five different ML models, including four traditional ML models and one end-to-end DL model, were explored in this work. They were chosen from different popular model classes to allow reasonable generality in our conclusions. Feature descriptors in the four traditional ML models have all been applied to study micrographs by different researchers. The DL model (WRN) has not been applied to micrographs but showed great performance in natural image classification tasks⁴⁴. Some models, like the CNNs based ones, usually correspond to better model accuracies. However, we note that different feature descriptors have different complexity and require different amounts of training data and computation resources. The optimal feature to use depends on the specific purpose and the available resources.

Binarized statistical image features (BSIF)³² is a binarized texture featurization algorithm that encodes image texture information efficiently within a compact vector. Gallagher et al.³⁰ have shown that BSIF captures the microstructural texture of TATB particles reasonably well and achieved a 0.13 MAPE for the UCS prediction. The most important hyperparameter in BSIF is the convolution filter set that binarizes input images. We used the

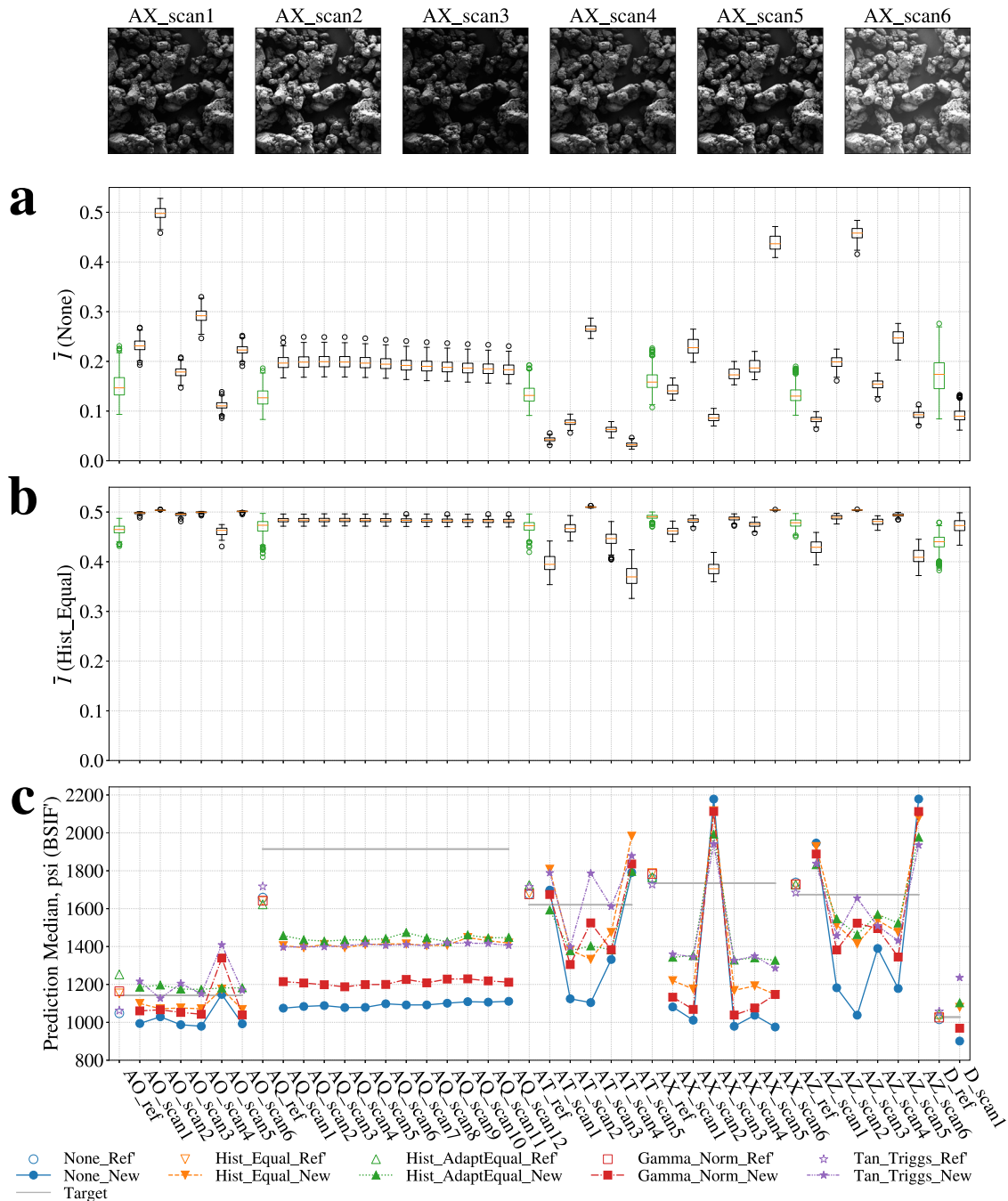


Fig. 8 Example images and statistics of the new dataset. The top row gives an example of how the same sample area (microstructure) appears in different scans according to various microscope settings. **a–c** show new lot statistics: **(a)** pixel intensity distributions of the new lot images, **(b)** image pixel intensity distributions after histogram equalization, **(c)** prediction median (BSIF model) of each new scan. Ref refers to the prediction median of the corresponding old lot in the reduced dataset. Box plots in **(a)** and **(b)** follow the same conventions as in Fig. 2.

pre-learned $11 \times 11 \times 8$ filter set provided by Kannala et al.³² and implemented the code following the examples given by the authors. Customized convolution filter sets were also learned from our SEM images but showed similar performance as the pre-learned filters.

Vector-of-Locally-Aggregated-Descriptors (VLAD)^{40,58} is a local key-point-based featurization algorithm. VLAD is inspired by the classic Bag-of-Words (BoW) algorithm³⁹ and performs well with relatively compact features⁴⁰. DeCost et al.²⁹ showed that VLAD can encode microstructures of ultrahigh carbon steel (UHCS) SEM images and achieved a 96.8% accuracy in one of their classification tasks. The detailed formula of the VLAD algorithm can be found elsewhere^{40,58}. In our work, local descriptors were implemented with the OpenCV library⁵⁹ and clustering algorithms were implemented with the Scikit-learn library⁶⁰. Note that the highly

popular SIFT local descriptor²² was not implemented because it is no longer open-source. Detailed hyperparameter tuning results are given in Table 1a. The clustering step of VLAD can take a prolonged time. Marginal performance improvement is not the focus of this work, so we set a time limit of 24 h and stopped all experiments even if they have not converged. This is why some fields of Table 1a are left blank.

CNN-based transfer learning features have less clear physical meanings than hand-crafted local descriptors but show supreme performance in many vision tasks^{61,62}. DeCost et al.²⁹ compared VLAD features built from key-point descriptors and VGG16, a popular CNN architecture, convolution layer outputs. A better performance was achieved with the CNN-VLAD (98.9%) feature in the microstructure classification task. Ling et al.²⁸ also showed that VLAD descriptors encoded from the VGG16 network

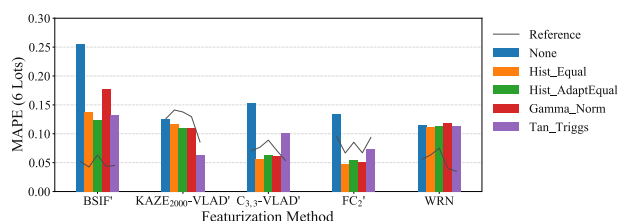


Fig. 9 Average performances of the ML models, trained on original datasets, and evaluated on the new dataset. References refer to average performances of the same six lots in the original data. Prime superscript denotes that the model was trained on the 6000-image reduced dataset. No prime superscript means the model was trained on the full dataset.

convolution layer outputs serve as effective features for various classes of micrographs. Our featurization steps are similar to those described by DeCost et al.²⁹ and Ling et al.²⁸. The CNN part was implemented with the PyTorch framework⁶³ and the VLAD part was implemented with the Scikit-learn library⁶⁰. Hyperparameter tuning results are given in Table 1b. Note that outputs from VGG16 low-level convolution layers have huge sizes and can present a great challenge to the clustering step of VLAD codebook learning. For example, the $C_{2,2}$ output of one single image has size [112, 112, 128] and occupies ~6 MB memory. To overcome this difficulty, we constrained ourselves to the reduced 6000-image dataset and used a randomly sampled subset of the total transfer learning features to learn the codebook. The subset sizes for $C_{2,2}$, $C_{3,3}$, $C_{4,3}$, $C_{5,4}$ are 0.2, 0.4, 0.8, 1.0.

Apart from CNN convolution layers, a more straightforward way to featurize images with pre-trained CNNs is to use the fully connected (FC) layers^{61,64,65}. In the field of micrograph analysis, Kitahara et al.²⁰ showed features based on FC outputs of VGG16 serve well in a classification task of surface defect SEM images. There are two FCs in the VGG16 model. We found that the second fully connected layer (FC_2) gives a slightly better performance than FC_1 in our TATB UCS prediction task and report results with on FC_2 .

End-to-end WRN implementation details

We also trained end-to-end DL models from scratch with our SEM data. Optimal network architecture design is not the focus of this work, so we screened several popular network architectures (AlexNet⁶⁶, DenseNet169⁶⁷, ResNet50⁶⁸, SqueezeNet⁶⁹, VGG16⁴²) following the default implementation in the PyTorch⁶³ package and a WRN following an open-source repository⁵⁰. The final model choice was WRN. A widen factor of 2.0 was used. In the data-preparation step, micrographs were resized to [352, 352] and normalized to 0.5 mean and 0.5 standard deviations. The batch size was set to 32. The Adam optimizer⁷⁰ was used, with a 0.001 learning rate and 0.0 weight decay. A 9:1 (training:validation) split ratio was adopted to further split the 29 lots of training data, and the model performance was evaluated using the validation data after each epoch. About 40 epochs were run for each model. The model with the best validation error during the 40 epochs was saved to evaluate the test data performance.

Intensity normalization

As for the intensity normalization algorithms, we implemented histogram equalization and (contrast limited) adaptive histogram equalization using the OpenCV library⁵⁹ with default parameters. The implementation of the Tan-Triggs normalization followed steps given by Tan et al.⁴⁶ and used the default parameter settings recommended by the authors.

Model evaluation

The test of model performance followed the leave-one-out routine as adapted by Gallagher et al.³⁰. For each of the 30 lots, a different model is built by using a lot of interest as test data and the remaining 29 lots as training data. The performance of each test lot is reported in absolute percent error (APE), which is computed using the ground truth UCS of the lot and the median UCS prediction for images within the lot. Performance of a method is reported in mean absolute percent error (MAPE), which is computed as the average APE of the involved lots.

DATA AVAILABILITY

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

CODE AVAILABILITY

Code for (image intensity) gamma normalization available at: https://github.com/xiaotingzhong/image_intensity_gamma_normalization. Other code used for these findings is available upon reasonable request.

Received: 4 February 2021; Accepted: 19 August 2021;

Published online: 05 October 2021

REFERENCES

- Bostanabad, R. et al. Computational microstructure characterization and reconstruction: review of the state-of-the-art techniques. *Prog. Mater. Sci.* **95**, 1–41 (2018).
- Ramprasad, R., Batra, R., Pilia, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 1–13 (2017).
- Ma, W. et al. Image-driven discriminative and generative machine learning algorithms for establishing microstructure-processing relationships. *J. Appl. Phys.* **128**, 134901 (2020).
- DeCost, B. L. & Holm, E. A. A computer vision approach for automated analysis and classification of microstructural image data. *Comput. Mater. Sci.* **110**, 126–133 (2015).
- Azimi, S. M., Britz, D., Engstler, M., Fritz, M. & Mücklich, F. Advanced steel microstructural classification by deep learning methods. *Sci. Rep.* **8**, 1–14 (2018).
- Bulgarevich, D. S., Tsukamoto, S., Kasuya, T., Demura, M. & Watanabe, M. Automatic steel labeling on certain microstructural constituents with image processing and machine learning tools. *Sci. Technol. Adv. Mater.* **20**, 532–542 (2019).
- Masci, J., Meier, U., Ciresan, D., Schmidhuber, J. & Fricout, G. Steel defect classification with Max-Pooling Convolutional Neural Networks. in *The 2012 International Joint Conference on Neural Networks (IJCNN)*. 1–6 (IEEE, Brisbane, QLD, 2012).
- Chowdhury, A., Kautz, E., Yener, B. & Lewis, D. Image driven machine learning methods for microstructure recognition. *Comput. Mater. Sci.* **123**, 176–187 (2016).
- Kondo, R., Yamakawa, S., Masuoka, Y., Tajima, S. & Asahi, R. Microstructure recognition using convolutional neural networks for prediction of ionic conductivity in ceramics. *Acta Mater.* **141**, 29–38 (2017).
- Cang, R., Li, H., Yao, H., Jiao, Y. & Ren, Y. Improving direct physical properties prediction of heterogeneous materials from imaging data via convolutional neural network and a morphology-aware generative model. *Comput. Mater. Sci.* **150**, 212–221 (2018).
- Kim, D. et al. Artificial intelligence for the prediction of tensile properties by using microstructural parameters in high strength steels. *Materialia* **11**, 100699 (2020).
- Unnikrishnan, S., Donovan, J., Macpherson, R. & Torrey, D. Machine learning for automated quality evaluation in pharmaceutical manufacturing of emulsions. *J. Pharm. Innov.* **15**, 392–403 (2020).
- Lapuschkin, S., Binder, A., Montavon, G., Müller, K. & Samek, W. Analyzing classifiers: fisher vectors and deep neural networks. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2912–2920 (IEEE, Las Vegas, NV, 2016).
- Arrieta, A. B. et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020).
- Guidotti, R. et al. A survey of methods for explaining black box models. *ACM Comput. Surv. CSUR* **51**, 1–42 (2018).
- Griffin, B. J. A comparison of conventional Everhart-Thornley style and in-lens secondary electron detectors—a further variable in scanning electron microscopy. *Scanning* **33**, 162–173 (2011).
- Zhou, W., Apkarian, R. P., Wang, Z. L. & Joy, D. Fundamentals of scanning electron microscopy (SEM). in *Scanning Microscopy for Nanotechnology* (eds Zhou, W. & Wang, Z. L.) 1–40 (Springer, 2006).
- Malti, A. C., Dembélé, S., Piat, N., Arnoult, C. & Marturi, N. Toward fast calibration of global drift in scanning electron microscopes with respect to time and magnification. *Int. J. Optomechatronics* **6**, 1–16 (2012).
- DeCost, B. L., Jain, H., Rollett, A. D. & Holm, E. A. Computer vision and machine learning for autonomous characterization of AM powder feedstocks. *JOM* **69**, 456–465 (2017).
- Kitahara, A. R. & Holm, E. A. Microstructure cluster analysis with transfer learning and unsupervised learning. *Integrating Mater. Manuf. Innov.* **7**, 148–156 (2018).

21. Maddern, W. et al. Illumination invariant imaging: applications in robust vision-based localisation, mapping and classification for autonomous vehicles. In *Proceedings of the Visual Place Recognition in Changing Environments Workshop, IEEE International Conference on Robotics and Automation (ICRA)*. Vol. 2–3 (IEEE, Hong Kong, 2014).
22. Lowe, D. G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**, 91–110 (2004).
23. Singh, B. & Davis, L. S. An analysis of scale invariance in object detection SNIP. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA 3578–3587 (IEEE, Salt Lake City, UT, 2018).
24. Wang, X., Han, T. X. & Yan, S. An HOG-LBP human detector with partial occlusion handling. in *2009 IEEE 12th International Conference on Computer Vision*. Japan 32–39 (IEEE, Kyoto, 2009).
25. Tsutsui, K. et al. A methodology of steel microstructure recognition using SEM images by machine learning based on textural analysis. *Mater. Today Commun.* **25**, 101514 (2020).
26. Strzelecki, M. & Materka, A. On sensitivity of texture parameters to smooth variations of local image intensity and contrast. in *2014 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. 48–53 (Poznan, 2014).
27. Materka, A. & Strzelecki, M. On the effect of image brightness and contrast nonuniformity on statistical texture parameters. *Found. Comput. Decis. Sci.* **40**, 163–185 (2015).
28. Ling, J. et al. Building data-driven models with microstructural images: generalization and interpretability. *Mater. Disco.* **10**, 19–28 (2017).
29. DeCost, B. L., Francis, T. & Holm, E. A. Exploring the microstructure manifold: image texture representations applied to ultrahigh carbon steel microstructures. *Acta Mater.* **133**, 30–40 (2017).
30. Gallagher, B. et al. Predicting compressive strength of consolidated molecular solids using computer vision and deep learning. *Mater. Des.* **190**, 108541 (2020).
31. Schwartz, A. J., Kumar, M., Adams, B. L. & Field, D. P. *Electron Backscatter Diffraction in Materials Science* Vol. 2 (Springer, 2009).
32. Kannala, J. & Rahtu, E. BSIF: binarized statistical image features. in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. 1363–1366 (IEEE, Tsukuba, 2012).
33. Alcantarilla, P. F., Bartoli, A. & Davison, A. J. KAZE Features. Computer Vision – ECCV. in *European Conference on Computer Vision*. (eds Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y. & Schmid, C.) 214–227 (Springer, 2012).
34. Rublee, E., Rabaud, V., Konolige, K. & Bradski, G. ORB: an efficient alternative to SIFT or SURF. in *2011 International Conference on Computer Vision*. 2564–2571 (IEEE, Barcelona, 2011).
35. Agrawal, M., Konolige, K. & Blas, M. R. CenSurE: center surround extremas for realtime feature detection and matching. in *European Conference on Computer Vision*. (eds Forsyth, D., Torr, P. & Zisserman, A.) 102–115 (Springer, 2008).
36. Calonder, M., Lepetit, V., Strecha, C. & Fua, P. BRIEF: binary robust independent elementary features. in *European Conference on Computer Vision*. (eds Daniilidis, K., Maragos, P. & Paragios, N.) 778–792 (Springer, 2010).
37. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28**, 129–137 (1982).
38. Murphy, K. P. *Machine Learning: A Probabilistic Perspective* (MIT Press, 2012).
39. Csurka, G., Dance, C. R., Fan, L., Willamowski, J. & Bray, C. Visual categorization with bags of keypoints. *Workshop Stat. Learn. computer Vis., ECCV* **1**, 1–2 (2004).
40. Arandjelovic, R. & Zisserman, A. All about VLAD. in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1578–1585 (IEEE, Portland, OR, 2013).
41. Deng, J. et al. ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255 (IEEE, Miami, FL, 2009).
42. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. in *International Conference on Learning Representations* (San Diego, CA, 2015).
43. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
44. Zagoruyko, S. & Komodakis, N. Wide residual networks. in *Proceedings of the British Machine Vision Conference (BMVC)* (eds Wilson, R. C., Hancock, E. R. & Smith, W. A. P.) 87.1–87.12 (BMVA Press, 2016).
45. Pizer, S. M. et al. Adaptive histogram equalization and its variations. *Comput. Vis. Graph. Image Process.* **39**, 355–368 (1987).
46. Tan, X. & Triggs, B. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Trans. Image Process.* **19**, 1635–1650 (2010).
47. Kaur, M., Kaur, J. & Kaur, J. Survey of contrast enhancement techniques based on histogram equalization. *Int. J. Adv. Comput. Sci. Appl.* **2** (2011).
48. Cordero, Z. C., Knight, B. E. & Schuh, C. A. Six decades of the Hall–Petch effect—a survey of grain-size strengthening studies on pure metals. *Int. Mater. Rev.* **61**, 495–512 (2016).
49. Reza, A. M. Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement. *J. VLSI Signal Process. Syst. Signal Image Video Technol.* **38**, 35–44 (2004).
50. Kuen, Jason. WideResNet-pytorch. *GitHub Repository* <https://github.com/xternalz/WideResNet-pytorch> (2019).
51. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. (2018). <https://arxiv.org/pdf/1802.03426.pdf>.
52. Chandola, V., Banerjee, A. & Kumar, V. Anomaly detection: a survey. *ACM Comput. Surv.* **41**, 15:1–15:58 (2009).
53. Bien, J. & Tibshirani, R. Prototype selection for interpretable classification. *Ann. Appl. Stat.* **5**, 2403–2424 (2011).
54. Kim, B., Khanna, R. & Koyejo, O. O. Examples are not enough, learn to criticize! Criticism for Interpretability. *Adv. Neural Inf. Process. Syst.* (eds Lee, D., Sugiyama, M., Luxburg, U., Guyon, I. & Garnett, R.) **29** (Curran Associates, Inc. 2016).
55. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. in *Proceedings of The 33rd International Conference on Machine Learning*. (eds Balcan, M. F. and Weinberger, K. Q.) 1050–1059 (PMLR, 2016).
56. Amodei, D. et al. Concrete problems in AI safety. Preprint at <https://arxiv.org/pdf/1606.06565.pdf> (2016).
57. Lipton, Z. C. The Mythos of Model Interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue* **16**, 31–57 (2018).
58. Jégou, H., Douze, M., Schmid, C. & Pérez, P. Aggregating local descriptors into a compact image representation. in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 3304–3311 (San Francisco, CA, 2010).
59. Bradski, G. The OpenCV Library. *Dr Dobbs J. Softw. Tools* (2000).
60. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
61. Zheng, L., Yang, Y. & Tian, Q. SIFT Meets CNN: a decade survey of instance retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 1224–1244 (2017).
62. Yue-Hei Ng, J., Yang, F. & Davis, L. S. Exploiting local features from deep networks for image retrieval. in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 53–61 (IEEE, Boston, MA, 2015).
63. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. in *Advances in Neural Information Processing Systems*. (eds Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. & Garnett, R.) 32 8026–8037 (Curran Associates, Inc., 2019).
64. Sharif Razavian, A., Azizpour, H., Sullivan, J. & Carlsson, S. CNN features off-the-shelf: an astounding baseline for recognition. in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 512–519 (IEEE, Columbus, OH, 2014).
65. Babenko, A., Slesarev, A., Chigorin, A. & Lempitsky, V. Neural codes for image retrieval. in *European Conference on Computer Vision*. (eds Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T.) 584–599 (Springer 2014).
66. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
67. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2261–2269 (IEEE, Honolulu, HI, 2017).
68. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778 (IEEE, Las Vegas, NV, 2016).
69. Iandola, F. N. et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. (2016). <https://arxiv.org/pdf/1602.07360.pdf>.
70. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. in *International Conference on Learning Representations* (San Diego, CA, 2015).

ACKNOWLEDGEMENTS

The authors would like to thank Donald Loveland, Jize Zhang, and Piyush Karande for prototype codes and helpful discussions. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and was supported by the LLNL-LDRD Program under Project No. 19-SI-001.

AUTHOR CONTRIBUTIONS

T.Y.-J.H. and B.G. conceived the project. E.R. collected the experiment image data. T.N.M. designed the gamma normalization. X.Z. ran the computation experiments. X.Z., K.E., B.G., and T.Y.-J.H. analyzed the results. All authors discussed the results and contributed to the writing of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-021-00616-3>.

Correspondence and requests for materials should be addressed to T.Yong-Jin Han

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021