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

Xiaoting Zhong, Brian Gallagher, Keenan Eves, Emily Robertson, T. Nathan Mundhenk and T. Yong-Jin Han

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 classes, 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. Researcher have shown that micrograph-based ML models can achieve highly accurate material classification, defect detection, property prediction, and material quality monitoring.

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. 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 collection. 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 they may affect ML model predictions. To be clear, an ML model in this context means either a traditional ML model, in which the model is a combination of a feature extraction step and a prediction (regression) step, or an end-to-end deep-learning (DL) model, in which the feature extraction step and the prediction step are not explicitly separable. Feature extraction 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. 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 accurately. In the field of medical images, Strzelecki and Materka et al. also reported inhomogeneous brightness and contrast of magnetic resonance images (MRI) affect the calculation of statistical texture descriptors.

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

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.
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, 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, 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 published 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.

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 \( I \), \( i = 1, \ldots, 1,048,576 \), where 1,048,576 is the number of pixels in the image. \( 0 \leq I_{i} \leq 1 \) is equal to black and 1 is equal to white. Four image pixel intensity metrics, including mean \( \mu \), median \( \tilde{I} \), mode, and standard deviation \( \sigma_{I} \), of \( \{ I_{i} \} \) were calculated to characterize the overall image pixel intensity level from perspectives of brightness and contrast. \( \tilde{I}, I, \mu, \sigma_{I} \), and mode are designed to capture the overall image brightness level. \( \mu \) is intended to capture the image contrast. Four example images...
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 $I$ and $l$ capture the overall image brightness reasonably well while mode does not. $s_i$ is also a proper approximation of the image contrast. Variations of $I$, $l$, and $s_i$ are generally similar, and we will only report $I$ as the image pixel intensity metric in the remaining of this document.

The distributions of $I$ in the 30 lots are shown in the second row of Fig. 2. It can be seen that $I$ 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 volume. 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 particles. 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 $<I>$, $<l>$, and $<s_i>$, in which $<x> = \sum_{i=1}^{n}x_i/n$ 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 $<I>$ 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), a local key-point-based featurization (KAZE, ORB, STAR detector), the number of clusters in the VLAD step.
The predictor in all four traditional ML models was kept as the same random forest regressor\(^43\) (RF) used by Gallagher et al.\(^30\) without fine-tuning. These four traditional ML models will be referred to by their featurization choice (BSIF, KAZE\(_{2000,VLAD},\) C\(_{1,3^3,VLAD},\) and FC) 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 \(I\) from small (dark) to big (bright), and then cut into ten groups according to deciles of \(I\). Each decile group has approximately the same number of images. The average UCS prediction of the lot image decile group, \(\langle y_{BSIF,\gamma} \rangle_{\gamma} / \gamma\), is calculated for every decile group of each lot. \(y_{\gamma}\) is the ground truth UCS of the lot \(\gamma\). The UCS prediction of one image. Angle bracket denotes average, and \(\langle y_{BSIF,\gamma} \rangle_{\gamma, d}\) is the average BSIF model prediction for images within the decile level \(d\) of lot \(\gamma\). 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 \((I\) decile level = 0\)) and the brightest \((I\) 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 equalization\(^45\), gamma normalization, and Tan-Triggs normalization\(^46\). Histogram equalization is one of the most common image enhancement algorithms.\(^47\) 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 normalization\(^46\) 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 \(\langle I \rangle\) (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 (\(\langle I \rangle\)) 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 \(I\) (the pixel intensity score of an image) and \(y\) (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 corr(\(I, y\)), the correlation between image pixel intensity \(I\) and image UCS prediction \(y\), 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 | KM32 | KM64 |
| ORB 500 | 0.229 | 0.235 | 0.229 | 0.233 |
| 1000 | 0.221 | 0.228 | 0.215 | 0.226 |
| 2000 | 0.212 | 0.221 | 0.207 | 0.216 |
| BRIEF 500 | 0.250 | 0.263 | 0.250 | 0.257 |
| 1000 | 0.243 | 0.257 | 0.237 | 0.250 |
| 2000 | 0.226 | 0.243 | 0.208 | 0.212 |
| KAZE 500 | 0.189 | 0.202 | 0.184 | 0.191 |
| 1000 | 0.178 | 0.188 | 0.174 | 0.183 |
| 2000 | **0.164** | 0.172 | \ | \ |

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_{i,j} 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.

| (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) | DenseNet169 | ResNet50 | WRN | SqueezeNet | VGG16 |
|-----|-------------|----------|-----|-----------|------|
| AlexNet | 0.151 | 0.187 | 0.136 | **0.115** | 0.271 |
| 0.240 | 0.271 | 0.240 | 0.240 | 0.240 |

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 (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_{i,j}, VGG16_FC_{i,j}, 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
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 image intensity and 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 contrast. 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...
local noises 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 cast 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, anomaly detection, and prototype (representative data points) identification. Other interesting and emerging paths include uncertainty quantification and explainable machine learning.

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 tasks. Some models, like the CNNs based ones, usually correspond to better model accuracies. We 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 tasks. 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)** is a binarized texture featureization algorithm that encodes image texture information efficiently within a compact vector. Gallagher et al. 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

### Table 1: Average performances of the ML models with different intensity normalization routines.

| Featurization Method | Note | Hist_Equal | Hist_AdaptEqual | Gamma_Norm | Tan_Trigger |
|----------------------|------|-------------|-----------------|------------|-------------|
| BSIF                 | 0.05 | 0.08        | 0.46            | 0.55       | 0.22        |
| KAZE                  | 0.06 | 0.35        | 0.34            | 0.34       | 0.33        |
| ViAD                  | 0.07 | 0.34        | 0.36            | 0.36       | 0.31        |
| FC2                   | 0.08 | 0.36        | 0.37            | 0.37       | 0.33        |
| WRN                  | 0.09 | 0.37        | 0.35            | 0.36       | 0.35        |

**Fig. 6** Correlations between lot averaged intensity (\(\bar{I}\)) and lot microstructure characteristics (1st to 5th column), and correlations between image intensity (\(I\)) 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.
pre-learned 11 × 11 × 8 filter set provided by Kannala et al. and implemented the code following the examples given by the authors. Customized convolution filters were also learned from our SEM images but showed similar performance as the pre-learned filters.

Vector-of-Locally-Aggregated-Descriptors (VLAD) is a local key-point-based featurization algorithm. VLAD is inspired by the classic Bag-of-Words (BoW) algorithm and performs well with relatively compact features. DeCost et al. 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. In our work, local descriptors were implemented with the OpenCV library and clustering algorithms were implemented with the Scikit-learn library. Note that the highly popular SIFT local descriptor was not implemented because it is no longer open-source. Detailed hyperparameter tuning results are given in Table 1. 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 1 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. DeCost et al. 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. also showed that VLAD descriptors encoded from the VGG16 network.

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.
convolution layer outputs serve as effective features for various classes of micrographs. Our featureization steps are similar to those described by DeCost et al.\(^{29}\) and Ling et al.\(^{30}\). The CNN part was implemented with the PyTorch framework\(^ {43}\) and the VLAD part was implemented with the Scikit-learn library\(^ {60}\). 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 featureize images with pre-trained CNNs is to use the fully connected (FC) layers\(^ {61,64,68}\). In the field of micrograph analysis, Kitahara et al.\(^ {50}\) 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 (AlexNet\(^ {66}\), DenseNet\(^ {169}\), ResNet\(^ {50}\), SqueezeNet\(^ {69}\), VGG16\(^ {42}\)) following the default implementation in the PyTorch\(^ {63}\) package and a WRN following an open-source repository\(^ {28}\). The final model choice was WRN. A wider 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 optimizer\(^ {29}\) was used, with a 0.001 learning rate and 0.0 weight decay. A 9:1 (training:validation) split ratio was used.

### Intensity normalization

As for the intensity normalization algorithms, we implemented histogram equalization and (contrast limited) adaptive histogram equalization using the OpenCV library\(^ {50}\) 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 average 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.
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 Advances in Neural Information Processing Systems. (eds Lee, D., Sugiyama, M., Luxburg, U., Guyon, I. & Garnett, R.) 29 (Curran Associates, Inc., 2016).

24. Wang, X., Han, 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. Tsutsumi, 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. Disc. 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. Predictive compressive strength of consolidated molecular solids using computer vision and deep learning. Mater. Des. 163 (2019).

31. Schwartz, A. J., Kumar, M., Adams, B. L. & Field, D. P. Predicting compressive strength of consolidated molecular solids using computer vision and deep learning. Mater. Des. 190 (2021).

32. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning model library. in 6th Python in Science Conference. (eds Millman, B., Botros, R. & Wegmueller, M.) 25:1–25:12 (2017).

33. Agrawal, M., Konolige, K. & Blas, M. R. CenSurE: center surround extremas for realtime feature detection and matching. in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. 512–519 (IEEE, Boston, MA, 2015).

34. Zhang, L., Wei, K., Zhan, Z. & Feng, W. A. P. Learning to compress. in Advances in Neural Information Processing Systems. (eds Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E. & Garnett, R.) 29 8206–8207 (Curran Associates, Inc., 2016).

35. Zhang, L., Yang, Y. & Tian, Q. SIFT Meets CNN: a decade survey of instance retrieval. in 2017 IEEE International Conference on Computer Vision (ICCV). 3304–3311 (San Francisco, CA, 2017).

36. Amodei, D. et al. Concrete problems in AI safety. Preprint at https://arxiv.org/pdf/1606.06556.pdf (2016).

37. 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).

38. 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).

39. Bradska, G. The OpenCV Library. Dr Dobbs J. Softw. Tools (2000).

40. Pedregosa, F. et al. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).

41. 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).

42. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. in Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops. 512–519 (IEEE, Boston, MA, 2015).

43. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning model library. in Advances in Neural Information Processing Systems. (eds Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E. & Garnett, R.) 29 32 8206–8207 (Curran Associates, Inc., 2019).

44. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. Commun. ACM 60, 84–90 (2017).

45. Zhang, L., Wei, K., Zhan, Z. & Feng, W. A. P. Learning to compress. in Advances in Neural Information Processing Systems. (eds Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E. & Garnett, R.) 29 8206–8207 (Curran Associates, Inc., 2016).

46. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. Commun. ACM 60, 84–90 (2017).

47. 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).

48. 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).

49. Razavian, A. et al. CaffeNet: an implementation of Convolutional Neural Networks for Computer Vision. in Proceedings of the 2014 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.

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences
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