doi:10.1017/S1431927622011370
Microsc. Microanal. 28 (Suppl 1), 2022

Synthetic Data for Machine Learning and Novel Edge Detection to Measure Particle Size Distributions in TEM

Eoin Walsh[1] and Andy Stewart[1,2*]

[1.] Department of Physics, University of Limerick, Limerick, Ireland.
[2.] Department of Chemistry, University College London, London, UK.

* Corresponding author: andy.stewart@ucl.ac.uk

Machine Learning (ML) has seen significant performance improvements, due to breakthroughs in algorithm design, such as Deep Learning [1], improvements in computing power, and an abundance of data to train ML algorithms. One popular subcategory of ML is Supervised ML [2]. The premise of Supervised ML is that an algorithm is trained using a pre-labelled dataset. The trained algorithm is then be used to make inferences about data outside of the training dataset.

This has proved to be a successful application of ML for inference, but it has some limitations. For example, the data undergoes manual labelling before algorithm training. This is a time-consuming task. Another problem is a dearth of labelled data with which to train a neural network. ImageNet [3] is a benchmark dataset used for testing ML algorithms in the computer vision community. It consists of 3.2 million images from 5247 labelling categories. No benchmark datasets of similar size and extent exist in microscopy. Most datasets consist of 10s to 100s of images, with no labelling of what is in these images as part of their metadata, which would be useful for ML image classification applications. ML segmentation algorithms require even more sophisticated labelling than this. Segmentation algorithms require per pixel labelling for each image in the training dataset.

One potential solution to a dearth of such data is to make use of synthetic data. In Transmission Electron Microscopy (TEM), this is generated data, which closely resembles data outputs from a TEM. Synthetic data has the advantage of being automatically segmented and labelled during data generation. The generated dataset is used to train a ML algorithm, which can segment particles of interest in experimental TEM images. Synthetic data to train ML algorithms has already seen success when applied to autonomous driving and the segmentation of nuclei in cells [5, 10].

In this work, we demonstrate how a synthetic image generator can be used to train ML algorithms that detect particles in TEM produced images. The TEM images consist of silica particles on a lacey-carbon grid substrate, with an ultrathin layer of amorphous carbon beneath. We also demonstrate the measurement of the particle's dimensions post detection, using a bespoke edge detection method, and compare these measurements to those carried out manually by a domain expert.

The ML algorithms detailed in this work were trained using synthetic data alone. A synthetic data generator was developed to produce images that resemble features in the TEM data, including variations in magnification, brightness, and texture. The synthetic data generator creates images using a series of image manipulation and shape generation techniques, such as the generation of Voronoi tessellations to mimic the lacey-carbon grid, and the use of a Sobel filter to recreate the lacey-carbon grid edge effects. These techniques were carried out using common python packages, such as OpenCV [6], SciPy [7] and NumPy [8]. Figure 1 shows an example of a synthetic TEM image.

https://doi.org/10.1017/S1431927622011370 Published online by Cambridge University Press

Using this synthetic data generator, a user defined number of images and labelling masks can be generated. To account for the range of magnifications in the TEM data, three separate algorithms were trained, each one trained with synthetic images for a particular magnification range. The U-Net ML architecture [9] was used for each algorithm and each one was trained on 1,000 training images and 300 validation images from the synthetic generator.

Once the algorithms had been trained successfully, analysis of the TEM data was carried out by passing the data through one of the three ML algorithms. The TEM micrograph dataset consists of 32 images at 2k resolution, with 4 different magnifications present. Each image passes to one of the three ML algorithms depending on the image's magnification. The ML particle prediction for each image is then passed through a filtering process to remove any erroneous features in each prediction. 102 particles out of 115 visible in the images were detected by the ML algorithms, an 88.7% detection rate. Figure 1 shows an example of a TEM image, with two particles having been detected.

The detected particles are then measured using a bespoke edge detection technique. Ninety cross-sections of each particle are taken at evenly spaced degrees of rotation azimuthally. Each cross-section is then analyzed to estimate the particle's edges. The cross-section is pre-processed using an average kernel and by finding the derivative between consecutive points. The perimeter of each particle is found by getting the best fit circumference using all the edge detections for a particle. This fitting is done using a statistical technique known as Functional Data Analysis (FDA). The best fit bounding box dimensions for this edge estimate provides the data for the major and minor axis lengths of each particle. Figure 2 shows edge predictions for two particles.

To gauge the accuracy of the major and minor axis measurements, the particles were manually measured too and so a comparison could be made. The manual measurements for the 115 particles gave an average major axis length of 182.96nm ± 26.97nm and an average minor axis length of 170.13nm ± 27.42nm. The combined particle detection and measurement program gave an average major axis length of 186.93nm ± 23.41nm and an average minor axis measurement of 173.06nm ± 19.13nm. The automated measurements are within the error margin of the manual measurements, demonstrating the preciseness of the automatic edge detection method.

We have shown how a synthetic image generator combined with supervised ML can be used to detect and segment particles of interest in TEM micrographs. The measurements obtained from this method compare with those carried out manually by a human expert [11].

https://doi.org/10.1017/S1431927622011370 Published online by Cambridge University Press

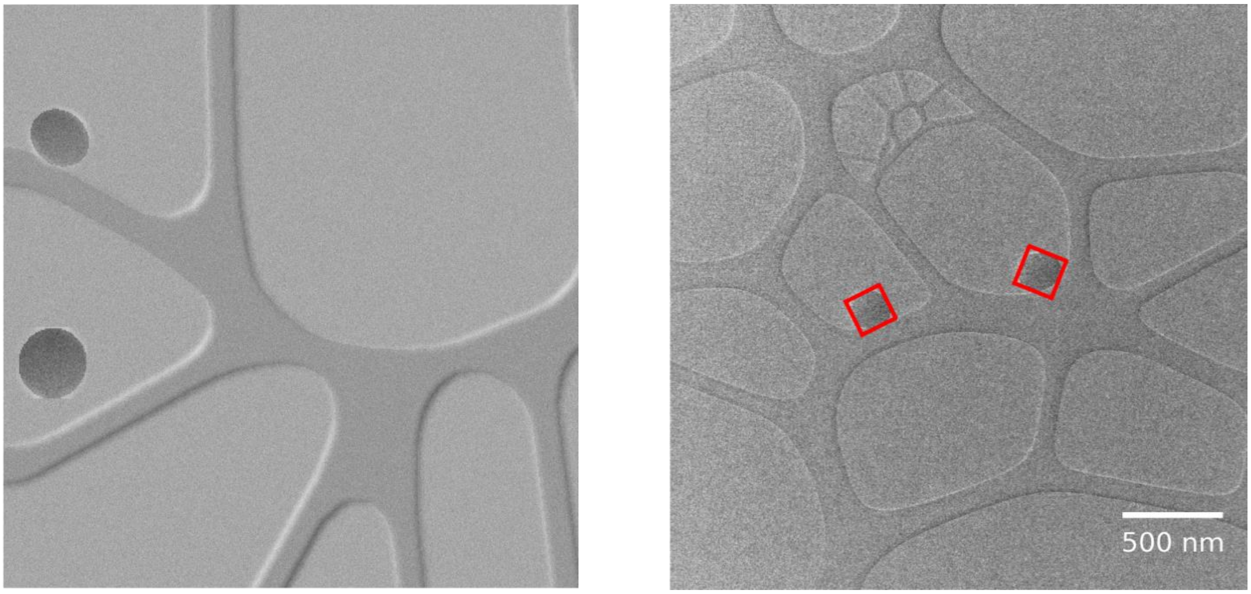

Figure 1. An example of a synthetically generated TEM image (left) and a TEM micrograph with two particles detected using ML within it.

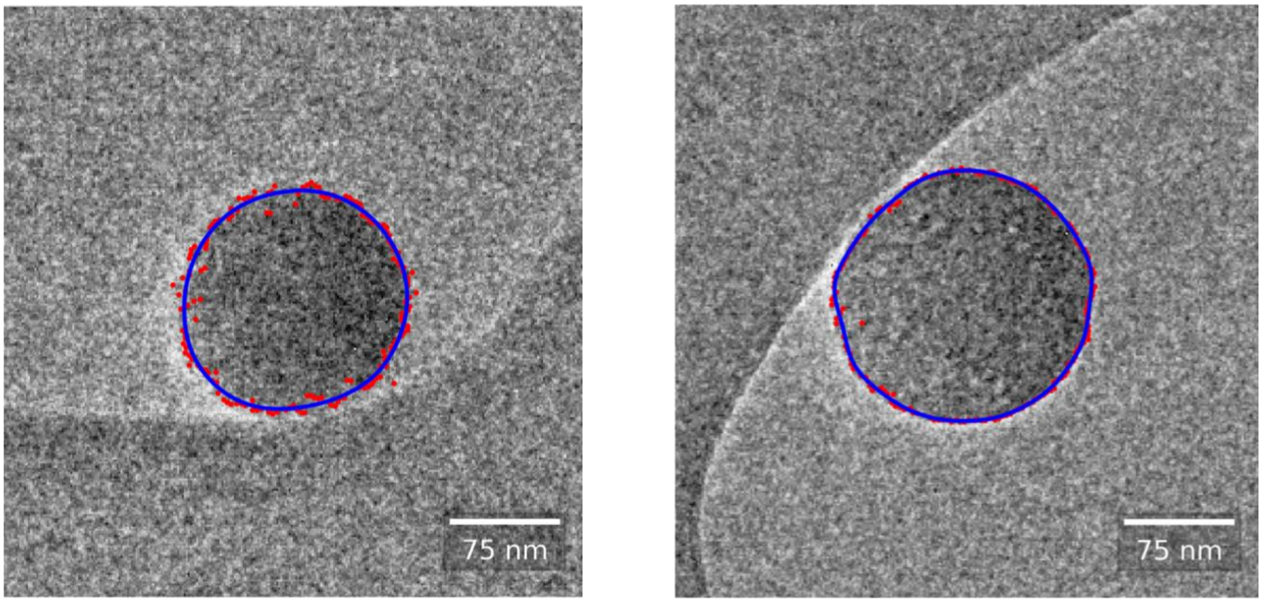

Figure 2. The results of the edge detection program for two separate particles. The red data points are the particle edge points detected for each cross-sectional particle profile; the blue line is the best-fit line for these particle edge detections.

References:

[1] A Krizhevsky, I Sutskever and GE Hinton, Advances in Neural Information Processing Systems Journal **25** (2012), p. 84. doi:10.1145/3065386

https://doi.org/10.1017/S1431927622011370 Published online by Cambridge University Press

[2] A Singh et al., 3rd International Conference on Computing for Sustainable Global Development (2016), p. 1310.
[3] J Deng et al., IEEE conference on computer vision and pattern recognition (2009), p. 248.
[4] Y Lecun et al., Proceedings of the IEEE **86**(11) (1998), p. 2278. doi:10.1109/5.726791.
[5] B. Osiński et al., IEEE International Conference on Robotics and Automation (2020), p. 6411.
[6] G. Bradski, Dr. Dobb's Journal of Software Tools (2000).
[7] P Virtanen et al., Nature Methods **17** (2020), p. 261, doi:10.1038/s41592-019-0686-2
[8] CR Harris et al., Nature **585** (2020), p. 357, doi:10.1038/s41586-020-2649-2
[9] O Ronneberger et al., Medical Image Computing and Computer-Assisted Intervention (2015), p. 234.
[10] KW Dunn et al., Scientific Reports **9** (2019), doi:10.1038/s41598-019-54244-5

[11] This abstract has emanated from research conducted with the financial support of Science Foundation Ireland under grant no. 18/CRT/6049.

https://doi.org/10.1017/S1431927622011370 Published online by Cambridge University Press