A Machine Learning Model to Automate Quantitative Evaluation of Line Pipe Microstructures

Peter Martin[1], Nathan Switzner[1], Joel Anderson[1], Joshua Stuckner[2], Owen Lopez-Oneal[3], Sophia Curiel[3], Peter Veloo[3]

[1]RSI Pipeline Solutions, [2]NASA Glenn Research Center, [3]Pacific Gas and Electric Company

Pipeline Pigging and Integrity Management Conference

February 12-16, 2024

Organized by

Clarion Technical Conferences

https://doi.org/10.52202/072781-0023

https://doi.org/10.52202/072781-0023

Abstract

Pipeline operators in the United States are increasingly relying upon materials verification programs (MVP) to establish the properties of pipelines lacking reliable records. The ongoing MVP at the Pacific Gas and Electric Company (PG&E) applies external nondestructive testing (NDT) to exposed line pipe to gain insight into the grade, vintage, and manufacturing method of the pipe. PG&E supplements the standard NDT methods for composition, strength, and geometry with the nondestructive collection of surface microstructures using metallographic replicas. The microstructures are quantitatively evaluated to determine the ferrite grain size and fraction of pearlite (dark constituent). These are then used in conjunction with other measured characteristics to support determination of grade and vintage, and to identify populations of similar pipes. Automating the quantitative evaluations is of interest because the manual evaluations are labor intensive and subject to variability associated with evaluator skill, judgement, and fatigue.

Traditional methods for automating image analyses are often challenged by small variations in sample or image quality that are ubiquitous in metallographic microstructure images. Machine learning (ML) models have been shown to be more robust, but training these models typically requires hundreds or thousands of manually pre-processed images. This creates a high initial investment that impedes practical implementation in an operational environment. Recently, pre-training ML models with a large number of generic images has been shown to substantially reduce the required number of application-specific training images. This work will describe the performance of an open-source ML model pre-trained on a database of over 10^5 microscopy images and subsequently 'finetuned' on 17 line pipe microstructures. The training and validation of the model will be described, and criteria for automated screening of discrepant results will be proposed and validated. Results from automated evaluations will be compared to corresponding manual evaluations from more than 170 microstructures from more than 50 line pipes. The automated results will be shown to be generally equivalent to the manual results, and a few outlier results will be examined in more detail to illustrate opportunities to improve performance in a next-generation version of the model.

Background and motivation

As part of their materials verification program, The Pacific Gas and Electric Company (PG&E) may collect metallographic images of the microstructures in line pipe steels using both field microscopy and nondestructive surface replicas. Since the observed microstructures result from complex interactions between chemistry and thermomechanical processing route, they can provide a 'fingerprint' that corresponds to properties within a known range and can be often used to draw an equivalency (or exclude one) between groups of pipes. Efforts to date have demonstrated that routine analysis of microstructure can contribute valuable insights during materials verification and MAOP reconfirmation [1], including:

- Verifying reliability of reported installation date and pipe grade from existing records [2,3]
- Determining the equivalency of features with and without reliable records
- Establishing the manufacturing process
- Verifying proper surface preparation for chemical and strength analysis
- Interpreting the validity of NDT for yield strength (YS) and assessing the possible mismatch with YS from destructive testing [4]

https://doi.org/10.52202/072781-0023

In addition, quantitative characterization of the microstructure can allow for estimation of physical properties via correlations developed using a database of pipes with known properties and quantified microstructures [5,[6]]. To accomplish this, PG&E currently performs quantitative analyses of line pipe microstructures using standard manual methodologies. The methodologies include comparison and counting methods for both ferrite grain size (GS) and percentage of dark constituent (%DC) [7]. Ideally, a typical analysis process for a single microstructure (i.e., one test location) takes the average result from multiple (3 or more) evaluators applying one or two methods on multiple (5) images. Multiple images are needed to ensure representative sampling of the microstructure, multiple evaluators to avoid observation bias and process drift, and multiple methods to enable real-time self-checking of results (the methods should agree if performed properly). When combined with subject matter expert (SME) oversight and annual retraining, this approach results in reliable and reproducible results; however, experience has revealed that the manual evaluations can become overly resource intensive as the volume of images to be analyzed increases. In addition, without SME oversight the results are subject to a risk of inconsistent measurements and/or process drift. As a result, automation of the process is of interest as a means to decrease resource demand and improve process stability.

In theory, the image processing tools required to automate measurement of GS and %DC are relatively straightforward: counting particles (grains) and dark pixels. In practice, implementation is challenged by several aspects of the specific microstructures and the quality of the images, Figure 1. Grain boundaries vary widely in appearance (thickness, color, density) and often have gaps that a human can interpolate but that create challenges for automated differentiation of discrete grains. In addition, the appearance of dark constituent (DC) ranges from dark and quasi-uniform to mottled with colorful bright regions, creating challenges with respect to reproducibly defining which pixels to count as DC. It can also be unclear how to differentiate between grain boundaries and DC since they both appear dark in the images. Finally, the surface replicas often have artifacts that aren't present in traditional metallographic samples/images, such as bubbles that appear as bright spots. These issues create challenges for traditional image processing because it relies on 'thresholding', the process of assigning a brightness or color level to differentiate one phase from another. Resolving these challenges generally requires human intervention, and it is often faster to manually analyze the microstructure (especially by comparison methods) than to optimize the image for automated processing.

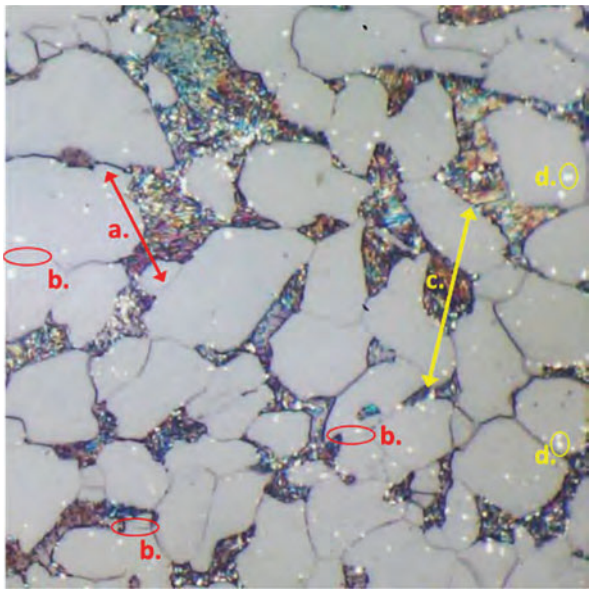

Figure 1. Example of several quality issues typical of microstructure images collected by nondestructive surface replication: (a) different grain boundary appearances, (b) gaps in grain boundaries, (c) different DC appearance, and (d) replication bubbles.

Recently, a machine learning (ML) method has been proposed to process (segment) the images to enable subsequent automated analysis by these standard image processing tools [8]. The method, which is implemented in Python, uses a convolutional neural network (CNN) that has been pre-trained on a database of over 10^5 microscopy images to minimize the

https://doi.org/10.52202/072781-0023

number of training images required to finetune the model for line pipe microstructures. The outputs of the ML model are segmented images in which the features of interest are identified by a unique color, one for DC and one for ferrite grains. Analysis of the segmented images to determine the ferrite grain size and %DC is then effectively automated using standard image processing tools such as ImageJ. The combined steps of pre-processing the images, segmenting them using the ML model, analyzing them in ImageJ, and compiling and quality checking (QC'ing) the results in MS Excel are referred to below as the 'automated' method.

Experimental methods

Collection of microstructure images

Microstructure images were collected from nondestructive replicas taken from pipe surfaces. Replication nondestructively records the topography of a metallographic specimen (the pipe surface) for subsequent laboratory examination by creating a negative relief of the prepared surface with a plastic film [7]. It is especially valuable when field conditions preclude direct imaging of the pipe surface, for example when the pipe is vibrating or has a small diameter. Microstructural images taken from replicas were used in this investigation because they are consistent with field (in-ditch) results accessible by nondestructive methods. Preparation of the pipe surfaces for replication is the same as for standard field metallography [9,10]. The pipe surface is first cleaned, the top surface is then removed (buffed) to a depth of ≈0.010 in. to mitigate the potential effects of near-surface decarburization, the surface is then polished with diamond paste to a finish of 1 μm. After polishing, the surface is etched with ≈5% nital to expose the microstructure.

Once the pipe surface is properly prepared and the microstructure has been exposed by etching, a cellulose acetate tape is wetted with acetone and pressed onto the etched surface where it is allowed to harden (typically within a few minutes). Subsequently, the back of the acetate tape is painted with a permanent black marker to enhance imaging contrast, and the tape replica is removed from the pipe surface and mounted on a glass slide for transportation and imaging in a standard laboratory metallograph. Images are typically collected at magnifications of 200x and 500x, and scaling is indicated by calibrated scalebars incorporated into the images.

Manual evaluation of microstructures

The automated microstructure evaluations were validated, in part, by comparison to extensive manual evaluations performed as part of the MVP development and implementation at PG&E. The manual evaluations assessed %DC and ferrite grain size using both counting and comparison methods adapted from ASTM standards [5,7]. Manual grain size evaluations used the methods described in the ASTM E112, Section 10 (comparison procedure) and Section 11 (counting method) [11]. %DC evaluations used the counting method from ASTM E562 and a comparison method developed internally by PG&E and RSI based on the ASTM E112 grain size comparison method [5,12].

Automated evaluation of microstructures

The automated microstructure analysis is enabled by the use of a ML model to segment the microstructure images for subsequent analysis using standard image processing tools. The ML model

https://doi.org/10.52202/072781-0023

Table 1. Characteristics of the 17 pipes used for training.

Sample	Year	Seam	Grade	OD, in.	Wall, in.
1	1947	SMLS	Grade B	16.000	0.313
2	1947	SMLS	Grade B	16.000	0.313
3	1947	SMLS	Grade B	8.625	0.277
4	1949	SMLS	Grade B	16.000	0.312
5	1961	SMLS	X46	12.750	0.500
6	1961	SMLS	X46	12.750	0.500
7	1966	SMLS	Unknown	16.000	0.368
8	1966	SMLS	Unknown	8.625	0.522
9	1981	ERW	Unknown	8.625	0.188
10	1986	SMLS	Grade B	8.625	0.322
11	1987	SMLS	Unknown	4.500	0.237
12	1991	SMLS	Unknown	8.625	0.322
13	1991	SMLS	Unknown	8.625	0.322
14	2007	SMLS	Grade B	16.000	0.313
15	2014	SMLS	Grade B	4.500	0.337
16	2014	SAWL	X52	36.000	0.500
17	2017	ERW	X52	24.000	0.500

applies a supervised deep learning approach implemented through a CNN using an encoder-decoder architecture. The model architecture is UnetPlusPlus and the encoder is se_resnext50_32x4d. These details are described further in [8] and example code is available at [13]. The primary novelty of the approach is that the model was pre-trained on a database of over 10^5 microscopy images, called MicroNet, which mitigates the need for an extensive set of manually segmented application-specific images for final training (i.e., 'finetuning'). Instead, the pre-trained model was optimized for evaluation of line pipe microstructures by finetuning with 17 manually segmented line pipe microstructures. The benefit of pre-training is significant since preparation of manually segmented training images requires 2 to 4 hours per image, and because the number of available images is often limited from a practical perspective. For example, PG&E has collected approximately 600 microstructure images in the last two years of its MVP.

The training images used for finetuning the model consisted of microstructures collected from surface replicas of line pipe steels, except for one transverse cross-section that was included to provide an example of a very coarse hot-rolled microstructure, Sample 15 in Table 1. The training images were selected to provide a representative range of image qualities, ranging from high contrast with few replication artifacts to low contrast with poor grain boundary resolution and numerous artifacts, Figure 2. This was to ensure that the training set would be representative of images from actual in-ditch inspections, both in content and quality. The pipes from which the microstructures were collected had manufacturing dates from 1947 to 2017 and Grades from B to X52 (with six unknown), Table 1. The long term training strategy was to initially focus on regular, more easily recognized hot-rolled microstructures, and to implement lessons-learned in a future update to include the more complex microstructures generated by quenching and other non-equilibrium processes. Thus, the microstructure types used for the training images in this work were mostly hot-rolled or normalized

https://doi.org/10.52202/072781-0023

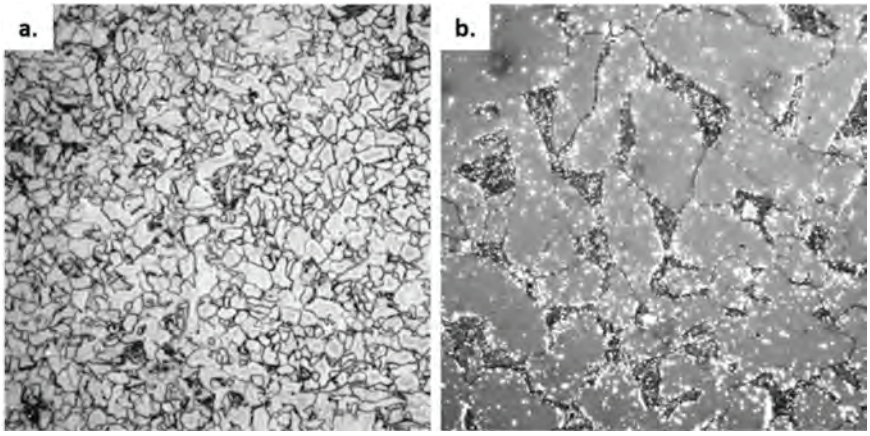

Figure 2. Examples of (a) high- and (b) low-quality images from the training set (samples 17 and 12, respectively).

(e.g., Figure 1 and Figure 2b), with two examples of TMCP (e.g., Figure 2a), and no examples of quench and tempered. The ferrite grain size and %DC determined by manual methods ranged from 12.0 to 7.7 G (MLI of 4.9 to 22.2 μm)[1] and 2.3% to 47.5%, respectively. These values are tabulated in Table 2, along with the weight % of C, Mn, Si and S.

Table 2. Composition, %DC, and grain size (G) of the 17 pipes used for training images.

Sample	C, %	Mn, %	Si, %	S, %	%DC	G
1	0.220	0.420	<0.001	0.022	20.3	8.7
2	0.180	1.040	0.220	0.009	8.2	10.2
3	0.250	0.810	<0.001	0.024	37.0	9.9
4	0.130	0.425	0.255	0.027	9.9	8.5
5	0.235	0.705	<0.001	0.0185	26.1	8.8
6	0.280	1.405	<0.001	0.0285	47.5	9.4
7	0.230	0.530	0.050	0.023	25.0	9.0
8	0.025	0.770	0.020	0.026	28.8	9.0
9	0.154	0.620	0.116	0.028	3.2	12.0
10	0.220	0.840	0.380	0.015	34.0	9.4
11	0.183	0.430	0.068	0.016	17.3	10.7
12	0.170	0.440	0.210	0.015	10.3	7.7
13	0.270	0.870	0.240	0.004	40.0	9.7
14	0.200	0.950	0.210	0.006	15.0	10.4
15	0.200	0.845	0.195	0.002	30.2	9.2
16	0.140	1.040	0.330	0.001	23.3	10.5
17	0.070	1.140	0.190	0.002	2.3	12.0

The training images were 512×512 pixel (px) slices taken from larger raw images, which were typically 1600×1200 px or 2588×1960 px. In some cases, the original images were rescaled prior to slicing in order to ensure a representative set of features in the 512×512 px slice. For training, each image was accompanied by a manually segmented copy to provide 'ground truth' for the model. The manual segmentation was performed in an open source image processing software (GIMP 2.10.34) by coloring the ferrite grains red (255,0,0)[2], the DC blue (0,0,255), and leaving the grain boundaries uncolored. The subsequent DC analysis determines the %DC directly from the segmented area of the DC; therefore, accurate outlining of the DC colonies during manual segmentation was important.

[1] Mean linear intercept (MLI, μm) = $320 \cdot 2^{(-G/2)}$.

[2] Red, Green, Blue (RGB) pixel values which range from 0 to 255 are indicated by (R, G, B).

https://doi.org/10.52202/072781-0023

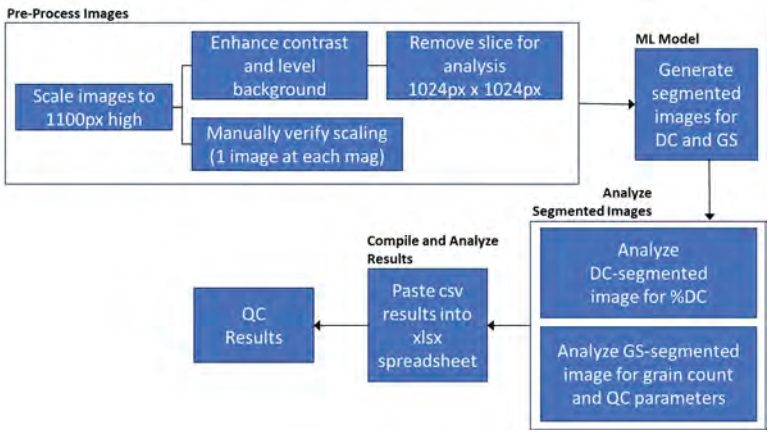

Figure 3. Process flow of the automated process for evaluating microstructures.

However, the subsequent grain size analysis relies on grain counting, not direct measurement of grain areas, because it is designed to duplicate the counting method outlined in ASTM E112. Thus, accurately outlining the (ferrite) grains was not critical as long as the grains could be counted. This was important for the manual segmentation because it mitigated the challenges associated with accurately tracing the width of the grain boundaries, which varies significantly across the various microstructures. Finally, the set of images was divided into training and validation groups of 12 and 6 images[3], respectively.

Finetuning the model was performed in Python as an iterative process that monitors changes in performance from one iteration (epoch) to the next. In this case, the performance of the model is related to a parameter that quantifies the difference between the manually and ML segmented images on a pixel-by-pixel basis. The training used a learning rate of 10^{-5}, which controls how fast the ML model parameters change with each epoch (a learning rate that is too low will take too long to train while a high learning rate will fail to converge). Early stopping with a patience of 15 was used, meaning that if model performance fails to improve for 15 consecutive iterations the training will stop and the optimized model will be saved. Early stopping prevents 'overfitting' where the model performs well on the training dataset, but poorly on the validation dataset.

The steps for practical implementation of the optimized automated model are shown schematically in Figure 3. Raw images are pre-processed in ImageJ, the image scaling (px/μm) is measured manually, the finetuned ML model is run to generate the segmented images, the segmented images are analyzed in ImageJ, and the results are compiled and evaluated in MS Excel. Figure 4 shows an example of the pre-processing steps. The raw image shown at left is scaled to 1100 px high and standard image processing tools are used to sharpen the image, enhance the contrast, and level the background to create the 'scaled' image shown at the center. Finally, the 1024×1024 px 'slice' shown at right is

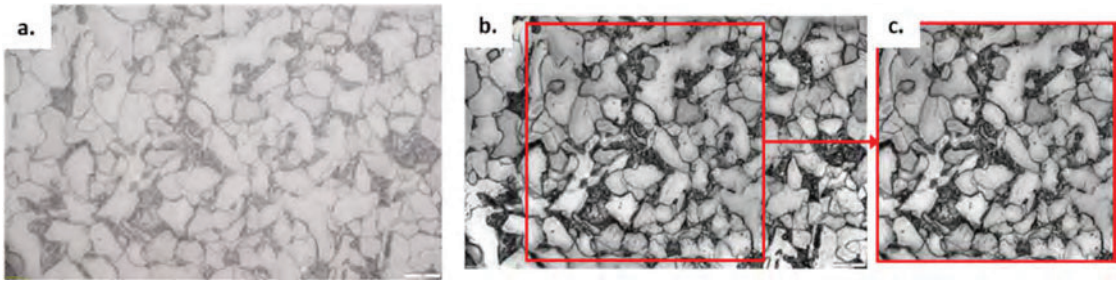

Figure 4. Example of the image pre-processing: (a) raw image, (b) same image after scaling, enhancing the contrast, and levelling the background, (c) 1024x1024 px slice used for analysis.

[3] One image from the training set was reused in the validation set after rotating 180°.

https://doi.org/10.52202/072781-0023

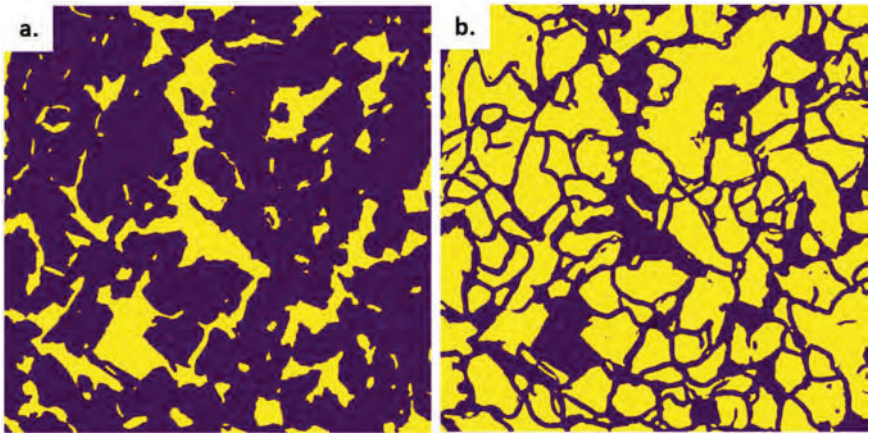

Figure 5. Segmented images generated by the ML model for the 1024x1024 px slice shown in Figure 4. The images highlight in yellow the (a) DC and (b) ferrite grains.

copied from the scaled image and saved for evaluation. These actions can be performed on batches of images by use of a suitable macro script. Manual measurement of the image scaling is easily performed in any image manipulation software by measuring the length (in pixels) of the scalebar in the scaled image, and it only needs to be performed on one image for each combination of magnification and image size.

The ML model is called in Python and outputs two segmented images for each 1024×1024 px input image. The segmented images highlight in yellow the colonies of dark constituent (DC) or the ferrite grains, respectively. Example segmented images corresponding to the preprocessed image from Figure 4 are shown in Figure 5. These images are analyzed in ImageJ to determine the %DC and the number of ferrite grains. Analysis of the DC segmented image comprises summing the area (in px^2) of the yellow-highlighted DC relative to the total number of pixels in the image (1024^2). That analysis considers only fields of 100 px^2 or larger in order to minimize errors associated with small particulates and image artifacts. The grain size segmented images are analyzed by counting the number of yellow-highlighted ferrite grains twice, once including and once excluding the grains that intersect the edges of the image. Those values are subsequently used, along with the image scaling, %DC, and image size, to determine the average grain size. The minimum particle (grain) size to be counted is again limited to mitigate overcounting due to noise in the image. In this case, the limit is based on an initial count performed with the limit set to 60 px^2. If that initial count exceeds 1000 grains, then the grains are small enough that the 60 px^2 threshold is likely to exclude a non-negligible portion of the actual grain size distribution. In that case, the minimum is decreased to 10 px^2 and the count is repeated. Both the %DC and grain counting pixel-limits were set by trial and error. These processes are easily implemented on a batch basis in ImageJ by calling the 'Analyze Particles' command from a macro script. That command is a pre-packaged macro that scans a binary image until it finds the edge of an object, outlines the object using the wand tool, measures it using the Measure command, fills it to make it invisible, then resumes scanning until it reaches the end of the image. The raw results and/or summary statistics can be tabulated and exported to a .csv file.

In practice, the numerical results from the image analyses are captured in a .csv file that tabulates the filename, the grain count including and excluding the edge grains ('total grains' and 'center grains', respectively), the percent of the image occupied by DC (%DC), the mean and standard deviation of the grain size in px^2, and the mean and standard deviation of the grain size excluding the largest

https://doi.org/10.52202/072781-0023

grain. The grain counts and the %DC are used to quantitatively determine the average grain size per ASTM 112 as follows:

1. Calculate the overall number of grains:

$$N = \#\ \text{center grains} + \left(\frac{\#\ \text{total grains} - \#\ \text{center grains}}{2}\right) \quad [1]$$

2. Calculate the area of ferrite by correcting the total image area for %DC and scaling:

$$A(mm^2) = \left(1 - \frac{\%DC}{100}\right)(1024\ px * 0.001 mm/\mu m * S)^2 \quad [2]$$

where S is the image scaling in units of μm/px as manually determined from the scale bar.

3. Calculate the number of grains per mm^2 from N and A:

$$N_A(grains/mm^2) = \frac{\text{grain count}}{\text{image area}} = \frac{N}{A} \quad [3]$$

4. Calculate the grain size G from N_A per ASTM E112:[4]

$$G = -2.9542 + 3.3219 \cdot \log_{10}(N_A) \quad [4]$$

The mean and standard deviation of the grain size, with and without the largest grain, reported in units of px^2 are not used in the grain size determination because the segmentation of the ferrite grains is not intended to be accurate from an areal perspective. As described above, the reported grain size is determined from the grain count; however, the grain areas in the segmented images are collected and used in the automated data screening described below. For both analyses, %DC area and grain counting, the results are also visually captured in the form of color masks with the original microstructure overlayed at approximately 40% transparency. These 'QC overlay' images provide a quick and convenient tool for visual qualitative verification of the results. Example overlays from analysis of the segmented images from Figure 5 are shown in Figure 6 for (a) DC and (b) grain size.

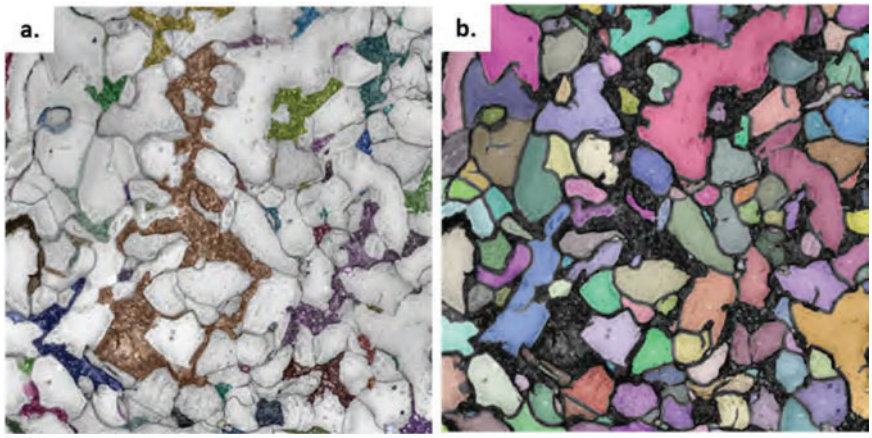

Figure 6. Original microstructures overlayed at 40% transparency on top of color masks showing the analyses of (a) DC, and (b) grain size for the segmented images from Figure 5.

[4] Note that grain size G is a log scale, with smaller values corresponding to larger grains and vice versa. From footnote 1, G values of 5, 10, and 15 correspond to MLI values of 57, 10, and 1.8 μm, respectively.

https://doi.org/10.52202/072781-0023

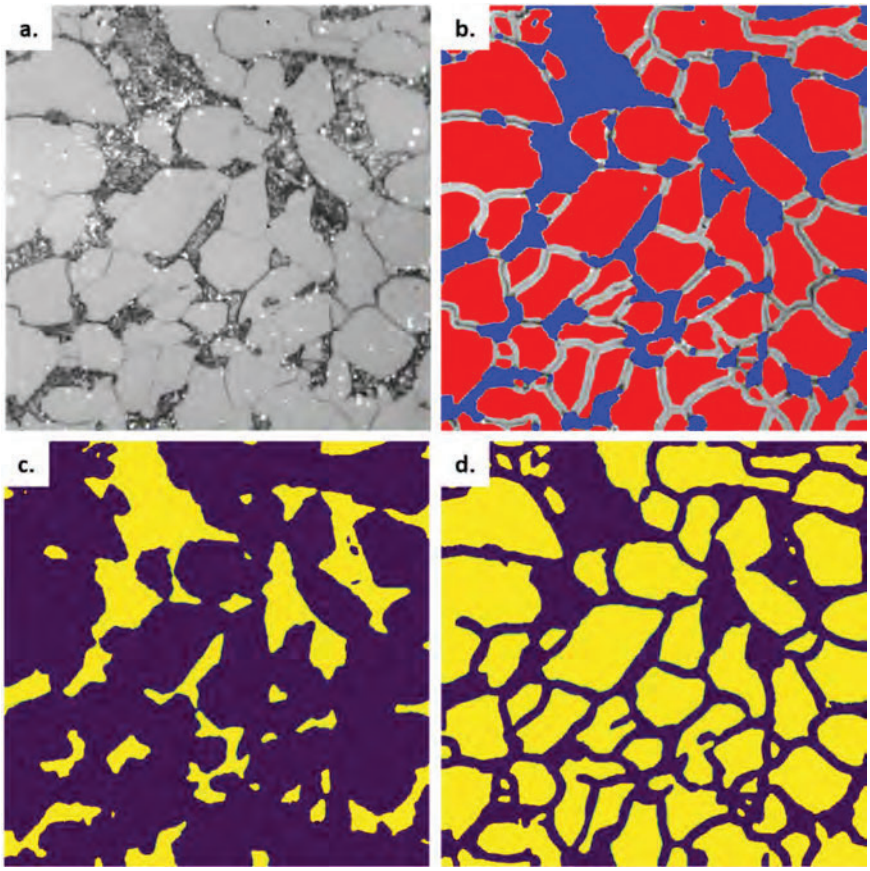

Figure 7. Example images from the ML model training set: (a) raw microstructure, (b) manually segmented microstructure, (c) ML segmented for DC analysis, and (d) ML segmented for grain size analysis.

Results

The automated evaluation consists of two primary components: i) automated segmentation by the ML model, and ii) automated analysis of the segmented images using traditional image processing tools and the calculations described by Eqs. 1 through 4. As a first step towards validation of the combined process, results from ML segmentation of the training and validation images were compared against results from manual segmentation of the same images. After finetuning, the process described in Figure 3 was used to generate segmented images from the raw microstructures used in the training and validation sets, as well as duplicate images rotated 180°. The grain size and %DC were then evaluated by the automated processes described above. For comparison, the same evaluations were applied to the manually segmented training and validation images. An example image-set is provided in Figure 7, which shows (a) a raw microstructure image from the validation set (sample 11), (b) the corresponding manually segmented image, and (c,d) the results from ML segmentation of the raw microstructure. The latter images highlight in yellow (c) the dark constituent (DC) and (d) the ferrite grains.

The quantitative evaluations of %DC from the manual and ML segmented images are compared in Figure 8a. The unity line is indicated in grey and the estimated repeatability of the manual measurement (±20% (relative) for %DC >5%, and 1% (absolute) for %DC ≤5%) is shown in red for reference. The results fall along the unity line and well within the repeatability of the equivalent

https://doi.org/10.52202/072781-0023

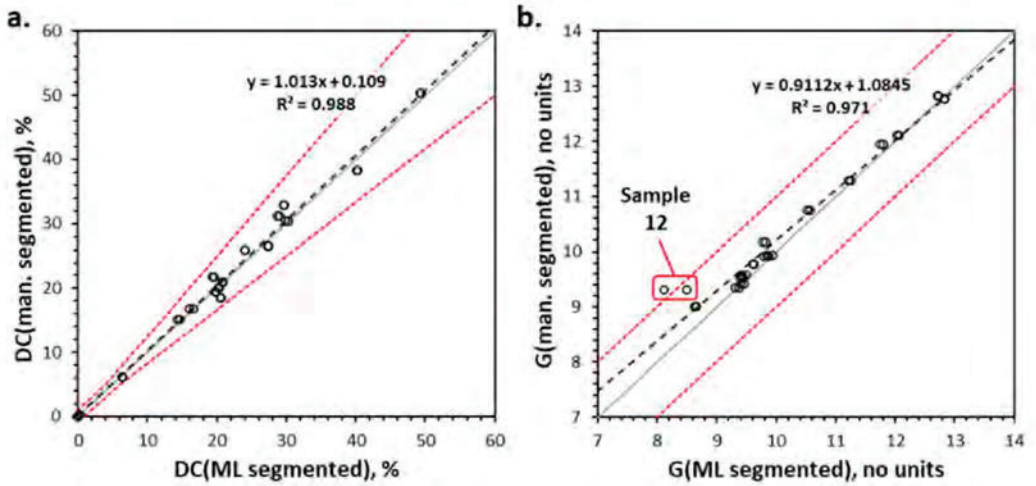

Figure 8. Results of (a) %DC, and (b) grain size from analysis of the training slices by manual versus automated segmentation (512x512 px).

manual measurements. The linear least squared regression (LLSR) has slope and intercept values of 1.013 and 0.109, respectively, and an R^2 of 0.988. These values confirm the visual interpretation that there is a strong linear correlation between the manual and ML results, with minimal scatter (R^2 close to1.0) and negligible systematic error (slope close to 1.0 and intercept close to 0.0). Note that a strong correlation between these results is expected because the ML model was trained using these images. A more direct evaluation of the ML results will be provided below by comparison to quantitative evaluations performed manually. However, the results from these training images are presented here to demonstrate that the ML model accurately reproduces the manual segmentation and, at least under ideal conditions, the automated evaluation process generates consistent results regardless of the source (ML vs. manual) of the segmentation images.

Figure 8b shows the corresponding results for grain size, where the unity line is again indicated in grey and ±1 G is indicated in red. The latter is again provided for reference since it corresponds to the repeatability associated with the manual grain size measurements discussed below [5]. The results suggest good agreement between the analyses, with most of the results falling along the unity line. One notable exception, sample 12 from Table 1, corresponds to an image for which the automated segmentation was inaccurate due to indistinct grain boundaries. The effect of the indistinct grain boundaries, which can occur due to under-etching and/or poor replication quality, is that the raw image lacks sufficient information for the model to resolve clusters of grains as distinct from each other. Consequently, the ML-segmented image combines multiple smaller grains into a single, large field that is subsequently counted as a single grain. From consideration of Eqs. 3 and 4, this undercounting of the grains leads to over-estimation of the grain size which manifests as a lower value of G. This is apparent in Figure 8b, where the grain size for sample 12 is reported to be 9.3 versus 8.1 to 8.5 for the manual and ML segmentation, respectively. Note that the two symbols circled in Figure 8b correspond the original (0°) and rotated (180°) orientations for the image from sample 12. Including the results from sample 12, the overall LLSR has slope and intercept values of 0.911 and 1.085, respectively, and R^2 of 0.971. This again confirms the visual interpretation that the correlation has low scatter and low systematic error. Combined, the results from Figure 8 suggest that, provided sufficient image quality, the ML segmentation accurately reproduces the manual segmentation, and the results from the subsequent analyses are consistent.

https://doi.org/10.52202/072781-0023

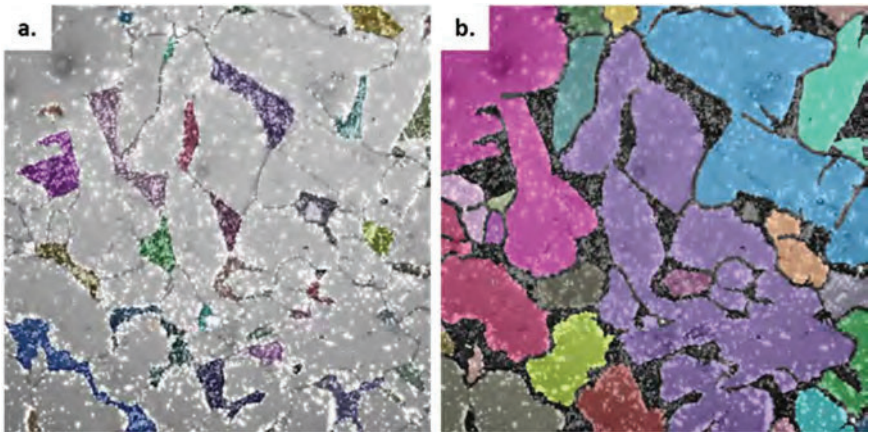

Figure 9. QC overlay images showing the original microstructure for the errant sample 12 feature highlighted in Figure 8 overlayed upon the color-mapped images: (a) DC, and (b) ferrite grains.

The QC overlays from the DC and grain size analyses of the ML-segmented images from sample 12 are shown in Figure 9a and Figure 9b, respectively. In the DC overlay, each distinct colony of DC has been assigned a different color, and the resultant colormap has been overlayed with the original microstructure. Visual inspection of the overlay image suggests accurate alignment between the DC in the original image and the colormap obtained from the segmented image. Similarly, in the grain size overlay each individual grain counted during the analysis has been assigned a different color. A cursory inspection of that image reveals several clusters of grains that have been assigned a single color, which means that the ML model was unable to resolve them as individual grains. This effect, discussed above with respect to Figure 8b, is most apparent as the central purple feature in the image, which covers at least 10 individual grains. This visual inspection constitutes an easy QC for the analysis. While visual QC of the overlay images will identify most issues associated with inaccurate results, it may not be practical to routinely rely on visual QC (i.e., human intervention) if a large number of images are to be processed. Therefore, criteria for automated identification of the most common sources of inaccurate grain size results have been developed.

During development of the evaluation process, it was observed that most cases of under-counting the number of grains (i.e., multiple grains being counted as a single large grain) resulted from indistinct grain boundaries and/or out-of-focus regions in the image. As discussed previously, the net effect of under-counting is to artificially attribute a relatively large area to one, or a very few, anomalously large grain(s) in the microstructure. Therefore, quantitative criteria based on maximum grain size have been implemented to automatically screen for results that should be rejected due to undercounting of the ferrite grains. The criteria evaluate the maximum grain size by determining whether either of the following conditions are met:

1. The largest ferrite grain is more than 10% of the non-DC area (i.e., more than 10% of the total area of the ferrite).
2. Excluding the largest ferrite grain changes the standard deviation of the grain size by more than 30%.

These criteria are evaluated from the mean and standard deviation of the grain size, with and without the largest grain, in units of px^2. Those parameters are, in turn, calculated from pixel-based measurements of the individual grain sizes that are tabulated as an intermediate part of the analysis in ImageJ. As stated above, the pixel-based grain sizes are used solely as screening criteria and are not

https://doi.org/10.52202/072781-0023

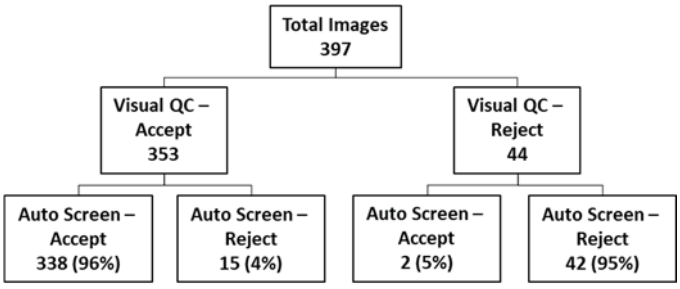

Figure 10. Comparison of Visual QC and Automated Screening from automated evaluations of 397 replica images. The Automated Screening agrees with the Visual QC for at least 95% of the images.

used for the actual grain size determination, which is based on grain count to maintain consistency with ASTM E112. Also, even if the segmentation and analysis properly delineate and count the ferrite grains, if any single grain represents more than 10% of the total area of the ferrite then it is unlikely that there are enough grains in the image to yield a representative grain size analysis. In that case, the image should not have been evaluated and the data should be rejected. No quantitative criteria have been developed for %DC yet, so %DC is assumed to follow grain size (i.e., if the grain size criteria identify a rejection, the data are rejected for both grain size and %DC).

The automated screening criteria were validated by comparison to visual QC of automated analyses from 397 line pipe replica microstructure images. For the automated analyses, the quantitative screening criteria were automatically evaluated during the final grain size calculations. For comparison, the visual QC was performed by briefly reviewing the grain size overlay images from each evaluation and manually assigning a 'reject' to any image with obvious discrepancies between the overlay and the original microstructure. The visual QC was performed by one of the authors 'blind', without any prior knowledge of the automated screening result. The results are illustrated schematically in Figure 10, which indicates that the visual QC accepted 353 evaluations and rejected 44. Within the 353 automated evaluations that were accepted by visual QC, the automated screening accepted 337 and rejected 15. Similarly, within the 44 analyses that were rejected by the visual QC, the automated screening rejected 42 and accepted 2. In both cases the miss rate (visual reject / auto accept or visual accept / auto reject) is less than ≈5%. Therefore, the automated screening described above accurately reproduces the visual QC more than 95% of the time. In addition, a review of the QC overlays for cases where the visual QC and automated screening criteria disagree provided some evidence that the automated process may be more reproducible (i.e., reliable) than the visual QC.

PG&E has selectively implemented a standardized process for manually performing quantitative microstructure evaluations as part of their MVP. During the development of that process, repeated manual evaluations were performed on a set of 42 line pipe microstructures with the goal of assessing evaluator-to-evaluator variation, measurement reproducibility, and process stability. As a result, those 42 microstructures have received an average of more than 10 evaluations per image and can be considered 'well-characterized'. Figure 11 compares the average manual results to the automated evaluations from those images. The unity line is shown in grey, and the repeatability limits for the manual evaluations are shown by the red lines. The automated evaluations of %DC generally provide accurate results when %DC is above approximately 10%, with most of the automated results falling within the repeatability limits of the manual measurements. For %DC below approximately 10%, the automated evaluations tend to over-predict relative to the manual measurements. This effect will be discussed in more detail below, as will the results from the outlier OL-1. For the grain size analysis shown in Figure 11b, the automated and manual evaluations are in good agreement. The screening criteria described above accepted all of the automated results, which are all within the repeatability limits of the manual evaluations. The LLSR analysis confirms a strong correlation (slope near 1.0

https://doi.org/10.52202/072781-0023

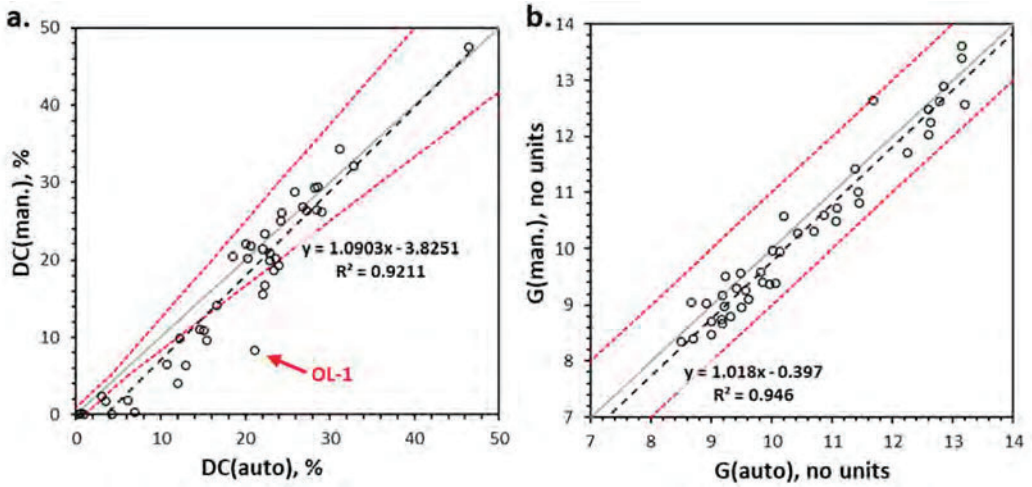

Figure 11. Manual versus automated evaluations of (a) %DC, and (b) grain size (G) for a set of 42 'well-characterized' pipe microstructures. These microstructures received an average of more than 10 manual evaluations per image. No results were rejected by the automated screening criteria.

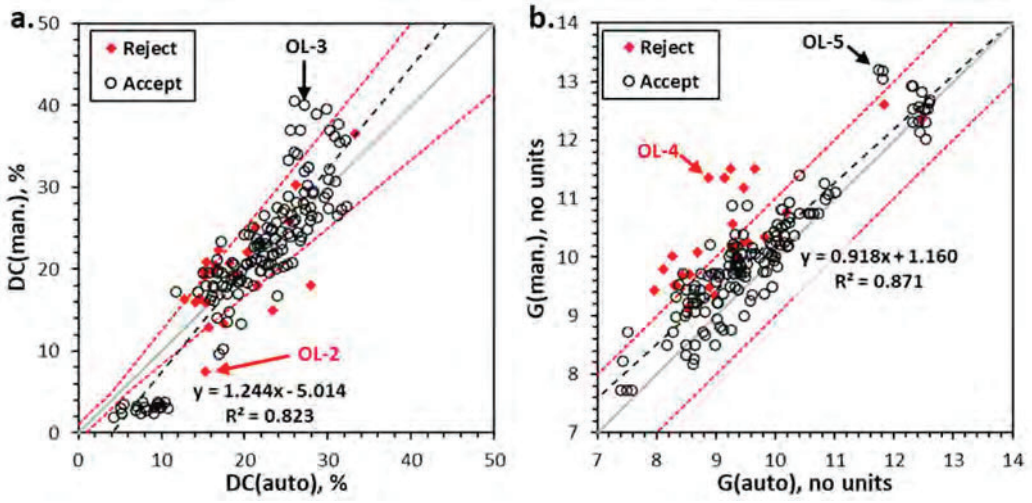

Figure 12. Manual versus automated evaluations of (a) %DC, and (b) grain size (G) for a set of 171 pipe microstructures. These microstructures received an average of fewer than 5 manual evaluations per image. Results rejected by the automated screening criteria are shown in red.

and intercept near 0.0) with low scatter (R^2 close to 1.0). These results suggest that the automated grain size measurements provide accurate results compared to the average of multiple grain size evaluations performed manually per ASTM E112.

Manual evaluations were also available for 171 of the 397 replica images used to develop the screening criteria described above. In these cases, however, the manual evaluations were performed as part of MVP data collection and not as part of the process development. As a result, practical resource constraints limited the number of evaluations to an average of fewer than 5 per image. Figure 12a compares the %DC results from the average of those manual evaluations to the automated results. The results rejected by the maximum grain size screening criteria described above, indicated by the red symbols, are distributed across the range of measurements with most of them falling within

https://doi.org/10.52202/072781-0023

the repeatability limits. This suggests that the grain size based screening criteria do not correlate well with the quality of the %DC result, and the strategy to reject %DC evaluations based on the corresponding grain size evaluation is likely to predominately exclude valid %DC results. Therefore, improved screening criteria for %DC will be evaluated during a future v2.0 development. The remaining (accepted) automated evaluations in Figure 12a are clustered around the unity line over most of the range. For %DC values between approximately 10% and 35%, the automated results tend to fall within the repeatability limits of the manual evaluations; however, when the %DC is less than approximately 10% the automated measurements tend to over-predict relative to the manual results. This is consistent with the behavior reported in Figure 11a, and it is a known behavior of the ML model that occurs in fine-grained, low-%DC microstructures when clusters of fine grains are interpreted as colonies of DC. The behavior will be addressed in a future v2.0 development by expanding the set of training images to include additional relevant fine-grained microstructure images. In addition, there is a tendency for the automated results to under-predict the %DC for values above 35%. These behaviors will be discussed below using the outliers indicated as OL-2 and OL-3 as examples. An LLSR analysis of the 'accepted' results yields slope and intercept values of 1.244 and -5.014, respectively, and an R^2 of 0.823. The deviation of the slope from 1.0, and the offset of the intercept from 0.0, reveal the impact of the over- and under-prediction at low and high percentages, respectively, while the R^2 value quantifies the amount of scatter in the results. The decreased R^2 relative to Figure 11a may indicate some additional variability in the manual measurements as a result of the smaller number of evaluations per image.

Figure 12b compares the corresponding grain size results from the same set of 171 images. The unity line is again shown in grey, the repeatability limits for the manual measurements are indicated by the red lines, and the results rejected by the automated screening criteria are indicated by the red symbols. The data show that the automated evaluations generally estimate grain size within the repeatability limits of the manual evaluations, that most of the values identified as 'rejected' fall outside that range, and conversely that most of the values falling outside that range are identified as 'rejected'. This suggests that the screening criteria appear to be effective for the grain size evaluations. In general, the automated results tend to slightly over-predict the grain size (smaller G, larger MLI) compared to the manual evaluations. This is likely to result from limited under-counting due to localized areas of indistinct grain boundaries in the images. The effect is more pronounced at larger grain sizes, where there are fewer grains in the images, so a few 'missed' grain boundaries have a larger effect on the average grain size. The scatter is fairly significant, with R^2 of 0.87, which probably reflects the range of image qualities in the dataset, but may also indicate some additional variability in the manual measurements as a result of the smaller number of evaluations per image. The outliers designated OL-1 through OL-5 were further evaluated by consideration of the corresponding QC overlay images.

Figure 13 shows several images from the outlier designated OL-1 in Figure 11a. For that outlier, the automated evaluation significantly over-predicted the %DC relative to the manual evaluation (21.1% vs. 8.3%). Figure 13 includes the microstructure image input to the ML model for analysis (a) and the QC overlay image output from the DC analysis (c). Also shown are magnified views of the regions indicated by the yellow squares in the two images (b, d). Figure 13a and Figure 13b show the microstructure image to have relatively high contrast, with a predominance of clearly defined light-grey ferrite grains and several dark, uniform colonies of DC. In addition, however, the image contains numerous instances of intermediate character that are darker than typical for the ferrite but lighter than the dominant DC colonies. Moreover, these features often exhibit internal structure that complicates interpretation for both the ML model and the manual evaluators. Overall, the QC overlay reveals that the ML model accurately segments the DC. The primary performance deficiency appears to be that the model struggles to identify small ferrite grains when they are embedded within

https://doi.org/10.52202/072781-0023

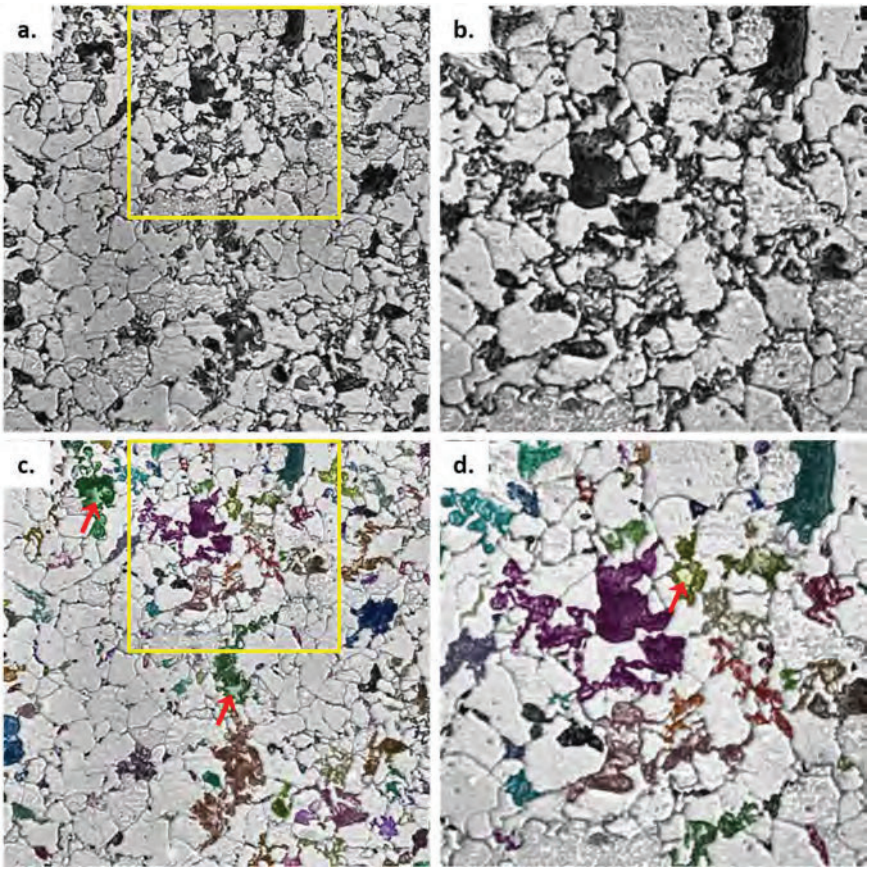

Figure 13. (a) Preprocessed image slice used for the automated evaluation of the outlier indicated as OL-1 in Figure 11a, (b) expanded view showing the highlighted area from (a), (c) DC overlay from the automated analysis, and (d) expanded view of the DC overlay corresponding to the highlighted area in (a) and (c).

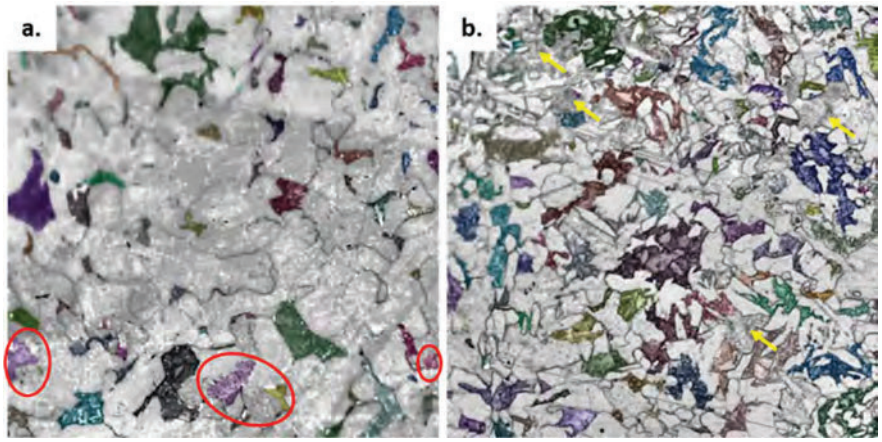

Figure 14. DC overlay images from the automated analyses of the outliers indicated as (a) OL-2 and (b) OL-3 in Figure 12a.

https://doi.org/10.52202/072781-0023

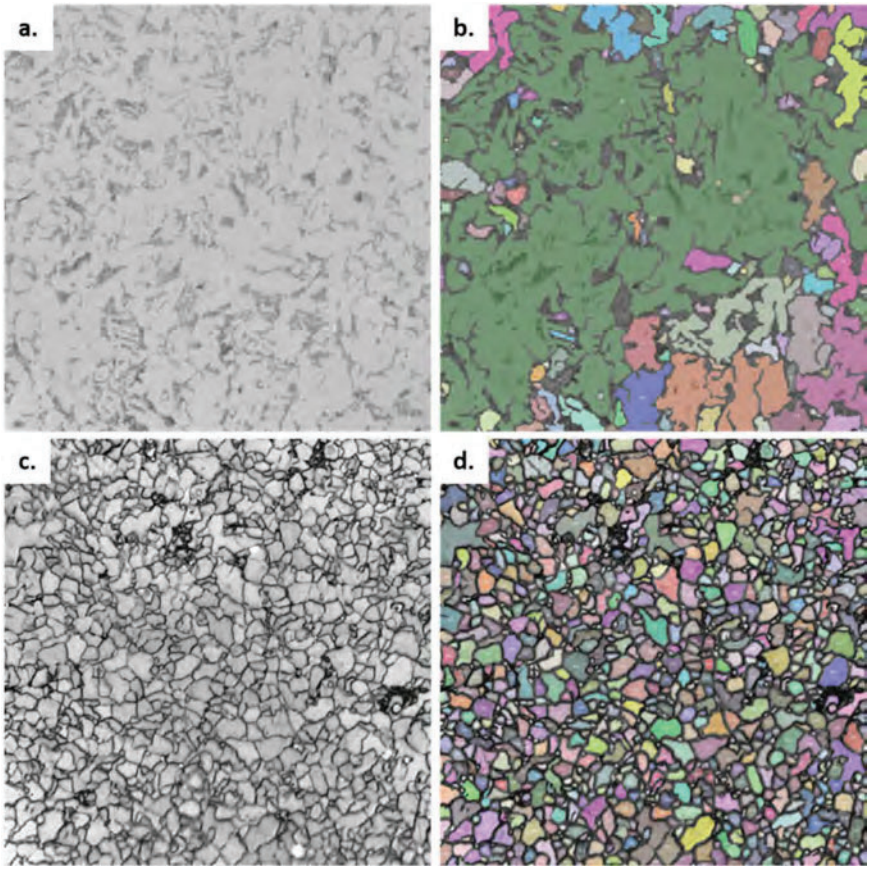

Figure 15. Grain size overlay images from the automated analyses of the outliers indicated as (a,b) OL-4 and (c,d) OL-5 in Figure 12b.

a larger field of DC. A few examples are indicated by the red arrows in Figure 13c and Figure 13d. Additional QC overlays are shown in Figure 14 for the %DC outliers designated OL-2 and OL-3 in Figure 12a. For OL-2, where the ML model over-predicts the %DC, Figure 14a shows the shade (darkness) of the ferrite to be inconsistent across the image. This causes localized areas of ferrite to be designated DC, as indicated by the colored fields circled in red. The poor resolution of the grain boundaries and the out-of-focus condition in the top half of the image caused the automated screening criteria to reject the evaluation, so the inaccurate %DC result is secondary. However, it is presented here in order to improve understanding of the model performance. The overlay image from outlier OL-3, shown in Figure 14b, illustrates the opposite effect when areas of DC appear lighter than expected. In many cases the shading (darkness) is closer to the ferrite than to the darkest DC colonies, and a few examples are indicated by the yellow arrows in the image. As a result of the low-contrast of these DC regions, they are difficult to properly categorize. The conditions leading to the results for OL-2 and OL-3 can probably be addressed by incorporation of a few additional images to the training set and by increasing the use of image augmentation during finetuning to "simulate" low quality or low-contrast images; however, these characteristics suggest low image quality that may be better mitigated by improved surface preparation (polishing and etching).

Finally, Figure 15 shows the initial microstructures (a,c) and QC overlays from the grain size analyses (b,d) corresponding to the outliers designated OL-4 and OL-5 in Figure 12b. For OL-4, which was rejected by the automated screening criteria, the initial microstructure shows poor contrast and almost no visible grain boundaries, Figure 15a. As a result, the entire central region of the image is

https://doi.org/10.52202/072781-0023

'counted' as a single grain in the overlay image, Figure 15b, and the automated evaluation over-predicts the average grain size (recall that smaller G corresponds to larger grains). In contrast, the images from OL-5, Figure 15c and Figure 15d, show well-defined, high-contrast grain boundaries and accurate segmentation. In both cases, OL-4 and OL-5, the visual QC confirms the outcome of the automated screening.

Summary and Conclusions

This work describes a process for automated evaluation of replica microstructures obtained from steel line pipe. While the evaluation is performed in several steps, each step requires minimal touch time by the operator. This significantly reduces the time required for evaluation relative to traditional manual methods, particularly when evaluation is performed on a large number of images. At the core of the process are two operations: i) segmentation of the microstructure to delineate the ferrite grains and the DC colonies, and ii) analysis of the segmented images to determine the count of ferrite grains and the percentage of the image (by area) occupied by DC. Those values are subsequently used to calculate the average grain size by a series of straightforward calculations.

The segmentation process is performed in Python using an ML model that was pre-trained on a database of over 10^5previously segmented microscopy images. Finetuning the model was performed on a small set of application-specific images comprising 17 manually segmented images of different line pipe microstructures. The segmented images are subsequently analyzed in ImageJ to determine the %DC and the number of ferrite grains. In addition, the ImageJ analysis outputs additional 'QC overlay' images in which the original microstructure is overlayed upon colormaps of the distinct ferrite grains or DC colonies. Finally, a pair of quantitative criteria were developed to enable automated screening of inaccurate grain size analyses based on the observation that inaccurate results were most often caused by undercounting grains due to poor image quality.

Results from the training images indicated that the automated segmentation generally reproduces the manual segmentation. Validation of the screening criteria performed by analysis of 397 microstructure images revealed that the automated grain size screening criteria duplicate the results from a visual QC more than 95% of the time. For 171 images with corresponding manual evaluations, the automated results tend to over-predict grain size slightly relative to, but are generally within the uncertainty limits of, the manual measurements. Similar comparison of the manual and automated results for %DC shows that the automated evaluation tends to over-predict when the %DC is below approximately 10%, but is within the uncertainty limits of the manual measurements over most of the relevant range of %DC (10% to 35%).

Future Work

Future development of a v2.0 of the automated evaluation process will address known deficiencies in performance and functionality. These include the overprediction at low %DC and the imperfect delineation of low-contrast colonies of DC. In addition, screening criteria will be developed for %DC to allow automated identification of unreliable results, and training for TMCP and quenched and tempered microstructures will be improved. Finally, the user interface will be streamlined so that all the functionality can be implemented from a single call in Python, rather than requiring the user to switch between ImageJ and Python.

https://doi.org/10.52202/072781-0023

References

[1] M Gould, B Amend, P Veloo, O. Oneal, R. Gonzalez, and N Switzner (2021) "The Role of In-Situ Metallography in Pipeline Integrity Management," PPIM 2021, Proceedings of The 33rd International Pipeline Pigging and Integrity Management Conference.
[2] N Switzner, P Veloo, M Rosenfeld, T Rovella, and J Gibbs (2020) "An approach to establishing manufacturing process and vintage of line pipe using in-situ nondestructive examination and historical manufacturing data," IPC2020 Proceedings of The International Pipeline Conference, American Society of Mechanical Engineers.
[3] J Gibbs, J Kornuta, O. Oneal1, P Veloo, and N Switzner (2022) "Using Nondestructive Testing and Statistical Analysis to Perform Material Property Verification and Align Records," PPIM 2022, Proceedings of The 34th International Pipeline Pigging and Integrity Management Conference.
[4] N Switzner, S Thorsson, J Kornuta, P Veloo, P Martin, T Rovella, and M Rosenfeld (2020) "Influence of line pipe steel microstructure on NDE yield strength predictive capabilities," PPIM 2020, The 32nd International Pipeline Pigging and Integrity Management Conference (PPIM), Houston, TX, USA (February 2020).
[5] LP Martin, NT Switzner, O Oneal, S Curiel, J Anderson, and P Veloo (2022) "Quantitative Evaluation of Microstructure to Support Verification of Material Properties in Line-Pipe Steels," IPC2022, Proceedings of the ASME 2022 14th International Pipeline Conference, September 26-30, Calgary, Alberta, CA.
[6] N Switzner, J Anderson, M Rosenfeld, J Gibbs, P Veloo, and R Gonzalez (2021) "Assessing steel pipe toughness using chemical composition and microstructure," PPIM 2021, Pipeline Pigging and Integrity Management Conference.
[7] B Amend, M Gould, P Veloo, O Oneal, R Gonzales, and N Switzner (2021) "In Situ applications in the Pipeline Industry," Materials Evaluation, 8, 791-796.
[8] J Stuckner, B Harder, and TM Smith (2022) "Microstructure segmentation with deep learning encoders pre-trained on a large microscopy dataset," NPJ Computational Materials, 8(1), 200.
[9] ASTM E3-11, "Standard Guide for Preparation of Metallographic Specimens," ASTM International, West Conshohocken, PA (2011).
[10] ASTM E1351-01 (Reapproved 2012), "Standard Practice for Production and Evaluation of Field Metallographic Replicas," ASTM International, West Conshohocken, PA (2012).
[11] ASTM E112, "Standard Test Methods for Determining Average Grain Size," ASTM International, West Conshohocken, PA (2013).
[12] ASTM E562-19, "Standard Test Method for Determining Volume Fraction by Systematic Manual Point Count," ASTM International, West Conshohocken, PA (2019).
[13] Stuckner, J. (n.d.). Nasa / Pretrained-Microscopy-Models. https://github.com/nasa/pretrained-microscopy-models. Accessed November 4, 2022.

https://doi.org/10.52202/072781-0023