Chemometrics and Intelligent Laboratory Systems 225 (2022) 104556

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Characterization of uncertainties and model generalizability for convolutional neural network predictions of uranium ore concentrate morphology

Cody A. Nizinski [a,b], Cuong Ly [b,c], Clement Vachet [c], Alex Hagen [b], Tolga Tasdizen [c], Luther W. McDonald IV [a,*]

[a] University of Utah Department of Civil and Environmental Engineering, Nuclear Engineering Program, 201 President's Circle, Salt Lake City, UT, 84112, United States
[b] Pacific Northwest National Laboratory, Richland, WA, 99352, United States
[c] Scientific Computing and Imaging Institute, 72 South Central Campus Drive, Room 3750, Salt Lake City, UT, 84112, United States

ARTICLE INFO

Keywords:
Uranium chemistry
Nuclear forensics
Machine learning
Image classification
Uncertainty quantification

ABSTRACT

As the capabilities of convolutional neural networks (CNNs) for image classification tasks have advanced, interest in applying deep learning techniques for determining the natural and anthropogenic origins of uranium ore concentrates (UOCs) and other unknown nuclear materials by their surface morphology characteristics has grown. But before CNNs can join the nuclear forensics toolbox along more traditional analytical techniques – such as scanning electron microscopy (SEM), X-ray diffractometry, mass spectrometry, radiation counting, and any number of spectroscopic methods – a deeper understanding of "black box" image classification will be required. This paper explores uncertainty quantification for convolutional neural networks and their ability to generalize to out-of-distribution (OOD) image data sets. For prediction uncertainty, Monte Carlo (MC) dropout and random image crops as variational inference techniques are implemented and characterized. Convolutional neural networks and classifiers using image features from unsupervised vector-quantized variational autoencoders (VQ-VAE) are trained using SEM images of pure, unaged, unmixed uranium ore concentrates considered "unperturbed." OOD data sets are developed containing perturbations from the training data with respect to the chemical and physical properties of the UOCs or data collection parameters; predictions made on the perturbation sets identify where significant shortcomings exist in the current training data and techniques used to develop models for classifying uranium process history, and provides valuable insights into how datasets and classification models can be improved for better generalizability to out-of-distribution examples.

1. Introduction

Convolutional neural networks (CNNs) are deep machine learning architectures that have demonstrated incredible performance on tasks involving images, such as classification. While many more detailed descriptions of convolutional neural networks and their deep learning can be found in the literature, in brief, supervised learning for CNNs involves passing a batch of labeled image examples through the network, computing how off the predictions were from the expected labels with a loss function, updating the convolutional kernel and fully-connected (FC) layer weights with the backpropagation algorithm, and then passing successive image batches until the network loss is minimized [1–3]. When making predictions, an image is passed through the convolutional and pooling layers which extract image features, next the fully-connected layers correlate extracted features to labels, and then an activation function computes a score for each image class [2]. For multi-class problems, a normalized exponential function known as SoftMax relates the FC layer's outputs (often called logits) to scores for each class that are normalized to sum to one [4]. In addition to the standard convolutional, pooling, and activation layers, residual neural network (ResNet) architectures utilize *skip connections* which retain the outputs of previous layers, allowing for better classification performance with deeper CNNs than previously possible [5].

Deep learning models can also be developed to learn without accessing all or any of the class ground truth labels in what is known as semi-supervised learning and unsupervised learning, respectively. In

* Corresponding author.
E-mail address: luther.mcdonald@utah.edu (L.W. McDonald).

https://doi.org/10.1016/j.chemolab.2022.104556
Received 13 December 2021; Received in revised form 22 March 2022; Accepted 30 March 2022

unsupervised learning for images with autoencoders, two sets of convolutional neural networks are used to first compress the information from images into a smaller dimensionality (encoding) and then reproduce the original image from this representation (decoding). While many varieties of autoencoders have been developed, learning of image representations is facilitated by implementing a loss function between the image reconstructions to the original images to iteratively update the encoder and decoder weights to produce better reconstructions [6–8]. Encoded image representations are often useful for other downstream tasks, such as clustering or training classification models.

The vector quantized variational autoencoder (VQ-VAE) learns discrete representations from the input data by replacing the encoder outputs with its most similar vector in a codebook; Oord et al. (2017) claims these discrete representations can produce better latent space features than continuous variational autoencoders, whose decoders can become autoregressive and bypass the need for good latent representations to reconstruct data [8]. A later VQ-VAE adaptation used a hierarchical encoder/decoder architecture to improve latent space representations at multiple image scales [9]. After training the autoencoder, image features can be extracted by passing the image through the encoders, quantizing features with the codebook vectors, and then essentially performing a histogram operation by counting the number of occurrences for each codebook vector in an image [10]. The ResNet-34 and VQ-VAE-2 architectures respectively serve as the starting networks for supervised and unsupervised learning in this paper.

One of the primary goals of machine learning algorithms is generalization, or the ability to make good predictions on unseen examples. Past advances in image classification include improvements to CNN architectures, regularization methods to prevent over-fitting, optimization functions for updating weights, and training practices to achieve state-of-the-art generalization and classification accuracies. Other recent work has sought to gain a deep understanding of what are often considered "black box" algorithms; such efforts have focused on fairness by eliminating biases, robustness to adversarial attacks, explainable artificial intelligence (XAI), uncertainty quantification (UQ), and the ability to identify out-of-distribution (OOD) examples [11–20]. This paper explores uncertainty quantification and predictions on OOD data in the context of nuclear forensics analysis using the surface morphology of uranium ore concentrates (UOCs) from a collection of scanning electron microscopy (SEM) images.

When considering uncertainty quantification for classification models, one must first consider the sources of the uncertainty. *Epistemic uncertainty* comes from a lack of knowledge in a model's parameters, such as a machine learning classifier's weights after training on an incomplete dataset [13,21]. On the other hand, the source of *aleatoric uncertainty* is the variance naturally present in the observed data, such as differences seen in the surface morphology across several micrographs belonging to the same class or even within a single micrograph [13,21]. Several methods for UQ in deep neural networks have been developed to help determine when to trust a set of predictions, and to improve the interoperability of the results. These methods include, but are not limited to, deep ensembles, Gaussian processes, Bayesian neural networks, and Monte Carlo (MC) dropout [22–27]. Among the most popular prediction uncertainty method is MC dropout, which uses variational inference and requires few modifications to existing state-of-the-art neural network architectures or training procedures [26]. Dropout is a well-known regularization technique that randomly "drops" neurons in the hidden layers of a neural network during training, which can prevent the network from relying too much on a single data feature [28]. MC dropout works by continuing to leave out connections during inference while making several predictions on a single test example, resulting in a distribution of SoftMax scores for each class; the statistical variance of the SoftMax scores for each class can be considered the "per-class" uncertainty [26].

On a "per-image" basis, uncertainty can be measured by the information entropy for the set of predictions [29]. Shannon information entropy *H(X)* is calculated by taking the negative sum of each class's SoftMax score $P(x_i)$ multiplied by the logarithm of that score for a set of n class predictions X (Eq. (1)). When using the base-2 logarithm, information entropy takes units of *bits* to quantify how much information is needed to communicate a set of predictions. A more surprising prediction has a higher information entropy and indicates a higher degree of uncertainty. This metric for prediction uncertainty can be applied to a single prediction with no statistical variance, as is done without MC dropout or ensembles, which makes Shannon information entropy particularly useful for comparing different inference methods.

$$H(X) = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i) \tag{1}$$

During nuclear forensics investigations, a variety of analytical techniques might be utilized to gather information about the anthropogenic origins of an unknown nuclear material that was illicitly trafficked or smuggled.[30] Such analytical techniques might include scanning electron microscopy (SEM) to observe microstructure, powder X-ray diffractometry (P-XRD) to determine crystal structure, inductively couple mass spectrometry (ICP-MS) to measure elemental composition and trace-level impurities, and alpha-particle spectrometry or gamma-ray spectroscopy to analyze the isotopic composition of the nuclear material. [30–33] The information gathered from these techniques, and others, can be combined to produce "signatures" or "fingerprints" that indicate how, when, and where the material was produced.[30]

Surface morphology analysis with SEM micrographs has shown the potential to be a powerful signature in the nuclear forensics toolbox [34–37]. Materials with the same crystal structures, as determined by P-XRD analysis, can have significantly different microstructures if synthesized by different processes. Fig. 1 shows UOCs from five common precipitation routes – ammonium diuranate (ADU), ammonium uranyl carbonate (AUC), magnesium diuranate (MDU), sodium diuranate (SDU), and uranyl peroxide (UO_4) – that have been converted to UO_3, U_3O_8, and UO_2. Each synthesis route has distinct particle morphology size and shape characteristics that could be used as a signature in determining their process history; detailed morphology descriptions for these materials can be found elsewhere in the published literature [38].

Other factors outside of the precipitation reaction can also affect particle microstructure of uranium ore concentrates. Such factors include aging, the presence of elemental impurities, physical mixtures of UOC powders, calcination conditions, starting material, and the uranium solution from which the UOC was precipitated [39–46]. Many previously published UOC morphology studies have trained convolutional neural networks to supplement qualitative descriptions and quantitative particle segmentation morphology measurements. While often achieving high classification accuracies, the scope of each classifier was narrow, often only using images from a single precipitation route and oxidation state in the training and testing data sets. When multiple precipitation routes and oxidation states were considered by Ly et al. (2020), the materials were pure and unaged; images were collected with standardized SEM parameters [47]. Girard et al. (2021) implemented unsupervised representation learning with a VQ-VAE for feature extraction, paired with supervised models – including a multi-level perceptron (MLP) – that were trained using the VQ-VAE codebook histograms for various downstream tasks, such as synthetic route classification, calcination temperature classification, and SEM acquisition magnification [10]. This VQ-VAE + MLP implementation and fully-supervised convolutional neural networks achieved comparable accuracies on 12-way synthetic route classification of uranium oxides produced from different precipitates [10,47].

With systems as complex as those influencing the surface morphology characteristics of nuclear materials, it is imperative that any classifier trained should be able to generalize to out-of-distribution data sets. Additionally, SEM data collection parameters are known to affect the appearance of the image; biases from micrograph acquisition could lead to so-called *texture-shape cue conflicts* when using a CNN to make predictions [48]. By evaluating the classification results on OOD image sets,

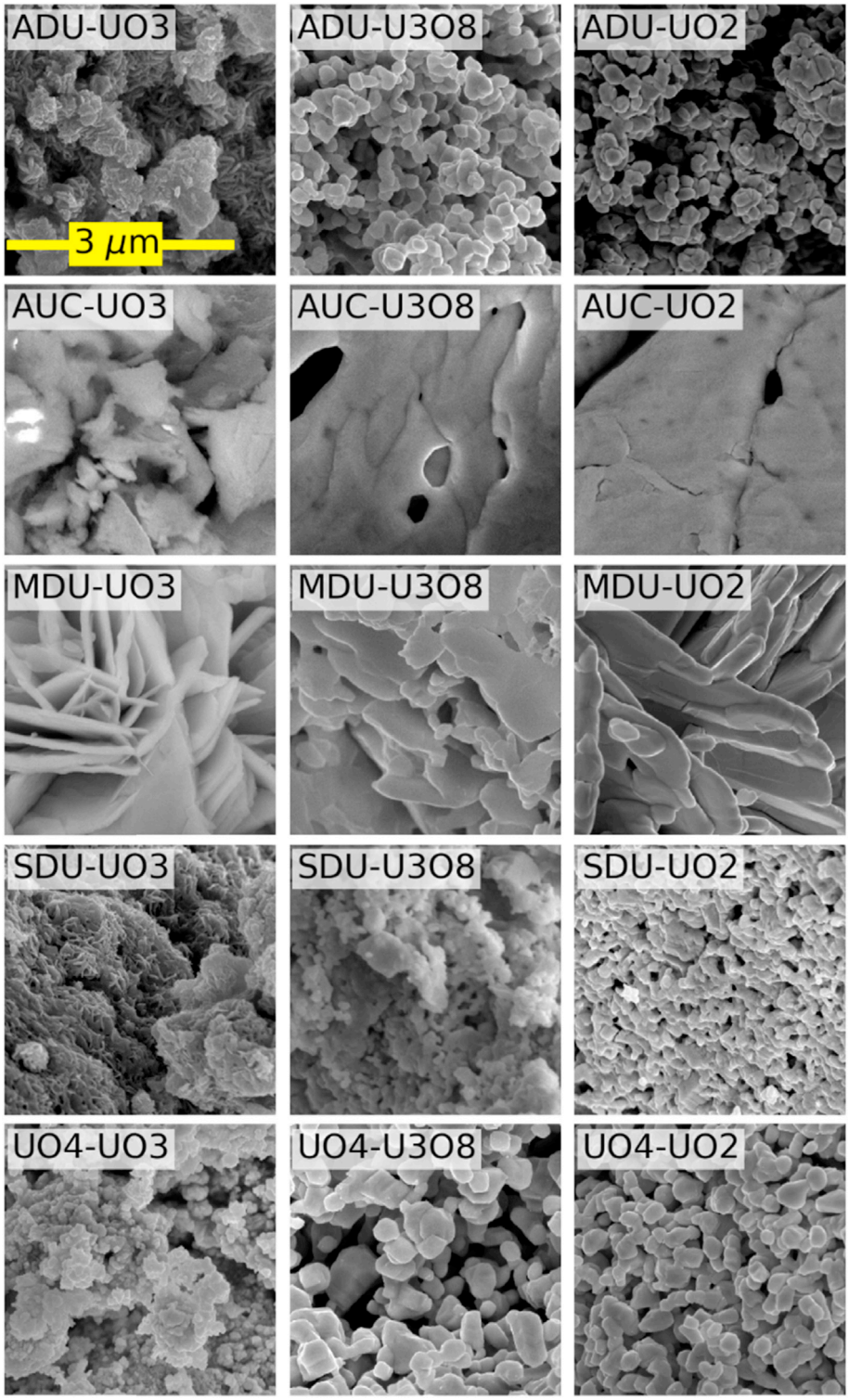

Fig. 1. Characteristic scanning electron micrographs for UO_3, U_3O_8, and UO_2 from the ammonium diuranate (ADU), ammonium uranyl carbonate (AUC), magnesium diuranate (MDU), sodium diuranate (SDU), and uranyl peroxide (UO_4) precipitation routes. The scale bar pictured applies for all micrographs.

the most problematic material and data collection perturbations can be identified, and specific strategies for improving CNNs for nuclear forensic tasks can be developed. Additionally, characterizing the uncertainty quantified for CNN predictions of UOC images will help identify which predictions should or should not be trusted by nuclear forensics investigators.

2. Methods & materials

The ResNet-34 residual neural network architecture starting with weights pretrained on the ImageNet dataset was used for each experiment [3,5]. The final layer of the network was replaced with a layer consisting of global max pooling (GMP) and global average pooling

(GAP) concatenated, followed by a dropout regularization layer, a fully-connected layer with 1000 units, and an output layer with the number of units corresponding to the number of classes for that model. The output layer used the SoftMax activation function for predicting labels.

Hyperparameter tuning using 5-fold cross validation (CV) was used to identify the optimal settings for training the model. Two training phases were used while learning the model. For the first 25 training epochs (passes through all the training data) only the weights of the fully connected layers were updated, with the pre-trained ResNet-34 wt frozen. All weights were updated for the final 50 epochs. The adaptive momentum (Adam) optimizer with decaying learning rates was used [49]. After hyperparameter tuning, initial learning rates of 0.0002 and 0.0001 were set for the first and second training phases, respectively. Image data augmentation during training included flipping the image across its horizonal and vertical axes and random crops to 224 × 224 pixels with 3 color channels, which corresponds to the default input size of the ResNet-34 architecture [5]. The deep learning models were implemented using the Keras API (version 2.3.1) with the TensorFlow backend (version 2.1.0). Training and inference were executed using a single NVIDIA RTX 2060 (6 GB) GPU.

The VQ-VAE was trained by tuning the learning rate and commitment loss term hyperparameters; the final model was trained with a commitment loss of 1.0 and an Adam optimizer (learning rate of 0.001) for 25 epochs. The VQ-VAE latent count and size were 128 and 256, respectively. The unsupervised model was trained with the same augmentation policy, though crop sizes of 512 × 512 pixels were used with grayscale images. To relate the VQ-VAE features to image classes, a 16-way MLP classifier consisting of 5 hidden layers with 1024 nodes each was trained using normalized codebook histograms extracted from random crops of training set images by the trained VQ-VAE. 5-fold CV was used to determine the best learning rate (Adam, 0.0005) and number of epochs (200) for the supervised MLP. The PyTorch machine learning API was used to implement the VQ-VAE model and its supervised neural network classifier. Training and inference were carried out on a single NVIDIA Tesla V100 Tensor Core GPU (32 GB).

2.1. Uncertainty estimates for CNNs

Machine learning classifiers, whose respective training datasets are described in Table 1, were used to evaluate model uncertainty and generalizability. The SEM micrographs corresponding to each of these models were collected with a FEI Nova NanoSEM 630 scanning electron microscope at a horizontal field width (HFW) of 6.13-μm and a resolution of 1024 × 943 pixels, meaning the 224 × 224-pixel training crops had a width of 1.34-μm. These images were collected of pure, unaged, and unmixed uranium materials belonging to 5 synthetic routes (ADU, AUC, MDU, SDU, and UO_4) and 3 calcination products (UO_3, U_3O_8, and UO_2). The synthesis and data collection has previously been described in the published literature [38,47,50].

The 5-class U_3O_8 model consisted only of SEM images taken of a single oxide (U_3O_8) and had 5 labels corresponding to the precipitation route used to synthesize these samples. The 16-class model treated each combination of synthetic routes and calcination products as individual labels, i.e., ADU-UO_3 and ADU-UO_2 each had their own image class. As per Ly et al. (2020) the ammonium uranyl carbonate UO_2 materials were divided into direct (AUCd-UO_2) and indirect (AUCi-UO_2) classes [47]. The image data for each model was partitioned using a stratified split with 80% of the total images belonging to a training subset and 20% belonging to the holdout (test) subset. The 5-fold CV accuracy was 0.962 ± 0.020 for the 5-class U_3O_8 model, 0.885 ± 0.019 for the 16-class CNN, and 0.870 ± 0.024 for the 16-class MLP trained on VQ-VAE codebook histograms. Training accuracy and loss curves can be found in the Supplemental Information.

To characterize the neural network and image uncertainties, the trained 5-class U_3O_8 model was used to make predictions on its holdout set using different inference methods (Table 2), which utilize either a single 224 × 224-pixel crop at the center of the image (#1, #3) or multiple 224 × 224-pixel random crops from the image (#2, #4). By evaluating these methods with (#3, #4) and without (#1, #2) Monte Carlo dropout, the contributions of aleatoric uncertainty from the image itself and epistemic uncertainty from variation inference can be estimated. Inference method #1 serves as a baseline for comparison, making only a single prediction without MC dropout on a single image crop, resulting in no variance with respect to class SoftMax scores. For all other inference methods SoftMax scores were averaged over each random crop and MC dropout trial associated with a test image, and the label with the greatest mean SoftMax score was considered the predicted label. Inference on the U_3O_8 holdout set was repeated for each method using n values from 1 to 100 to evaluate the convergence of prediction uncertainty and classification accuracy.

2.2. Model generalizability

In addition to the holdout image sets corresponding to each trained model, predictions will be made on image sets out-of-distribution from the training data to determine how models trained with "unperturbed" datasets perform on uranium ore concentrate process history perturbations or data collection perturbations. The in-distribution holdout set splits and OOD perturbation datasets used to evaluate model generalizability are described in Table 3. Where applicable, the corresponding publications that describe the uranium morphology of interest have been listed. U morphology studies represented in the perturbation sets often only used a single precipitation route, typically uranyl peroxide, which means the effects of all process history perturbations on all precipitation routes cannot be evaluated at this time. However, by identifying which changes to UOC process history are most challenging to accurately predict, future surface morphology research needs can be more efficiently prioritized.

Predictions made on sets containing the image scale perturbation were made twice: with and without adapting the image scale to match the crops. As with HFW 6.13-μm images, unscaled predictions used random crops of 224 × 224 pixels for the CNN or 512 × 512 pixels for the VQ-VAE, though the actual physical area of sample represented would not be equal to what was seen in the training crops. HFW-scaled inference used an adaptive integer crop size C that corresponded to the physical width of what was seen during training (Eq. (2)); for example, a test image acquired at HFW 8.54-μm would use a 161-px wide crop (1.34-μm width) for prediction by the CNN or a 368-px wide crop (3.06-μm width) for the VQ-VAE feature extractor. Each of the C x C-px crops were then

Table 1
Description of training image data for 5-class U_3O_8-only model and 16-class models.

Model	Oxide(s)	Training Images					
		ADU	AUC	MDU	SDU	UO_4	Total
5-class U_3O_8	U_3O_8	104	102	96	117	80	499
16-class	UO_3	72	82	86	72	72	1336
	U_3O_8	104	102	96	117	80	
	UO_2	117	84(d), 64(i)[a]	73	73	72	

[a] AUC-UO_2 images split into the directly (d) and indirectly-reduced (i) classes.

Table 2
Summary of inference methods used to characterize prediction uncertainty.

#	MC Dropout	Crop Mode	Name	Passes/Image
1	No	Center	No MC-center	1
2	No	Random	No MC-random	n
3	Yes	Center	MC-center	n
4	Yes	Random	MC-random	n^2

Table 3
Summary of holdout (test) and OOD image datasets.

Dataset	Perturbation(s)	Images in Dataset	Reference(s)
U_3O_8 Holdout		124	[38,47]
Full Holdout		340	[38,47]
UO_3	Oxidation state	480	[38,47]
UO_2	Oxidation state	603	[38,47]
Humidity Aging	Artificial aging	142	[46,51]
High Temperature Aging	Artificial aging SEM used Image scale	500	[39,52]
UO_2Cl_2	Precipitation solution	101	[50]
Partial Reduction	Partial reduction (U_3O_8 to UO_2)	80	[a]U
Variable Temp.	Calcination temperature SEM used Image scale	247	[43]
800 °C Temp.	SEM used Image scale	103	[43]
Single Impurities	Impurities Unwashed precipitates Image scale	277	[40]
SX Impurities	Impurities Precipitation solution Partial calcination (UO_3 to U_3O_8)	180	[41]
Unwashed UO_4	Unwashed precipitates SEM used Image scale	184	[53]
Washed UO_4	SEM used Image scale	485	[53]
Different SEM	SEM used	15	[a]U
Bin (3.00, 4.57]	Image scale	300	[38,47]
Bin (4.57, 6.13]	Image scale	189	[42]
Bin [6.13, 7.70]	Image scale	300	[44]
Bin (7.70, 9.27]	Image scale	300	[44]
Bin (9.27, 10.83]	Image scale	93	[38,45]
Bin (10.83, 12.40]	Image scale	300	[38,47]

[a] U = unpublished data.

resized to the desired input size of *S* x *S* using the OpenCV *resize* function with bilinear interpolation. Scaling images by horizontal field width normalizes particle size features to those seen in the training images, which is expected to result in better predictions for test images across a wider range of acquisition scales.

$$C = int\left(\frac{train\ HFW}{test\ HFW} * S\ px\right) \quad \text{(Eq. 2)}$$

In past uranium morphology studies – particularly Schwerdt et al. (2019) and Ly et al. (2020) – the similarities and differences in surface morphology observed for the calcination products of a single synthetic route have been noted [38,47]. If a CNN trained with images of single calcination product (U_3O_8) can generalize enough to effectively make predictions on other calcination products, the scope of future morphology studies could possibly be scaled back; this will be tested by making predictions for the UO_3 and UO_2 datasets with the 5-class U_3O_8. The perturbation sets for UO_3 and UO_2 consist of images that make up the training and holdout sets for the 16-class models.

The humidity-aged perturbation set consists of uranium ore concentrate that have been artificially aged in a laboratory setting under a variety of conditions. 72 of the images are from images of UO_3 from the UO_4 synthetic route that were aged under fixed conditions with respect to aging time, temperature, and relative humidity (RH) [51]. The remaining 70 images are of U_3O_8 from the AUC or UO_4 routes that were aged with diel cycling of the temperature and relative humidity with 12-h "high" periods and 12-h "low" periods. The high and low cycles had setpoints of 45 °C/90 RH% and 25 °C/20 RH%, respectively [46].

The high temperature aging set aged U_3O_8 and UO_2 from uranyl peroxide UOCs in an atmosphere-controlled furnace under variable time, temperature, and O_2 partial pressure [39]. Surface morphology characterizations showed qualitative and quantitative changes for the aged UO_2 materials, particularly seen by decreasing circularity after aging under more extreme conditions; no significant changes were seen in the aged U_3O_8 [39]. Crystallographic changes for these materials were also quantified using P-XRD [52]. These micrographs were collected at HFW 8.53-μm using an FEI Quanta 600 FEG scanning electron microscope [39].

The uranyl chloride (UO_2Cl_2) set contains images of UO_4 materials precipitated from a solution of uranyl chloride rather than the uranyl nitrate solutions used to synthesize the unperturbed data sets. All images of the materials in this set are of the U_3O_8 calcination product acquired at a HFW of 6.13-μm. The morphology of UO_4 precipitates from the uranyl chloride solution was more platelet-like with some rounded and acicular microparticles, than the entirely acicular uranyl nitrate precipitates; the calcined U_3O_8 materials from either route had similar shapes, though the UO_2Cl_2 UOCs were somewhat larger in size [50].

The partial reduction perturbation set includes SEM images of uranium oxides that were not fully reduced from U_3O_8 to UO_2 in a 5% hydrogen atmosphere at 560 °C with a dwell time of 8 h. Rietveld refinement of the P-XRD determined that materials from the UO_4 route were 77.8 wt% α-U_3O_8 and 22.2 wt% UO_2. The AUC route UOCs were almost completely reduced to UO_2, with only 1.9 wt% belonging to the α-U_3O_8 phase. Supplemental Table 1 shows the Rietveld refinement lattice parameters for each of these materials. All SEM images in this set were acquired with a horizontal field width of 6.13-μm. The largest phase present was used as the oxidation state for the 16-class label for sets with partial calcination or reduction perturbations, e.g., the partially reduced UO_4 with 77.8 wt% α-U_3O_8 was assigned the label "UO_4–U_3O_8."

The variable calcination set consists of α-U_3O_8 materials from the uranyl peroxide route synthesized using calcination temperatures between 600 °C and 750 °C, whose images where acquired with an FEI Quanta 600 FEG SEM at HFWs ranging from 3.20-μm to 6.40-μm [43]. The 800 °C calcination image set features images of materials synthesized at the same calcination temperature as the training set U_3O_8 materials (800 °C) that were collected using the same SEM settings as the variable calcination set at HFWs between 3.20-μm to 8.53-μm [43]. UO_4 materials calcined at lower temperatures were shown to have smaller particle sizes and a somewhat lower circularity [43].

Hanson et al. (2019) has previously characterized the materials present in the single impurities set, which studied the effect of a single impurity added to the uranyl nitrate before precipitation as UO_4 and calcination to U_3O_8 [40]. This perturbation set consisted of 227 images with HFWs ranging from 3.40-μm to 9.88-μm, and included materials doped with calcium, zirconium, and magnesium impurities. Materials containing each of these materials qualitatively resembled the control materials, and most quantitative particle features were found to be statistically different [40]. A convolutional neural network was trained to classify images by the impurities present and achieved an accuracy of 83.84% [40]. Predictions on this image set were made with and without scaling the HFW of the image crops to the size of the training image crops.

Images in the SX impurities dataset were originally described in Nizinski et al. (2020), which compared the surface morphologies of UOCs extracted from uranium ore and purified by two different solvent extraction (SX) routes [41]. The materials from this study were precipitated as ADU from uranyl sulfate solutions and then partially calcined between the α-UO_3 phase (~70 wt%) and α-U_3O_8 phase (~30 wt%), as determined by Rietveld refinement of P-XRD spectra [41]. ICP-MS showed that the final UOCs were between 80 wt% and 90 wt% uranium, with sodium (0.65–1.65 wt%) and calcium (~0.80 wt%) as the most abundant elemental impurities [41]. The calcined UOCs from either purification route were described as agglomerations of smooth and mostly spherical microparticles, and a CNN was trained to distinguish materials by SX route at an accuracy greater than 97% [41].

The insufficient washing of uranium ore concentrates following their precipitation is known to result in drastically different and more complex surface morphologies than their washed counterparts [53,54]. Separate

perturbation sets were created for unwashed and washed uranyl peroxide UOCs that had been calcined to UO_3. Micrographs in either set were collected at HFW 8.53-μm using an FEI Quanta 600 FEG SEM [53]. The washed UO_4 set serves as a baseline to distinguish how the process history and data collection perturbations each affect the performance of the classification models.

To evaluate the effects of only SEM settings on convolutional neural network predictions, 15 SEM acquisitions were made using an FEI Helios NanoLab 650 SEM. U_3O_8 calcined from UO_4 precipitates were imaged at a HFW of 6.13-μm using beam voltages of 7.00-kV, which falls within the 5.00-kV to 10.00-kV range seen for the training images, and 2.00-kV, which does not. Acquisitions using 2.00-kV appeared to be grainier and have sharper contrast than those at 7.00-kV, but particle sizes and shapes appeared to be the same. This visual difference is due to the lower penetration depth of electrons at lower beam voltages, which reveals more information about the surface of the sample than deeper penetrating electrons. A comparison of micrographs acquired by the FEI Nova NanoSEM 630 and FEI Helios 650 scanning electron microscopes at varying beam voltages can be seen in Supplemental Fig. 1.

SEM micrographs of pure, unaged, and unmixed materials were acquired at horizontal field widths between 3.00-μm and 12.40-μm, excluding the training HFW of 6.13-μm. Images in this range were grouped by HFW into 6 bins with a width of 1.57-μm. Data with HFWs of 3.06-μm or 12.30-μm consisted primarily of the same UOCs as the training images [38,47]. The data in other bins was sourced from previous studies that characterized materials that would be considered unperturbed in the scope of this paper [39,42,43,45,53]. Each group of images was predicted with and without HFW-scaling.

3. Results & discussion

3.1. Uncertainty estimates for CNNs

Predictions were made on the 5-class U_3O_8 holdout set using each inference method described in Table 2. The number of predictions n for inference on each image varied between 1 and 100 for the inference methods using MC dropout or random image crops. The dropout probability was initially set to $P = 0.20$, which is equal to the dropout regularization probability determined by hyperparameter tuning. The resulting classification accuracies with uncertainties calculated by the Wilson score interval at a 1σ confidence interval (CI) and per-image average Shannon information entropies can be seen in Fig. 2(a) and (b), respectively. A classification accuracy of 0.952 was seen when making a single prediction without MC dropout on the center crop of each image ("No MC-center"), and the resulting mean information entropy was 0.190 bits per holdout set image. Predictions made on the holdout set using MC dropout with center crops ("MC-center") had a classification accuracy equal to the baseline accuracy of 0.952 for all number of MC predictions made per image. However, as n increased, so did the Shannon entropy from 0.166 bits/image at $n = 1$ until converging to around 0.220 bits per image by $n = 10$.

The random crop inference methods with ("MC-random") or without ("No MC-random") Monte Carlo dropout initially saw classification accuracies below the baseline accuracy for $n = 1$. As n increased, the classification accuracy for the No MC-random predictions converged back to the baseline accuracy at $n = 25$ and above. The MC-random predictions exceeded the baseline accuracy, peaking with an accuracy of 0.968 for $n = 10$ and $n = 25$, before decreasing back to the baseline. The information entropies for the random crop predictions were significantly higher than predictions made with only crops as the center of the image, converging around 0.300 bits/image for No MC random and 0.315 bits/image for MC-random.

These results indicate that higher classification accuracies are seen when using random image crops paired with Monte Carlo dropout during inference. Consequently, this method also results in the highest prediction uncertainties, representing both aleatoric (data) uncertainty from random crops and epistemic (model) uncertainty from MC dropout [13,21]. The uncertainties appear to converge around $n = 25$, which equates to 625 forward passes through the CNN for each prediction; this requires ~24.5 s to make predictions on each image using a single NVIDIA RTX 2060 GPU. The baseline inference method ("No MC-center") needed only one-twentieth of a second per image for predictions; the $O(n)$ inference methods took ~1.0 s/image for $n = 25$. The complete time complexity comparison can be seen in Supplemental Fig. 2.

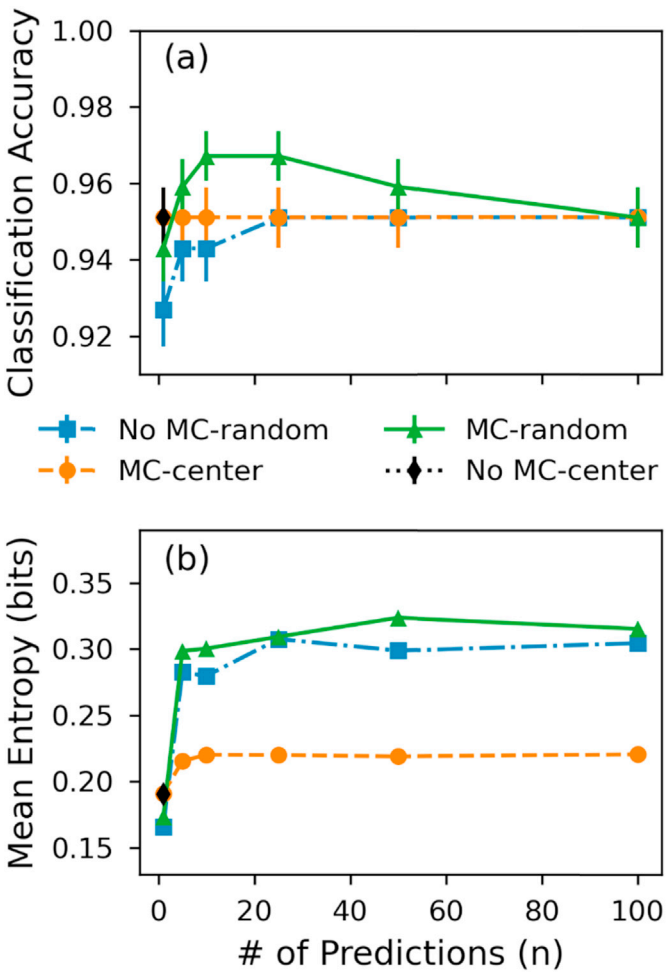

Fig. 2. (a) Classification accuracy on holdout set represented as the Wilson score interval at 1σ confidence interval and (b) mean image information entropy for each prediction method as a function of number of predictions n. Random crop methods show lower accuracy but lower uncertainties when using fewer predictions per image before converging around n = 25. The higher uncertainties of the Monte Carlo dropout + random crop inference methods may indicate the inclusion of epistemic and aleatoric uncertainties for predictions.

Using a dropout probability of 0.20, it is apparent that random image crops, representing aleatoric uncertainty, provides the largest contribution to the final prediction uncertainty. Past research has noted that CNN uncertainty estimates using MC dropout variation inference are often

Table 4
Classification accuracy and information as a function of MC dropout probability from P = 0.20 to P = 0.80

Dropout Prob. (P)	Accuracy	Mean Entropy (bits)
0.20	0.968 ± 0.007	0.309
0.35	0.960 ± 0.007	0.342
0.50	0.952 ± 0.008	0.396
0.65	0.952 ± 0.008	0.490
0.80	0.952 ± 0.008	0.700

miscalibrated [55,56]. Table 4 shows the classification accuracy and mean information entropy for MC-random $n = 25$ predictions made on the holdout set for dropout probabilities between 0.20 and 0.80. At $P = 0.80$ the mean information entropy is 0.700 bits/image, more than twice the uncertainty seen at $P = 0.20$. A small trade-off in classification accuracy exists, since fewer of the image features extracted by the trained CNN are available for predictions as more connections are dropped; the classification accuracy at $P = 0.80$ only decreases to an accuracy of 0.952. This suggests that CNN prediction uncertainty estimates for uranium ore concentrate morphology images could possibly be scaled by adjusting the dropout probability, though calibration of uncertainty estimates remains more elusive.

Since both Monte Carlo dropout and random image crops are stochastic inference methods, some deviation in predictions is expected each time predictions are made on a set of images. To quantify the consistency of stochastic inference, 10 sets of predictions were made on the holdout data using the trained model with MC-random inference ($n = 25$, $P = 0.20$). The results (Table 5) show a range of classification accuracies from 0.952 to 0.968, with a mean and standard deviation of 0.956 ± 0.008. This would indicate that the benefit of MC-random inference over the baseline is less than previously shown in Fig. 2 and Table 4. The relative standard deviation was 0.837% for the classification accuracy and 1.27% for the per-image Shannon information entropy suggests that $n = 25$ is sufficient for consistent and reproducible results by MC dropout variational inference. Moreover, the standard deviation in classification accuracy empirically seen with multiple trials (0.008) was well-aligned with the reported 1σ Wilson score intervals (0.007–0.008).

While related work has noted practical success with the implementation and calibration of variational inference methods across various data domains, other efforts are still seeking to establish a sound theoretical basis for Monte Carlo dropout [55–59]. Sicking et al. (2020) empirically demonstrated correlated systems that were more complex in nature – latent activation layers that combine Gaussian and non-Gaussian, exponentially-tailed distributions differently for train, test, and OOD image data – than previously assumed, which may complicate the base assumption that MC dropout can be used for Bayesian approximation [26,59]. Future research characterizing the statistical distributions of deep learning models and the discovery of new distribution functions, such as the continuous Bernoulli and continuous categorical distributions, will likely be required to further improve the performance, tractability, and quality of UQ for deep learning models [59–61].

3.2. Model generalizability

Since UO_3 and UO_2 materials were included in the 16-class training sets, predictions on these perturbation sets were only made by the 5-class U_3O_8 model. The classification accuracy for the UO_3 set was 25.0%, compared to 90.2% for the UO_2 set. In qualitative descriptions of these materials, Schwerdt et al. (2019) notes the similarities in surface morphology, particularly the formation of elongated platelets, seen in the ADU, MDU, and SDU starting materials and low-temperature calcined UO_3 [38]. Sintering was shown to occur for these UOCs during the 800 °C calcination to U_3O_8, which formed sub-rounded grains for the SDU route, spherical particles for ADU, and more rounded platelets for MDU; little to no qualitative change in surface morphology was seen for these materials during the reduction of U_3O_8 to UO_2 in a hydrogen atmosphere [38].

Table 5
Consistency of 10 replicates of MC-random ($n = 25$, $P = 0.20$) predictions on the holdout set with respect to classification accuracy and uncertainty. RSD indicates relative standard deviation.

	Accuracy	Mean Entropy (bits)
Minimum	0.952	0.307
Maximum	0.968	0.318
Mean ± 1σ	0.956 ± 0.008	0.314 ± 0.004
RSD (%)	0.837%	1.27%

The confusion matrix of the UO_3 predictions by the 5-class U_3O_8 classifier shows significant confusion for every label (Fig. 3). Images of ADU and MDU UOCs were nearly always predicted as SDU. UO_4 and SDU images were predicted as either MDU or SDU. The AUC-UO_3 images were most often incorrectly predicted as belonging to the MDU or SDU route. In contrast, the UO_2 perturbation saw relatively little confusion, which can be explained by the qualitative similarities seen in U_3O_8 and UO_2 materials from each precipitation route. This confirms the need to thoroughly characterize surface morphologies at each oxidation state along uranium's process history when developing nuclear forensics datasets, and that training models on a single product does not produce generalizable models. As such, only the 16-class models were used for predicting the remaining OOD sets.

Each 16-class model made predictions with high accuracy on its respective holdout image set, though a higher accuracy was seen with the fully supervised model (0.929) over the VQ-VAE (0.856); confusion matrices for the 16-class CNN and MLP trained on VQ-VAE feature codebooks can be seen in Fig. 4. The supervised CNN showed the most confusion in overpredicting the examples as UO_3 calcined from an MDU starting material. The incorrectly predicted examples from the MLP were more dispersed, and most of the confusion was between images of various uranium oxides synthesized from the ammonium uranyl carbonate starting material.

The classification accuracies of predictions made by each model on the out-of-distribution image data sets can be seen in Table 6; Wilson score intervals at a 1σ CI can be found in Supplemental Table 3. For sets that include the scale on the image as a data perturbation, the classification accuracy is listed for both the unscaled and HFW-scaled predictions. Confusion matrices, which help visualize a classifier's correct and incorrect predictions, can be found in the Supplemental Information for each classifier and image set. In general, the classification accuracy on OOD datasets was very low, with both classifiers failing to achieve above random chance (1/16 = 6.25%) for six of the 17 OOD sets. These sets include both aging sets (humidity and high temperature), the partially reduced U_3O_8 to UO_2, U_3O_8 calcined at variable temperatures, materials extracted from uranium ores then purified by solvent extraction, and unwashed uranyl peroxide precipitates. Three of these datasets (high temperature aging, variable temperature, and unwashed UO_4) were collected on a different SEM than the training data. Two of the sets (partial reduction and SX impurities) contained images of materials that were not fully converted to a single uranium oxide phase. Apart from the partially reduced UO_2 set, each of these OOD sets with predictions below chance contained more than one perturbation from the training data, which highlights the need for greater understanding of compound effects in uranium process history and SEM data collection when developing machine learning datasets and classifiers.

Predictions on the humidity aging OOD set by each 16-class model had a classification accuracy of 2.1%. The confusion matrix for the CNN predictions shows that micrographs of aged materials from the uranyl peroxide and ammonium uranyl carbonate synthetic routes were overwhelmingly misclassified as belonging to the magnesium diuranate route (Supplemental Fig. 3). In contrast, the predictions for aged UO_4 materials by the MLP had misclassifications across nearly all other precipitates. The MLP's predictions of the aged U_3O_8 from AUC materials reported a classification accuracy of 0.0%, but 16 of the 35 predictions went to classes for other uranium oxides produced from AUC (12 for UO_2 directly reduced from AUC and 4 for UO_3 calcined from AUC), which means that 46% of aged AUC micrographs were predicted to the correct precipitation route. Supplemental Fig. 4 shows the relative similarity between unaged and diel-cycling aged AUC compared to UO_4, which may indicate that classification models might have an easier time predicting aged materials from the more unique starting morphologies, like those seen for AUC and MDU [46].

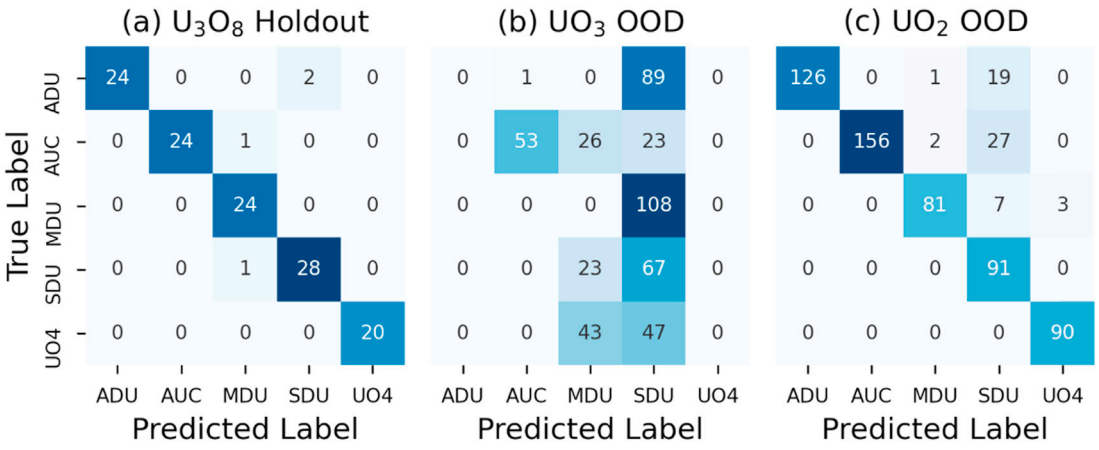

Fig. 3. Confusion matrices for predictions made on the (a) U_3O_8 holdout set, (b) UO_3 perturbation set, and (c) UO_2 perturbation set by the 5-class model trained only on U_3O_8 image data. Predictions on the UO_3 set show that the MDU and SDU routes were the most frequent mispredictions.

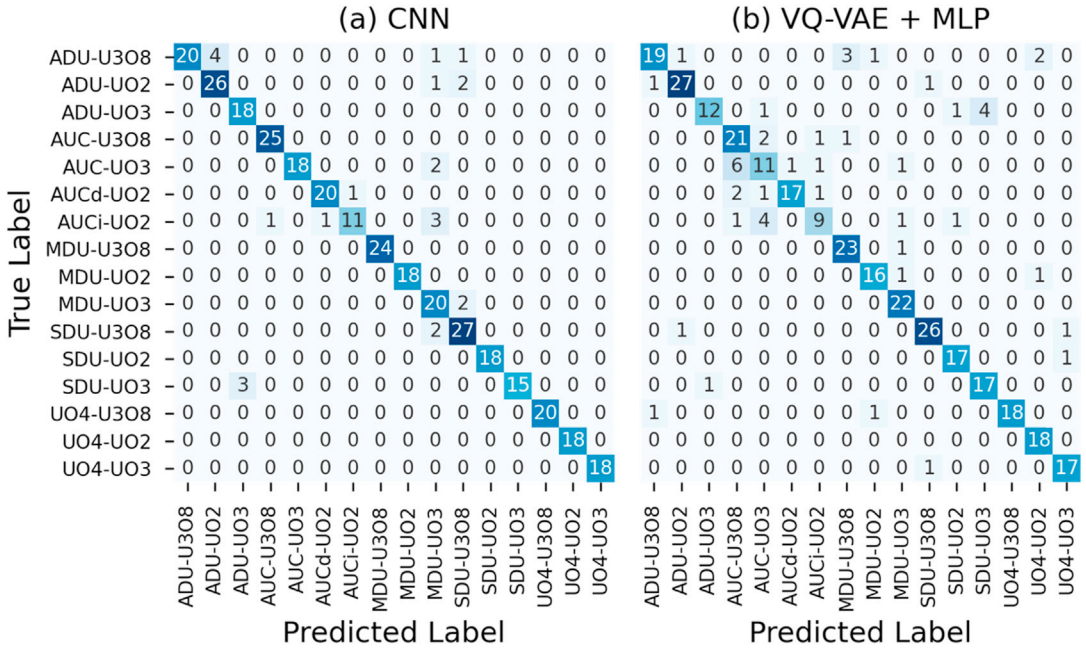

Fig. 4. Confusion matrices for (a) 16-class CNN and (b) VQ-VAE + MLP predictions made on the holdout set using random crops with MC dropout inference (n = 25).

No correct predictions were made for the high temperature aging set by either classifier. The unscaled and scaled CNN predictions were biased heavily towards the ADU-UO_3 and SDU-UO_3 classes, respectively (Supplemental Fig. 5). As with the humidity aging set, confusion in predictions by the MLP trained on unsupervised VQ-VAE features were spread across more labels.

Micrographs of UO_4 precipitated from UO_2Cl_2 solutions and calcined to U_3O_8 were very accurately predicted by the CNN model, with a classification accuracy of 97.0%. The same image set predicted by the multilevel perceptron trained on autoencoder feature histograms had a significantly lower classification accuracy (41.6%). Supplemental Fig. 6 shows that most of this confusion was towards the MDU-U_3O_8 class, though the exact reason for this cannot be determined.

Varying calcination conditions, as seen in the partial reduction and variable calcination temperature sets, also proved to be challenging OOD sets for the classifiers. Confusion in the partially reduced UO_4 or AUC-route U_3O_8/UO_2 indicated a heavy bias towards the MDU routes by both classifiers (Supplemental Fig. 7). Predictions by the CNN saw a much higher classification accuracy when the same calcination temperature as the training set is used (38.8%) than with other temperatures (4.5%), despite the fact that α-U_3O_8 was synthesized in each case [38,43]. It may be beneficial to slightly vary these conditions when synthesizing future materials for uranium oxide morphology dataset development in order to introduce more variance within each image class.

The unscaled predictions from the supervised CNN had the highest classification accuracy (37.2%) on the single impurities data set; the unscaled VQ-VAE + MLP predictions had an accuracy of 22.0%. Both sets of unscaled predictions outperformed the HFW-scaled predictions, whereas the opposite trend was seen for nearly all other OOD sets acquired at varying image scales. Hanson et al. (2019) notes that SEM HFW's for each impurity element were individually selected to best represent each's unique particle morphology characteristics [40]. Table 7, which splits the predictions by impurity, shows that images from unwashed uranyl peroxide precipitates with calcium or magnesium impurities could be classified with higher accuracy than the washed precipitates despite higher impurity concentrations and surface morphologies that were more distinct from the pure control samples

Table 6
Summary of predictions made for holdout (test) and perturbation image datasets by the 16-class supervised CNN and 16-class MLP trained on features from the unsupervised VQ-VAE. A second set of predictions were made using HFW scaling by adaptive crop sized (Eq. (2)) when an OOD set's HFW was not equal to that of the training set, 6.13-μm. Corresponding confusion matrices and uncertainties calculated by the 1σ Wilson score interval have been included as supplemental information.

Dataset	16-class CNN		VQ-VAE + MLP		Supplemental Figure
	Unscaled	Scaled	Unscaled	Scaled	
Humidity Aging	0.021	–	0.021	–	3
High Temp. Aging	0.000	0.004	0.000	0.000	5
UO_2Cl_2	**0.970**	–	0.416	–	6
Partial Reduction	0.013	–	0.013	–	7
Variable Temp.	0.000	0.045	0.000	0.000	8
800 °C Temp.	0.029	**0.388**	0.000	0.000	9
Single Impurities	**0.372**	0.029	0.220	0.112	10
SX Impurities	0.000	–	0.000	–	11
Unwashed UO_4	0.000	0.000	0.022	0.017	12
Washed UO_4	0.008	**0.177**	0.014	0.017	13
Different SEM	**0.333**	–	0.133	–	14
Bin (3.00, 4.57]	0.590	**0.693**	0.343	0.613	15
Bin (4.57, 6.13]	0.280	**0.302**	0.153	0.148	16
Bin [6.13, 7.70]	0.073	**0.302**	0.061	0.061	17
Bin (7.70, 9.27]	0.110	**0.340**	0.051	0.000	18
Bin (9.27, 10.83]	0.355	**0.484**	0.070	0.070	19
Bin (10.83, 12.40]	0.717	0.577	0.433	**0.727**	20

Table 7
Unscaled predictions on the single impurity OOD set split by impurity element. Uncertainties calculated by the 1σ Wilson score interval are tabulated in Supplemental Table 3.

Impurity	Images	CNN Accuracy	VQ-VAE + MLP Accuracy
Ca (unwashed)	47	0.574	0.404
Ca (washed)	51	0.118	0.098
Mg (unwashed)	34	0.147	0.029
Mg (washed)	34	0.000	0.000
Zr (unwashed)	57	0.579	0.316
Zr (washed)	54	0.593	0.333

[62]. Supplemental Fig. 10 shows that U_3O_8 from UO_4 containing elemental impurities were often classified as various oxides from MDU or UO_2 from UO_4. The latter further supports the earlier results showing that models may not be able to easily discriminate between the relatively similar morphologies seen in U_3O_8 and UO_2 from the same precipitation route.

Confusion matrices for predictions on the SX impurities data set overwhelmingly show sodium diuranate labels as the most common incorrect prediction (Supplemental Fig. 11); it is unclear to what extent the mispredictions as SDU ore concentrates can be attributed to the most abundant impurity element in these UOCs (sodium), or to the similarities in morphology between SDU and ADU material that have been calcined to UO_3 [38,41]. The 16-class CNN made 7/180 predictions as ADU-U_3O_8, which is below random chance; the VQ-VAE + MLP made 20 predictions as ADU-U_3O_8 and 5 predictions as ADU-UO_2, for a 13.9% classification when only the correct precipitation route is concerned.

Predictions on the 15 micrographs acquired using the FEI Helios 650 scanning electron microscope were 33% accurate for the 16-class CNN. The 5 correct predictions made by each classifier had been acquired at a beam voltage of 7.00-kV. Each incorrectly classified image was acquired using a 2.00-kV beam voltage and was predicted as ADU by the 5-class models or ADU-UO_2 by the 16-class model, rather than the true label of UO_4 calcined to U_3O_8 (Supplemental Fig. 14). The VQ-VAE feature-trained MLP predicted UO_2 or U_3O_8 from UO_4 for four of five 7.00-kV images, and UO_2 from SDU for each 2.00-kV image and the remaining 7.00-kV image. While the size of this OOD data set (15 images) was relatively small, consistent trends were apparent between the two models. Recent advances in domain adaption methods for image classifiers may provide a way to train models only on morphology features without recognizing biases related to data collection parameters.

The classification accuracy plotted against HFW bin shows a "U" shape for both the 16-class CNN and MLP trained on VQ-VAE features, with higher classification accuracies at the bins furthest from the 6.13-μm scale of the training data (Fig. 5). This trend was the opposite of what was hypothesized: that image scales nearer the training data would result in better predictions. However, the images acquired at horizontal field widths of 3.06-μm and 12.3-μm came from the same set of samples synthesized for Ly et al., 's 2020 multi-magnification study, which indicates that image scale plays a smaller role than the inherent variance in other pure, unaged uranium oxides synthesized under slightly different conditions [47].

Implementing HFW-scaling during inference (as described by Eq. (2)) improved classification accuracy on most, but not all, OOD sets. Whereas the fully supervised 16-class CNN saw the greatest relative improvement with respect to classification accuracy using HFW-scaling for the intermediate HFW bins (4.57-μm to 10.83-μm), the MLP classifier had little-to-no improvement for these bins but significantly higher classification accuracies for scaled predictions on the (3.00-μm, 4.57-μm] and (10.83-

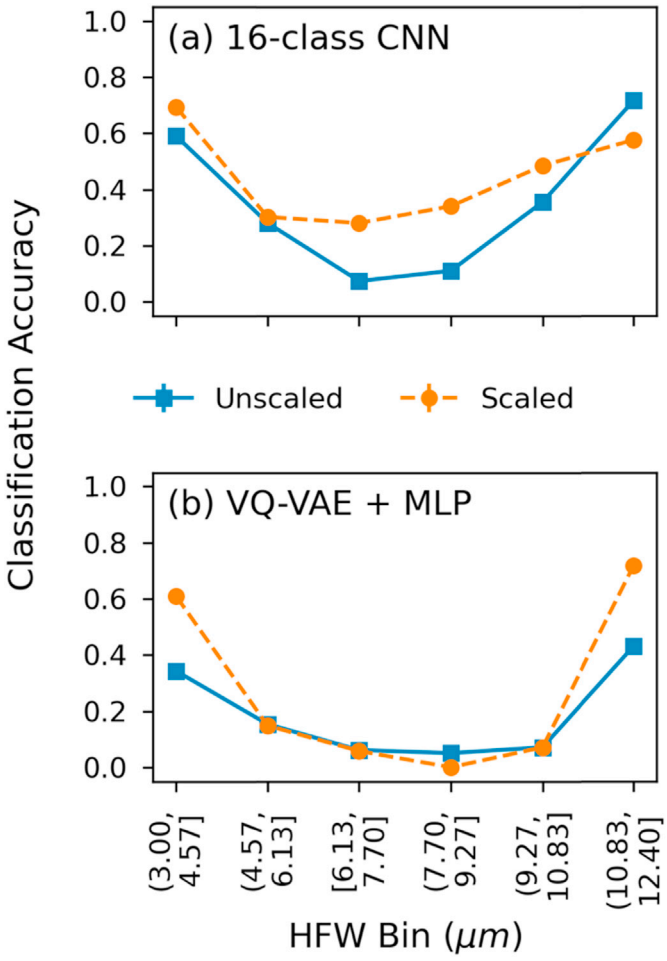

Fig. 5. Predictions on images acquired at HFWs ranging from 3.00-μm to 12.40-μm (excluding the training HFW of 6.13-μm) for the (a) 16-class CNN and (b) MLP trained on VQ-VAE features. Higher classification accuracies for the HFW bins furthest from the training HFW can be attributed to the fact that images in these sets come from the same materials as those imaged at 6.13-μm for the training data. This indicates that image scale has less impact on model performance than the variance seen in other pure, unaged materials that have been synthesized and analyzed for other research efforts than the training set.

μm, 12.40-μm] bins. Confusion matrices for each model's HFW-scaled and unscaled predictions can be found in Supplemental Figs. 15–20. As with the holdout set, the 3.06-μm and 12.30-μm predictions show confusion mostly between adjacent classes of the same precipitation route. The intermediate HFW bins, which consisted largely of UO_4-route micrographs, had confusion across the ADU, MDU, and SDU uranium oxide labels.

3.3. Evaluation of uncertainties for OOD sets

Sections 3.1 and 3.2 presented results for the implementation of uncertainty quantification with MC dropout and random image crops for CNNs and the performance of classification models on out-of-distribution data sets, respectively. One purpose of investigating uncertainty estimates for neural network predictions is to determine what the model is and is not confident in, and when the end-user can trust the model's confidence in its predictions. In an ideal scenario, one would see low uncertainties for a set of predictions and believe that those predictions were correct; predictions with high uncertainties could be considered low-confidence and face additional scrutiny from human domain experts. However, biases and errors from many sources – insufficient training data, mislabeled training data, overfit models, domain shift, OOD data, etc. – can lead to models that are under confident for correct predictions and overconfident for incorrect predictions, which makes knowing what to trust troublesome.

To evaluate the quality of the classifiers' UQ, the Shannon information entropies were computed for each holdout and out-of-distribution set predictions made by the 16-class CNN and MLP from VQ-VAE features. Mean entropy values for each set's correctly and incorrectly predicted examples were plotted with the color of plot markers indicating the classification accuracy on that set (Fig. 6). OOD sets with 0% classification accuracy were excluded from this figure; entropy values for all sets are tabulated in Supplemental Table 4. Accurately predicted sets can be seen clustered towards the bottom left of the plot for either classifier, which shows lower uncertainties for correct predictions than for incorrect predictions. For the supervised CNN, these sets include the holdout set, uranyl chloride set, and several HFW bin sets. Likewise, the VQ-VAE + MLP classifier showed confident correct predictions for the holdout set and several HFW bins.

For OOD sets predicted with lower classification accuracies the CNN-predicted sets generally had higher mean information entropies for correctly *and* incorrectly predicted examples on low-accuracy OOD sets, which indicates a lower level of confidence for all examples in sets where the classification accuracy metric measured poor performance. This would mean all predictions could be rejected due to low confidence. Some exceptions to this trend, such as the partial reduction set (1.2% classification accuracy) and the humidity aging set (2.1%), showed incorrect example entropies that were lower than those for correctly predicted examples. In contrast, OOD sets predicted with low classification accuracies by the MLP had lower entropies for incorrect predictions than correct predictions. In such cases the low information entropy values of incorrect predictions may result in false positive examples being accepted by the end-user. Better generalizing classifiers and well-calibrated UQ methods will be needed to avoid such pitfalls in the future.

Significant challenges were seen in making predictions and for obtaining reasonable uncertainty estimates on out-of-distribution data with either fully supervised convolutional neural networks or networks that were trained on features extracted by unsupervised autoencoders. However, the results were not entirely unexpected. The ability to produce models that can generalize to different kinds of OOD data is problematic for state-of-the-art machine learning classification models and remains a very active area of research. Recent work by Power et al. has utilized small algorithmic datasets to better probe the mechanisms by which neural networks generalize, though these efforts have not yet been applied to larger image datasets [63]. For image data collected by different sources, domain adaptation might help develop classifiers that

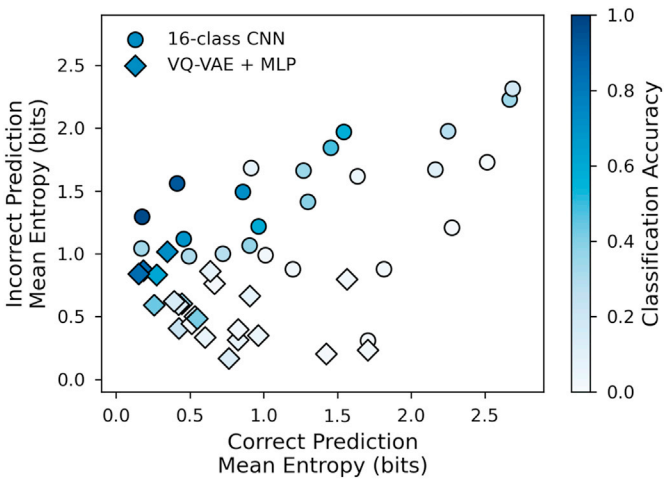

Fig. 6. Mean information entropy values of correct and incorrect predictions on OOD sets by the 16-class CNN and MLP trained on VQ-VAE features; darker plot marker shading represents higher classification accuracies. Accurately predicted sets from either classifier show low uncertainties for correct predictions and high uncertainties for incorrect predictions. The CNN classifier showed high uncertainties for all predictions from datasets that were not accurately predicted, whereas the VQ-VAE + MLP classifier tended to be overconfident (low uncertainties) on sets predicted with lower classification accuracies.

use image features relevant only to the image classes, and not the way in which they were collected. In the context of scanning electron micrographs, implementing domain adaptation for neural networks could likely reduce the depreciation in classifier performance related to SEM data collection parameters, which was particularly apparent with the FEI Helios SEM OOD set [64,65]. Specifically, test-time domain adaptation could be useful for classification models utilizing features extracted by unsupervised models.

While the overall performance of the fully supervised CNN and unsupervised VQ-VAE features used with a supervised MLP were similar, both eventually use large, labeled image sets to train a classifier. The full utility of the VQ-VAE features for uranium oxide morphology data outside of direct classification has not yet been explored. Oord et al. (2017) hypothesized that the discrete nature of VQ-VAE codebook histograms would be especially useful for categorical contexts, such as descriptions of images with natural language [8]. This holds particular relevance for relating morphology features to nuclear forensics morphology lexicon, which may help develop highly interpretable machine learning models [37]. Additionally, the discrete learned representations of the VQ-VAE are highly descriptive with a relatively low dimensionality, which may prove useful in few-shot learning (FSL) and human-in-the-loop machine learning contexts [66–69]. FSL models have previously shown a great ability to generalize using only a few – and in the case of zero-shot learning, zero – training examples for each class. Human-in-the-loop implementations for nuclear forensics could synergistically pair the knowledge and insight of domain experts with the highly descriptive image representations to better characterize unknown materials.

4. Conclusions

Uncertainty quantification was successfully implemented for convolutional neural networks predicting the process history from SEM images. The combination of Monte Carlo dropout and random image crops creates per-class uncertainty estimates that capture both aleatoric and epistemic uncertainties, though provides no significant boost in classification accuracy, as reported by others. At 25 MC dropout predictions on 25 random image crops per SEM image, the uncertainties were found to converge, indicating reproducible UQ by variation

inference methods. The magnitude of the uncertainties was found to be scalable by adjusting the dropout probability without significantly affecting model performance; further investigation is needed before it can be determined whether uncertainty estimates by MC dropout can truly be calibrated.

Making predictions on OOD image datasets using a classifier trained with uniformly acquired SEM images of unperturbed UOCs from common processing routes demonstrated the shortfalls of relying on a narrow set of training data, and highlighted areas where more surface morphology data is needed. Perturbations of lesser concern include the solution of precipitation and image scale, when considered alone. Partial conversions, aged materials, and chemical impurities all led to low classification accuracies when compared to the unperturbed holdout sets. The classifiers were also sensitive to the image collection parameters of the scanning electron microscope, even when the same samples were represented in the training images, signifying the importance of utilizing training data from multiple facilities. Domain adaptation neural networks should be explored for SEM image data of uranium oxides collected by different microscopes and acquisition parameters once sufficient data is collected.

Most present shortcomings with predicting on OOD data mainly stem from a shortage of available data, leading to significant data biases and a limited ability to generalize. Among the most pressing research questions are how chemical impurities, aging, partial conversions, and combinations of all other factors affect the UOC morphology for each precipitation product. Design-of-experiment studies might be able to quantitatively tease out the most significant effects. FSL and human-in-the-loop learning schemes could produce models with greater generalizability when only a few examples are available for each image class and perturbation. Further explainability might also be accomplished by combining visual explanations in natural language with nuclear material lexicon descriptions, allowing classification models to describe why a processing route was or was not predicted. In any case, staying at the forefront of deep learning advances will be crucial for developing the most generalizable and interpretable classification models for nuclear forensics using surface morphology signatures.

Author statement

Cody A. Nizinski: Conceptualization, Methodology, Formal Analysis, Investigation, Writing – Original Draft, Visualization.

Cuong Ly: Methodology, Writing – Review & Editing.

Clement Vachet: Methodology, Writing – Review & Editing.

Alex Hagen: Methodology, Formal Analysis, Writing – Review & Editing.

Tolga Tasdizen: Supervision, Funding Acquisition, Writing – Review & Editing.

Luther W. McDonald IV: Supervision, Project Administration, Funding Acquisition, Writing – Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the Department of Homeland Security, under Grant Award no. 2016-DN-077-ARI102 and the NA22 in the office of nuclear nonproliferation under: LA21-ML-MorphologySignature-P86-NTNF1b. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the funding agencies.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2022.104556.

References

[1] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (1998) 2278–2323.
[2] W. Rawat, Z. Wang, Deep convolutional neural networks for image classification: a comprehensive Review, Neural Comput. MIT Press J. 1 (2017) 2352–2449. September.
[3] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, 2012.
[4] A.F. Agarap, Deep Learning Using Rectified Linear Units, ReLU), 2018.
[5] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; IEEE Computer Society 2016–, 2016, pp. 770–778. December.
[6] D.P. Kingma, M. Welling, An introduction to variational autoencoders, Found. Trends Mach. Learn. 12 (2019) 307–392.
[7] C. Doersch, Tutorial on Variational Autoencoders, 2016.
[8] A. Oord, den van, O. Vinyals, K. Kavukcuoglu, Neural discrete representation learning, Adv. Neural Inf. Process. Syst. *2017* (2017) 6307–6316. *December*.
[9] A. Razavi, A. Oord, den van, O. Vinyals, Generating diverse high-fidelity images with VQ-VAE-2, Adv. Neural Inf. Process. Syst. 32 (2019).
[10] M. Girard, A. Hagen, I. Schwerdt, M. Gaumer, L. McDonald, N. Hodas, E. Jurrus, Uranium oxide synthetic pathway discernment through unsupervised morphological analysis, J. Nucl. Mater. 552 (2021) 152983.
[11] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: an overview of interpretability of machine learning, in: Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 80–89.
[12] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[13] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 30 (2017) 5574–5584.
[14] M. Du, F. Yang, N. Zou, X. Hu, Fairness in deep learning: a computational perspective, IEEE Intell. Syst. 36 (2020) 25–34.
[15] H.M.D. Kabir, A. Khosravi, M.A. Hosen, S. Nahavandi, Neural network-based uncertainty quantification: a survey of methodologies and applications, IEEE Access. Inst. Electr. Electr. Eng. Inc. 3 (2018) 36218–36234. June.
[16] Y. Ovadia, G. Research, E. Fertig, J. Ren, Z. Nado Google Research, S. Nowozin Google Research, J.V. Dillon Google Research, B. Lakshminarayanan, J. Snoek, Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift 32, 2019.
[17] N. Akhtar, A. Mian, Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE Access*, vol. 16, Institute of Electrical and Electronics Engineers Inc. February, 2018, pp. 14410–14430.
[18] D. Hendrycks, T. Dietterich, Benchmarking neural network robustness to common corruptions and perturbations, 7th Int. Conf. Learn. Represent. ICLR 2019 (2019).
[19] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, 5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc. (2016).
[20] X. Zhang, X. Xie, L. Ma, X. Du, Q. Hu, Y. Liu, J. Zhao, M. Sun, Towards characterizing adversarial defects of deep learning software from the lens of uncertainty, Proc. - Int. Conf. Softw. Eng. (2020) 739–751.
[21] A. Der Kiureghian, O. Ditlevsen, Aleatory or epistemic? Does it matter? Struct. Saf. 31 (2009) 105–112.
[22] B. Lakshminarayanan, A. Pritzel, C.B. Deepmind, Simple and scalable predictive uncertainty estimation using deep ensembles, in: 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017.
[23] Y. Gal, Z. Ghahramani, Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference, 2015.
[24] J. Miguel Hernández-Lobato, R.P. Adams, Probabilistic backpropagation for scalable learning of bayesian neural networks, in: Proceedings of the 32 Nd International Conference on Machine Learning, Lille, France, 2015.
[25] K. Shridhar, F. Laumann, M. Liwicki, Uncertainty Estimations by Softplus Normalization in Bayesian Convolutional Neural Networks with Variational Inference, 2018.
[26] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: representing model uncertainty in deep learning, 33rd Int. Conf. Mach. Learn. ICML 3 (2016 2015) 1651–1660.
[27] D. Tran, M.W. Dusenberry, M. van der Wilk, D. Hafner, Bayesian layers: a module for neural network uncertainty, in: Advances in Neural Information Processing Systems; Neural Information Processing Systems Foundation, vol. 32, 2018.
[28] N. Srivastava, G. Hinton, A. Krizhevsky, R. Dropout Salakhutdinov, A Simple Way to Prevent Neural Networks from Overfitting, vol. 15, 2014.
[29] C. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423.
[30] K. Mayer, M. Wallenius, I. Ray, Nuclear Forensics - A Methodology Providing Clues on the Origin of Illicitly Trafficked Nuclear Materials. *Analyst*, vol. 17, Royal Society of Chemistry March, 2005, pp. 433–441.

[31] K. Mayer, M. Wallenius, T. Fanghänel, Nuclear forensic science—from cradle to maturity, J. Alloys Compd. 444–445 (2007) 50–56.
[32] M.J. Kristo, S.J. Tumey, The state of nuclear forensics, in: Nuclear Instruments and Methods in Physics Research, Section B: Beam Interactions with Materials and Atoms; North-Holland, vol. 294, 2013, pp. 656–661.
[33] Z. Varga, J. Krajkó, M. Peńkin, M. Novák, Z. Eke, M. Wallenius, K. Mayer, Identification of uranium signatures relevant for nuclear safeguards and forensics, J. Radioanal. Nucl. Chem. 312 (2017) 639–654.
[34] I.L. Ray, A. Schubert, M. Wallenius, The concept of a "microstructural fingerprint" for the characterization of samples in nuclear forensic science, in: International Conference on Advances in Destructive and Non-destructive Analysis for Environmental Monitoring and Nuclear Forensics, International Atomic Energy Agency (IAEA), Karlsruhe, Germany, 2003, pp. 371–373.
[35] R. Porter, C. Ruggiero, D. Hush, N. Harvey, P. Kelly, W. Scoggins, L. Tandon, Interactive image quantification tools in nuclear material forensics. In *image processing: machine vision applications IV*, SPIEL 7877 (2011) 787708.
[36] C.A. Wilson, W.P. Adderley, A.N. Tyler, P. Dale, Characterising the morphological properties and surface composition of radium contaminated particles: a means of interpreting origin and deposition, Environ. Sci. Process. Impacts 15 (2013) 1921–1929.
[37] A.L. Tamasi, L.J. Cash, C. Eley, R.B. Porter, D.L. Pugmire, A.R. Ross, C.E. Ruggiero, L. Tandon, G.L. Wagner, J.R. Walensky, A.D. Wall, M.P. Wilkerson, A lexicon for consistent description of material images for nuclear forensics, J. Radioanal. Nucl. Chem. 307 (2016) 1611–1619.
[38] I.J. Schwerdt, C.G. Hawkins, B. Taylor, A. Brenkmann, S. Martinson, L.W. McDonald, Uranium oxide synthetic pathway discernment through thermal decomposition and morphological analysis, Radiochim. Acta 107 (2019) 193–205.
[39] A.M. Olsen, I. Schwerdt, A. Jolley, N. Halverson, B. Richards, L.W.M.D. Iv, A response surface model of morphological changes in UO2 and U3O8 following high temperature aging, Radiochim. Acta 107 (2019) 449–458.
[40] A.B. Hanson, R.N. Lee, C. Vachet, I.J. Schwerdt, T. Tasdizen, L.W. McDonald, Quantifying impurity effects on the surface morphology of α-U3O8, Anal. Chem. 91 (2019) 10081–10087.
[41] C.A. Nizinski, A.B. Hanson, B.C. Fullmer, N.J. Mecham, T. Tasdizen, L.W. McDonald, Effects of process history on the surface morphology of uranium ore concentrates extracted from ore, Miner. Eng. 156 (2020) 106457.
[42] S.T. Heffernan, N.C. Ly, B.J. Mower, C. Vachet, I.J. Schwerdt, T. Tasdizen, L.W. McDonald, Identifying surface morphological characteristics to differentiate between mixtures of U3O8 synthesized from ammonium diuranate and uranyl peroxide, Radiochim. Acta 108 (2019) 29–36.
[43] A.M. Olsen, B. Richards, I. Schwerdt, S. Heffernan, R. Lusk, B. Smith, E. Jurrus, C. Ruggiero, L.W. McDonald, Quantifying morphological features of α-U3O8 with image analysis for nuclear forensics, Anal. Chem. 89 (2017) 3177–3183.
[44] I.J. Schwerdt, A. Olsen, R. Lusk, S. Heffernan, M. Klosterman, B. Collins, S. Martinson, T. Kirkham, L.W. McDonald, Nuclear forensics investigation of morphological signatures in the thermal decomposition of uranyl peroxide, Talanta 176 (2018) 284–292.
[45] E.C. Abbott, A. Brenkmann, C. Galbraith, J. Ong, I.J. Schwerdt, B.D. Albrecht, T. Tasdizen, L.W. McDonald, Dependence of UO2 surface morphology on processing history within a single synthetic route, Radiochim. Acta 107 (2019) 1121–1131.
[46] A.B. Hanson, C.A. Nizinski, W. Luther, I. McDonald, Effect of diel cycling temperature, relative humidity, and synthetic route on the surface morphology and hydrolysis of α-U3O8, ACS Omega 6 (2021) 18426–18433.
[47] C. Ly, C. Vachet, I. Schwerdt, E. Abbott, A. Brenkmann, L.W. McDonald, T. Tasdizen, Determining uranium ore concentrates and their calcination products via image classification of multiple magnifications, J. Nucl. Mater. 533 (2020) 152082.
[48] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F.A. Wichmann, W. Brendel, ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, 7th Int. Conf. Learn. Represent. ICLR 2018 (2019).
[49] D.P. Kingma, J.L. Ba, Adam: a method for stochastic optimization, in: *3rd International Conference On Learning Representations, ICLR 2015 - Conference Track Proceedings*; International Conference on Learning Representations, ICLR, 2015.
[50] E.C. Abbott, H.E. O'Connor, C.A. Nizinski, L.D. Gibb, E.W. Allen, L.W. McDonald IV, Thermodynamic evaluation of the uranyl peroxide synthetic route on morphology, J. Nucl. Mater. (2021). *In press*.
[51] A.B. Hanson, I.J. Schwerdt, C.A. Nizinski, R.N. Lee, N.J. Mecham, E.C. Abbott, S. Heffernan, A. Olsen, M.R. Klosterman, S. Martinson, A. Brenkmann, L.W. McDonald, Impact of controlled storage conditions on the hydrolysis and surface morphology of amorphous-UO3, ACS Omega 6 (2021) 8605–8615.
[52] A.M. Olsen, I.J. Schwerdt, B. Richards, L.W. McDonald, Quantification of high temperature oxidation of U3O8 and UO2, J. Nucl. Mater. 508 (2018) 574–582.
[53] I.J. Schwerdt, A. Brenkmann, S. Martinson, B.D. Albrecht, S. Heffernan, M.R. Klosterman, T. Kirkham, T. Tasdizen, L.W. McDonald IV, Nuclear proliferomics: a new field of study to identify signatures of nuclear materials as demonstrated on alpha-UO3, Talanta 186 (2018) 433–444.
[54] N.B.A. Thompson, M.R. Gilbert, N.C. Hyatt, Nuclear forensic signatures of studtite and α-UO3 from a matrix of solution processing parameters, J. Nucl. Mater. 544 (2021) 152713.
[55] K. Fang, C. Shen, D. Kifer, Evaluating aleatoric and epistemic uncertainties of time series deep learning models for soil moisture predictions, Water Resour. Res. 56 (2019).
[56] M.-H. Laves, S. Ihler, K.-P. Kortmann, T. Ortmaier, Well-Calibrated Model Uncertainty with Temperature Scaling for Dropout Variational Inference, in: In 4th Workshop on Bayesian Deep Learning (NeurIPS 2019); arXiv, 2019.
[57] L. Zhu, N. Laptev, Deep and confident prediction for time series at uber, in: IEEE International Conference on Data Mining Workshops, ICDMW; IEEE Computer Society, 2017-November, 2017, pp. 103–110.
[58] M.-H. Laves, S. Ihler, T. Ortmaier, L.A. Kahrs, Quantifying the uncertainty of deep learning-based computer-aided diagnosis for patient safety, Curr. Dir. Biomed. Eng. 5 (2019) 223–226.
[59] J. Sicking, M. Akila, T. Wirtz, S. Houben, A. Fischer, Characteristics of Monte Carlo dropout in wide neural networks, in: ICML 2020 Workshop for Uncertainty and Robustness in Deep Learning, 2020.
[60] G. Loaiza-Ganem, J.P. Cunningham, The continuous Bernoulli: fixing a pervasive error in variational autoencoders, Adv. Neural Inf. Process. Syst. 32 (2019).
[61] E. Gordon-Rodriguez, G. Loaiza-Ganem, J.P. Cunningham, The continuous categorical: a novel simplex-valued exponential family, 37th Int. Conf. Mach. Learn. ICML PartF168147–5 (2020 2020) 3595–3605.
[62] A.B. Hanson, R.N. Lee, C. Vachet, I.J. Schwerdt, T. Tasdizen, L.W. McDonald, Quantifying impurity effects on the surface morphology of α-U_3O_8, Anal. Chem. 91 (2019) 10081–10087.
[63] A. Power, Y. Burda, H. Edwards, I. Babuschkin, Vedant, M. Openai, GROKKING: generalization beyond overfit-ting ON small algorithmic datasets, in: 1st Mathematical Reasoning in General Artificial Intelligence Workshop, ICLR, 2021, p. 2021.
[64] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, Adv. Comput. Vis. Pattern Recognit. 17 (2015) 189–209.
[65] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, D. Erhan, Domain separation networks, Adv. Neural Inf. Process. Syst. 29 (2016).
[66] Y. Wang, Q. Yao, J. Kwok, L.M. Ni, Generalizing from a few examples: a survey on few-shot learning, ACM Comput. Surv. 53 (2019).
[67] J. Snell, K. Swersky, R.S. Zemel, Prototypical networks for few-shot learning, Adv. Neural Inf. Process. Syst. *2017*– (2017) 4078–4088. December.
[68] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform. 3 (2016 32 2016) 119–131.
[69] M. Pirrung, A. Yankov, N. Hilliard, C.D. Corley, N. O'Brien, N.O. Hodas, SHARKZOR: human in the loop ML for user-defined image classification, in: International Conference on Intelligent User Interfaces, Proceedings, Association for Computing Machinery, 2018.