MDPI

Article

Deep Learning and AutoML for Automated Sorting of X-ray Scattering Patterns

Daniela Ushizima[1,2,3,†*], Flavio Araujo[4,†], Romuere Silva[4,†], Harinarayan Krishnan[1], Eric Roberts[1] and Alexander Hexemer[1]

[1] Lawrence Berkeley National Laboratory, Berkeley CA, USA
[2] University of California Berkeley, Berkeley CA, USA
[3] University of California San Francisco, San Francisco CA, USA
[4] Federal University of Piaui, Picos PI, Brazil
* Correspondence: dushizima@lbl.gov; Tel.: +1-510-486-4061 (D.U.)
† These authors contributed equally to this work.

Version February 10, 2021 submitted to J. Imaging

Abstract: X-ray scattering is an experimental technique that generates image patterns used to provide sub-nanometer structural information about materials, e.g. polymers. In order to capture the diffuse scattering patterns from disordered systems, distinct techniques have been developed, including small and wide angle X-ray scattering (SAXS and WAXS) and their respective surface-sensitive variation due to grazing-incidence known as GISAXS and GIWAXS. During a single, high-throughput experiment, these different techniques can be used interchangeably. Therefore, sorting and analyzing the acquired patterns require rapid processing seldom amenable to manual interaction. In this paper, we propose a set of computational tools ("camSortXS") exploiting machine learning to sort X-ray scattering patterns from large image datasets of energy critical materials: these materials were imaged using 4 techniques: SAXS, WAXS, GISAXS, and GIWAXS. The image datasets include thousands of patterns from real experiments performed by multiple users from a synchrotron-light beamline. Each scattering pattern undergoes featurization using 8 methods, making use of different architectures for deep learning, generating a total of 25 possible representations per pattern. The extracted features then serve as input to AutoML, an automated paradigm used to optimize 5 different classification models per featurization, which are then individually evaluated using 7 metrics of performance. Our analysis shows that different choices of convolutional neural networks architecture lead to sorting schemes with similar accuracy rates (over 97%) at diverse computing times, peaking with an accuracy of $99.11\% \pm 0.12\%$ These are promising results toward automating the sorting of patterns for metadata recovery, and enabling autonomous experiments for film design.

Keywords: Deep Learning; Classification; AutoML; X-ray Scattering

1. Introduction

X-rays have been broadly applied in the structure characterization of matter [1–5], for example, to inspect noncrystalline samples using X-ray scattering [6], which is an experimental technique that collects images by shining collimated X-ray beams through a material of interest. A two-dimensional detector captures the scattering patterns that result from interference between elements composing the sample structure [7]. The resulting scattering patterns reveal details about the physical structure and properties of targeted materials on the molecular and nano-scale, including crystalline unit cells [8], molecular and atomic spacing [9], nanostructure of human tissue [10], and surface roughness [11]. More recently, this imaging modality has been used to provide information about materials on a sub-nanometer scale, such as the stability and flexibility of sub-nanometer polymer-like wires used in

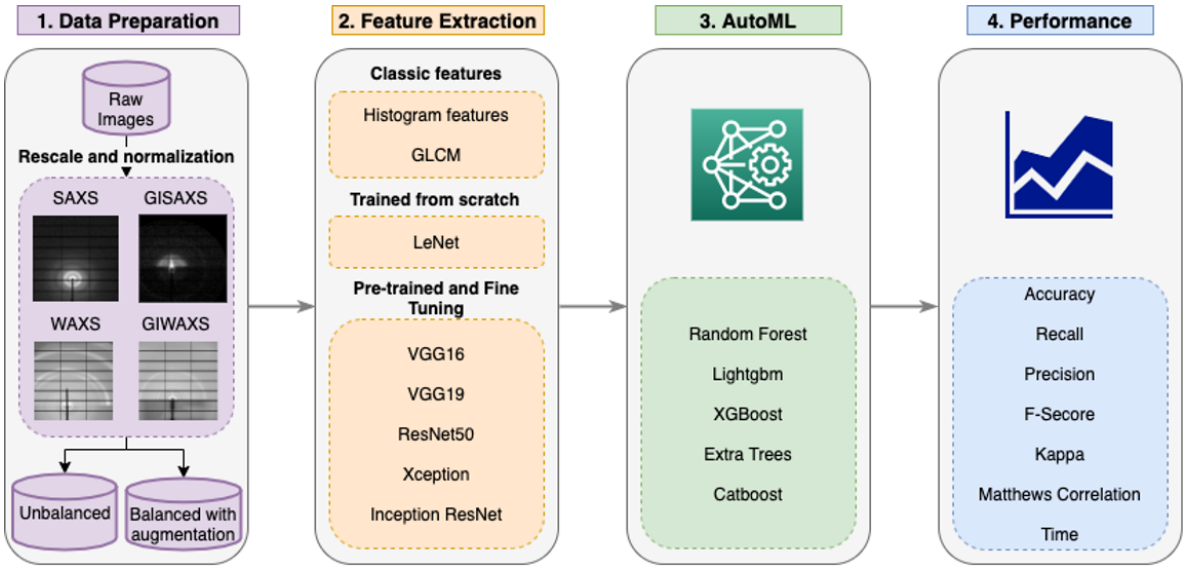


Figure 1. Diagram of camSortXS with main steps in sorting X-ray scattering data.

sub-nanometric material design [12,13] and the effects of sub-nanometer biomembrane thickness on intracellular transport [14,15] and cellular reactions to pathogens and drug designs [16].

In order to capture the diffuse scattering patterns from disordered systems, distinct techniques have been developed, including small- and wide-angle X-ray scattering, SAXS and WAXS, and their respective grazing-incidence variations known as GISAXS and GIWAXS, whose geometries differ due to the surface-sensitive variation. During a single high-throughput experiment [6,17], these different techniques can be used interchangeably. The data collection rates are challenging for manual interaction; real-time screening and analysis of the acquired patterns are required to keep pace. While human-interaction will continue to play an important role in the analysis process, the massive amounts of data, e.g. 10 petabytes per year [3], produced at some experimental facilities necessitates automation.

This paper proposes a set of algorithms for sorting scattering data from experiments into the different X-ray scattering techniques of SAXS, WAXS, GISAXS and GIWAXS, using deep learning and AutoML, an end-to-end automated machine learning tool [18], as illustrated in Figure 1. Our dataset includes thousands of patterns from real experiments performed by different users at the Lawrence Berkeley National Laboratory synchrotron light source. Each scattering pattern undergoes featurization using 8 methods for pattern description, including 6 deep learning architectures, as discussed in Section 3.1. Different combinations of training sets and feature extraction schemes are responsible for generating a total of 25 possible representations per pattern, once different training strategies are considered. These are input to AutoML (Section 3.2) that creates 5 different classification models for each of the featurizations, which are individually evaluated using 7 metrics of performance. Due to the combinatorial complexity of these different approaches, we use `dask` [19] to parallelize the classification using AutoML, as in Section 3.3. In order to improve interpretability of the results (Section 3.4), we show the Gradient-weighted Class Activation Mapping (Grad-CAM) for scattering patterns in each class. Finally, we summarize performance results in Section 4.

Our exploration shows that different choices of convolutional neural network (CNN) architectures lead to sorting schemes with similar accuracy rates (over 97%) at diverse computing times. It means that deeper neural networks performed as well as their shallower counterparts, but shallow ones are faster to retrain and require less labeled data. These are promising results and represent a first step toward the automation of sorting of patterns for metadata recovery and/or verification, and enabling autonomous [20] experiments for film design.

2. Background

2.1. Challenges in X-ray scattering

In energy-related research, thin-film polymeric materials play a major role in improving properties for batteries or organic photovoltaics (OPV). OPVs usually require thin layers of small molecules or polymers (usually below 100 nm thick) to optimize efficiency by balancing photon absorption versus the length that electrons and holes have to travel to reach an electrode. In addition, the structure inside the thin polymer layer requires an even smaller length scale on the order of just a few nanometers. The small structure allows the exciton created during the photon absorption, to diffuse to a charge separation area in the material [21].

These types of the materials are being investigate at the Advanced Light Source in Berkeley, one of brightest soft X-ray sources in the world, providing 40 beamlines with intense and coherent short-wavelength light for use in scientific experiments by researchers worldwide. As an example, grazing-incidence small angle X-ray scattering (GISAXS) enables the extraction of morphological information about both the thin film surface and the depth of embedded structures within the sample [22]. In tandem with extremely bright X-ray sources, the development of large and fast two-dimensional X-ray detectors provides an invaluable tool for in-situ time-resolved experiments, while creating a deluge in X-ray scattering data [23].

Roughly speaking, the choice of X-ray scattering experiment includes two different modes, namely small-angle scattering (S) and wide-angle diffraction (W). The approach depends on which structures, morphologies, or motifs are being targeted. For example, SAXS focuses on nanometer-size features while WAXS is more suitable for molecular and crystal structures of polymers. Most users usually employ more than one approach in order to understand the hierarchical structure of the compounds, e.g., combining SAXS and WAXS or GISAXS and WAXS. When the sample consists of a thin-film on a flat substrate, e.g. a silicon wafer, GISAXS or GIWAXS techniques are suitable choices to probe the surface as well as the internal structure of the material. Such experiments are essentially driven by the user, based on the properties of the sample; currently the manual mode of operation implies that the user will look at the recently acquired data and decide which scattering technique would work best next. In practice, a user adjusts the sample detector distance to a range that should suit the structure of the material. Initial experimental data are inspected by hand and a decision is made to continue and collect more data, or to change the experimental setup and adjust the sample detector distance and therefore the sold-angle of the collected data. The decision is driven by information content inside the initial data. For example, if the user notices several peaks (bright spots) just around the beam stop, but nothing else in the detector image, the user may choose a different small angle setup to better identify the peaks and extract more precise information for the experiment.

In order to begin the process of data reduction, it is important to identify the geometry of the experiment. However, when metadata is not available, the challenge is to automatically recover which technique was used. To exacerbate the problem of metadata recovery, multiple different techniques at the ALS have been intertwined for more than ten years, leading to a rich and invaluable, but somewhat chaotic data collection consisting of a mix of these image patterns.

To summarize, the main analysis tasks include:

- Indirect interpretation of physical properties from reciprocal-space data;
- The search for and recognition of features visually, such as peaks, arcs, rings, and rods;
- Verification of current X-ray scattering technique - if not appropriate, move detector; and
- Pattern measurements to infer sample properties, such as homogeneity, and decision about next steps in the experiment and analysis pipeline.

2.2. Machine Learning for Pattern Classification

Classification and ranking of scientific images [24–26], which includes the extraction of features and metrics from X-ray scattering patterns to infer the structure of probed materials, has become an

unfeasible manual effort across laboratories. Most previous works using deep learning to automate classification and sorting report results on simulated data [6,17,27,28], in part due to the lack of publicly-labeled datasets from experimental settings.

Options for generating simulated data include several open source programs, which model the underlying physics associated with the generation of X-ray scattering patterns, such as BornAgain [29] that is a research software used to simulate and fit grazing-incident small angle scattering (GISAS) reflexometry using both neutrons and X-rays, generally used in Europe [30]. Similarly, HipGISAXs [31] is a pattern simulation software, based on the Distorted Wave Born Approximation (DWBA) theory, and it is broadly used across American institutions to generate synthetic data.

Despite providing realistic data by computing scattering intensity patterns for a variety of sample orientations, morphologies, etc., using simulated data present some challenges: for example, the images are unrealistically "clean", often misleading to impressive accuracy rates, which may not generalize when transferring learning to real data [6]. Later work addressed the challenge of increasing variability of synthetic data by exploring different sources of noise, such as varying smear effects, Gaussian noise levels, Poisson shot noise, and multiple image resolutions [17]. These studies provided insights regarding data fluctuation and opportunities for pattern compaction, however the high accuracy from such models was still restricted to simulated data.

While a wide spectrum of problems have been tackled with CNNs, including application to categorization [32], to the best of the authors' knowledge, the work presented here is one of the first network designs for sorting X-ray scattering patterns into SAXS, WAXS, GISAXS and GIWAXS. Analogous data inputs were investigated by Zhong and Xu [33], who explored pattern representation to improve reconstruction from the reciprocal space; differently, and in contrast, we keep the scattering patterns in the reciprocal space, which are input to featurization using CNNs, followed by data compression. Instead, Zhu et. al [34] uses wavelet analysis to enhance spectral properties of inputs before classification with CNNs, however on a different science domain: fault diagnosis for hydraulic piston pump. Similarly to Fountsop et. al. [35], we explore architectures with different depths: one key difference is that they seek compressed models by considering data quantization for fast training and testing, while our proposal focuses on the searching for the best possible image representation (Section 3.1) to create robust models, and data compression (Section 3.2) of feature vectors that can have more than 4,000 descriptors.

To address these challenges, in the next section we describe our work on camSortXS, which is a set of essential methods to build digital infrastructure for automated identification of X-ray scattering imaging techniques: one important use is to apply the correct geometry change to the data when converting from pixels to q-space, which often requires knowledge about the sample detector distance and whether reflection or transmission geometry have been employed.

3. Materials and Methods

This paper creates deep learning models applied to a publicly-available dataset consisting of images of X-ray scattering patterns. The dataset contains ≈700 images per class without augmentation, with each image measuring 256x256 pixels and 64 KB in tiff file format. Access to materials for reproducibility are available on github [1].

There are 4 possible types of experimental techniques within the dataset, pictured below in Figure 2 and listed as follows:

- GISAXS: grazing-incidence small angle X-ray scattering;
- GIWAXS: grazing-incidence wide angle X-ray scattering;
- WAXS: wide angle X-ray scattering;
- SAXS: small angle X-ray scattering.

[1] Source code and documentation for camSortXS: http://bit.ly/camSortXS

Each of these datasets contain patterns coming from three or more users, as well as different types of materials. Neither the user nor the material type were disclosed or included as prior information to our experiments.

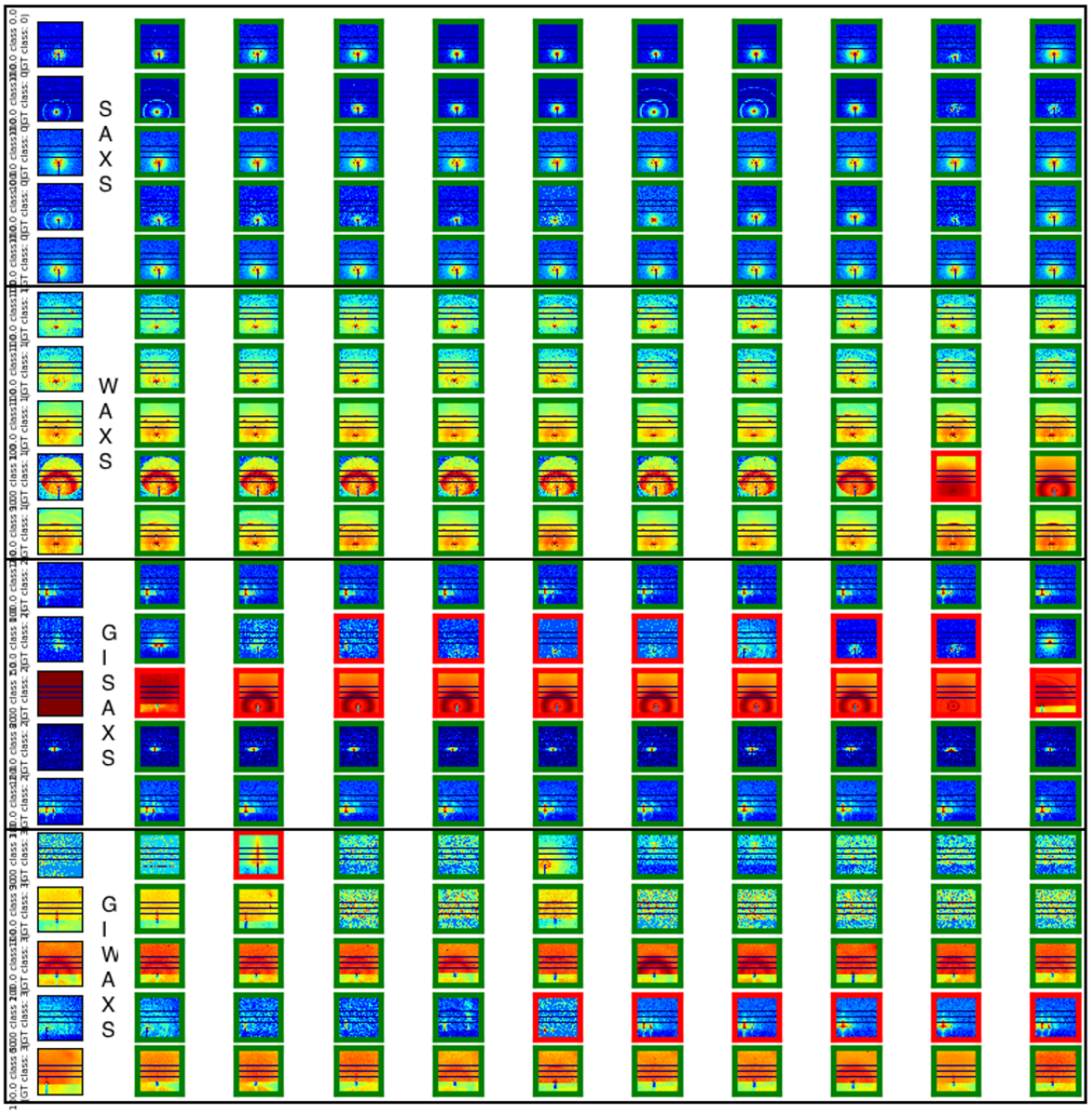


Figure 2. Randomly chosen X-ray scattering patterns (first column) are input to a similarity search with retrieval of top-10 matches: featurization is performed by LeNet and similarity by approximate KNN [6]. We emphasize successful cases with green borders, and failures with red borders.

3.1. Featurization: from Texture Descriptors to Deep Learning Fingerprints

3.1.1. Texture Descriptors

Gray Level Co-occurrence Matrix (GLCM)

This technique evaluates co-occurrences between pairs of pixels which follow a given pattern, distance, and angle [36]. The result is a co-occurrence matrix from which we compute the following image features: contrast, correlation, energy, homogeneity, angular second moment (ASM), and entropy.

GLCM has been used in cervical cell description [37], cytology [38,39], mammogram [40], and lung cancer detection [41].

Histogram Features (HF)

Histogram-based features consist of values for the average, entropy, energy, variance, skewness, kurtosis, and roughness computed from the image histogram. Despite its simplicity, histogram features have been broadly used for brain tumor detection [42], breast cancer detection [43], and vibration signal analysis for mechanical systems [44].

3.1.2. Deep Learning Fingerprints

Recent works typically employ three main different ways of using CNNs for pattern featurization with the intent of providing data fingerprints and classification. The first is a more time-consuming approach [45,46], in which the training is performed from scratch with a large set of data. The second is through transfer learning using pre-trained networks [47,48] trained in a large natural image database, such as ImageNet [49], which contains over 14 million images from 1,000 classes. Thus, the neural network can assimilate generic features, facilitating its application to small databases. The third is the fine-tuning technique that consists of continuing the hyperparameter adjustment training of a pre-trained network by using data from a new image set [6].

Yet another variation of each of these schemes involves the featurization portion of the CNN used as input to more sophisticated non-linear mappings; for example, one could substitute the several max-pooling operations and soft-max activation function for a support vector machine [50] to define a mapping from inputs to outputs, therefore unveiling a class prediction of an unseen object – this is the approached adopted in this paper, combined with the following featurization schemes: (a) LeNet [51] model trained from scratch, (b) VGG16, VGG19 [52], ResNet50 [53], Xception [54] and Inception Resnet [55] as pre-trained models on ImageNet, (c) their variants with fine-tuning. In both LeNet training and deep models fine-tuning, Stochastic Gradient Descent (SGD) was used for the optimization and the loss function was taken as the categorical cross-entropy, which have become straightforward approaches in multi-class classification problems. The next sections summarize important aspects of each CNN available within camSortXS, and how to obtain the feature vectors for each architecture.

Lenet

The Lenet architecture switches from fully connected to sparsely connected neurons, allowing feedback in real-time for most applications [45]. Due to its simplicity, it can manage recognition problems with a smaller number of examples during training in comparison with deeper CNNs, however it tends to be less accurate when dealing with complex recognition tasks. In this work, we used a Lenet with three convolutional layers with 64, 64, and 48 convolutional filters of size 3×3. There is a max-pooling with kernel size of 2×2 and a batch normalization layer after each convolutional layer. The classification network often has two fully connected layers with 192 and 64 neurons, with dropout value of 50% between them to generalize the learning process [56], followed by an output layer with k neurons whose response is modulated by a soft-max function. In this paper, we intercept the feature vector after the first fully connected layer.

VGG

The Visual Geometry Group (VGG) at the University of Oxford proposed the idea of decomposing big convolution kernels (11×11, 7×7 and 5×5) into smaller 3×3 convolution kernels [57]. According to the VGG architecture, multiple 3×3 convolutions stacked together can replicate larger convolution kernels, and there are more non-linear features (in terms of activation function) between them, even larger than with convolution. The difference between VGG16 and VGG19 CNNs reflects the number of layers: the first has 16, while the second has 19. In this architecture, the input size of the images is $224\times224\times3$ and the classification network has two fully connected layers with 4,096 neurons followed

by an output layer with soft-max function. The pattern representation follows a scheme similar to that described for LeNet, but here the feature vector is the output of the second fully connected layer.

ResNet50

ResNet50 is a 50-layer residual network. A common problem with deep neural networks is the repeated multiplication that occurs as the network is traversed deeper, resulting in an infinitely small gradient [58,59]. To address this problem, the residual module introduces a direct step from one layer to the next. Intuitively, these skip steps form a gradient highway where the computed gradients can directly affect the weights in the first layer, allowing updates to have a more meaningful effect [60]. As in VGG19, the input size of this architecture is 224×224×3. The ResNet50 does not have fully connected hidden layers. The sequence of convolutional and pooling layers results in 2,048 features which serve as input to the output layer with the soft-max function: we used the input to the soft-max as our 2,048-length feature vector.

Inception ResNet

Inception ResNet contains multiple sub-networks and a much deeper and wider architecture than the ResNet50. These hierarchical layers promote many levels of non-linearity needed for deeper, more extensive pattern classification, generally required for more complex inputs. This model is formed by inception modules combined with residual modules. An inception module is conceptually similar to convolutions (they are convolutional feature extractors), and empirically appears to be capable of learning richer representations with fewer parameters. The input size for this architecture is 299×299×3. As with ResNet50, the Inception ResNet does not have fully connected hidden layers. The sequence of convolutional and pooling layers results in 1,536 features before the soft-max function during model creation: we used these 1,536 values as features afterwards.

Xception

Xception is based on the Inception architecture, where separable convolutions replace the inception modules in depth. This process is named "depth-wise separable convolution", often referred as an inception module with a maximally large number of towers. As with Inception ResNet, the input size of this architecture is 299×299×3. The Xception does not have fully connected hidden layers, and the sequence of convolutional and pooling layers results in 2,048 features.

3.2. AutoML

Automated Machine Learning or AutoML frameworks offer strategies to generate robust data-driven models while minimizing numerous possible choices during key tasks, such as data preparation, model hyperparameter tuning, model selection, and model evaluation. Currently, there are several AutoML frameworks available for model deployment, including Auto Sklearn [61] and Auto Keras [62]. Our paper leverages an open-source AutoML package called PyCaret [18], which is a well-maintained framework that accelerates much of the exploration of possible models, with rapid testing of different statistical models, such as LightGBM, XGBoost, Random Forest, Extra trees, and Catboost, among others.

Before running such classification models, we perform dimensionality-reduction by running principal component analysis (PCA), enforcing all resulting feature vectors to have dimensionality equal to 5. PCA compresses the dataset onto a lower-dimensional feature subspace, which is empirically determined based on the realization that the top 5 components maintain most (over 95%) of the relevant information about the scattering patterns. Additional tests used fast independent component analysis (FastICA), whose results were similar to those using PCA, but at least twice as slow.

In Algorithm 1, each item of List_of_feature_vector_matrix is transformed using PCA. The next sections describe the classification models provided by the AutoML, which are responsible for statistical inference of different classes using the compressed pattern representation.

Algorithm 1: Python Pseudocode.

```
Input :
        Dask Configuration:
        - host config: "NERSC CORI".
        - nodes: "number of nodes".
        - cores: "number of cores".
        - conda_env: "Python Dask/Distributed + PyCaret environment".
begin
    Dask Client -> Launch Dask Scheduler and Workers using Dask Configuration;
    classic = [hist,glcm];
    unbalanced = [LeNet, VGG16, VGG19, Resnet, Xception, Inception];
    balanced = [LeNet, VGG16, VGG19, Resnet, Xception, Inception];
    augmented = [LeNet, VGG16, VGG19, Resnet, Xception, Inception];
    pre_trained = [VGG16, VGG19, Resnet, Xception, Inception];
    List_of_feature_vector_matrix = [classic, unbalanced, balanced, augmented,
     pre_trained];
    def dask_kernel(Feature_vector_matrix):
        List_of_models = [rf, lightgbm, xgboost, et, catboost];
        for each model in List_of_models do
            for each iteration of the model do
                perform operations: create(), tune(), update() -> new model;
            end
            write_stats_results_to_disk()
        end
        return None
    # Allocate and assign dask workers to resources;
    # Submit features to workers;
    Map List_of_feature_vector_matrix to remote workers;
    Dask Client [Parallel %Map] to Remote Cluster with (Dask Configuration,
     feature_vector_matrix);
end
```

3.2.1. LightGBM

LightGBM or Light Gradient Boosting Machine [63] implements an ensemble method that combines simple learners in a stage-wise approach to obtain a single prediction model. This gradient boosting framework uses tree-based learning algorithms to obtain high-quality prediction models. The main motivations to use this framework are the training speed, the higher efficiency associated with preserving suitable accuracy despite low memory usage, and the ability to handle large-scale data.

3.2.2. XGBoost

"Extreme Gradient Boosting" or simply XGBoost is another scalable tree-boosting framework invented by Tianqi Chen [37]. The term "Gradient Boosting" refers to previous work by Friedman [64] on greedy function approximation.

3.2.3. Random Forests (RF)

One of the motivations to use random forests is the high classification accuracy obtained through the creation of ensembles of trees [65]. RF drives the generation and selection of feature vectors that govern the growth of each tree in the ensemble, and it is known for preventing data overfitting, even with relatively small datasets.

3.2.4. Extra Trees (ET)

Extra trees, also know as Extremely Randomized Trees Classifiers [66], are ensemble methods similar to random forests on the basis that both consider a large number of decision trees to obtain the best prediction. The main motivation for using the ET method is its higher performance in the presence of noisy features, which is the case of most of scattering patterns.

3.2.5. Catboost

Catboost is a state-of-the-art open-sourced machine learning algorithm that implements gradient boosting on decision trees library [67]. This method generally yields competitive performance without the need for as much data for training as other methods. Our investigations have shown that this approach was the most time consuming of the ensemble methods adopted in this paper.

3.3. Parallelization and AutoML

Dask provides functions to scale python-based workflows in an easy way, distributing our computations across multiple cores. To create different prediction models using AutoML, we execute the `distributed` module to set up a scheduler with several worker processes. In order to speed up a portion of our computations, we created a specific routine (Algorithm 1) that runs on a high performance supercomputer: we used the National Energy Research Scientific Computing facility (NERSC) supercomputer Cori, which is a Cray XC40 with a peak performance of about 30 petaflops, comprised of 2,388 Intel Xeon "Haswell" processor nodes and 9,688 Intel Xeon Phi "Knight's Landing" (KNL) nodes.

3.4. Grad-CAM

Gradient-weighted Class Activation Mapping or Grad-CAM [68] is a technique used to visualize the activation maps of each class and explain which image parts or pixel sets contributed more to the final output of the CNN model. In other words, CAM verifies if specific image areas are associated with a particular class, allowing better model interpretability by the material scientists. It calculates the gradients of any target concept flowing into the final convolutional layer to produce a coarse localization map highlighting "important regions" associated with prediction.

To preview the X-ray scattering image classification and demonstrate Grad-Cam, Figure 3 shows heatmaps corresponding to the class activation maps of the four different scattering techniques. The maps show the well-known concepts expected in these imaging modalities, such as peaks, arcs, and rings, which is described in Section 4.

4. Results and Discussion

This section describes how to use the X-ray scattering data (Table 1) as input to the several architectures for feature extraction, including time for training the models, and fingerprinting each data sample (Table 2). Next, we show boxplots (Figure 4- 5), which summarize the comparisons among the different featurization schemes: notice that each interquartile (IQR) shows the metric dispersion over the 5 different AutoML classifiers. Finally, Table 3 highlights the best results in terms of six different metrics given each featurization scheme and the best AutoML classifier, as detailed below.

Table 1. Number of images for each class of generated datasets for training and testing of deep learning models: unbalanced (UB), balanced (BL) and balanced augmented (BLA).

Dataset name		WAXS	SAXS	GIWAXS	GISAXS
Train	UB	373	322	373	1756
	BL	322	322	322	322
	BLA	1288	1288	1288	1288
Test	UB	373	323	373	1756

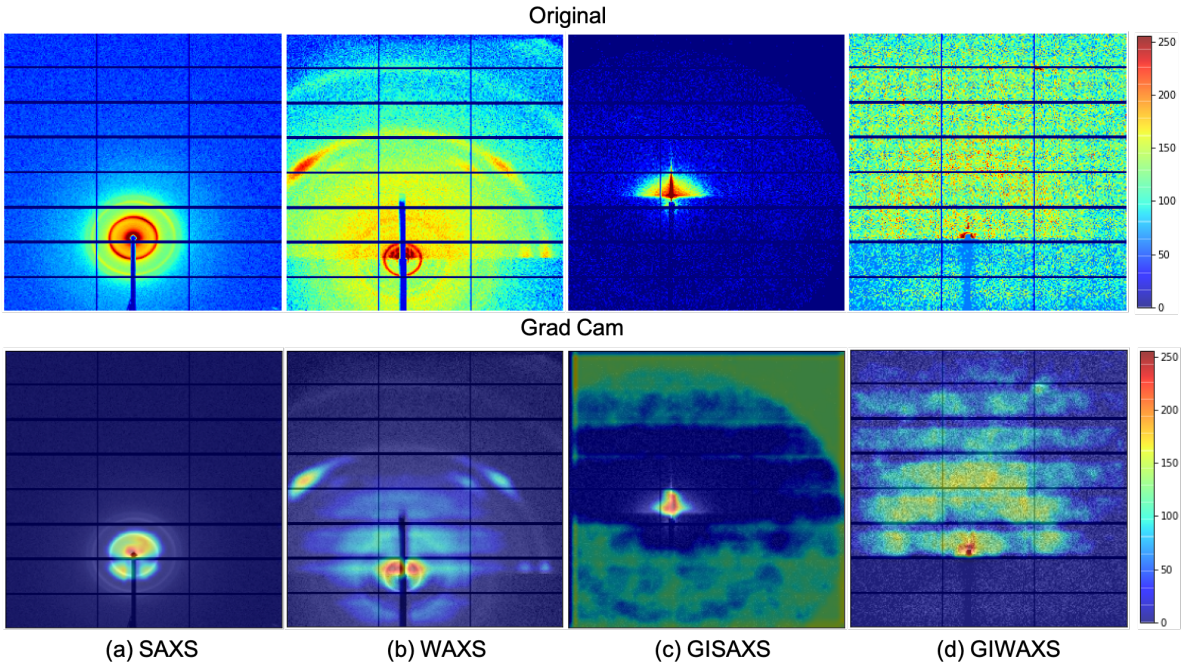


Figure 3. Visualization of the class activation maps using Grad-CAM: SAXS, WAXS, GISAXS and GIWAXS. The first row shows the original images, their respective Grad-CAM counterparts in the second row.

4.1. Dataset Creation and Augmentation

In order to compare the different classification approaches, we created models using X-ray scattering images as those shown in Figure 2. The database consists of 3,512 GISAXS, 746 GIWAXS, 746 WAXS, and 645 SAXS images, and we generate three distinct datasets from this database: (a) the Unbalanced (UB) dataset, which randomly splits samples such that 50% of the dataset is used for training and the other 50% for testing (number of images are listed in the first and last rows in Table 1); (b) the Balanced (BL) dataset, created to prevent training with uneven sizes, therefore selecting 322 images from each class, with the value 322 chosen because it evenly divides the maximum number of examples from the smallest class (second row in Table 1); (c) the Balanced Augmented (BL), an augmented training dataset, which increases the number of training samples to 1,288 by rotating each image in the BL dataset (third row in Table 1) with the angles: 0, 90, 180, and 270.

Table 2. Processing time for training and fine-tuning architectures.

Architecture	Train time (sec)			Feature extraction time (sec)
	Unbalanced	Balanced	Balanced augmentated	
LeNet	1283	587	2758	0.010
VGG16	1587	744	2822	0.020
VGG19	1814	846	3272	0.021
Resnet	2097	984	3872	0.022
Xception	2967	2251	5520	0.034
Inception Resnet	3832	2932	10546	0.052

4.2. Model Evaluation

These three datasets, UB, BL, and BLA are input to featurization as in Section 3.1, whose results are evaluated using 5 classifiers provided by the AutoML. Such evaluation considered cross-validation with k-folds ($k = 5$) to avoid biases on specific data subsets. Table 2 shows the processing time for training and fine-tuning during architectures evaluation.

Regarding the classic feature extraction methods, we computed the histogram-based features, which is a parameter-free approach, and the GLCM for distance values from 1 to 10, achieving the best results for distance equal to 1. The parameters used in the training of the LeNet are as follows: learning rate equal to 0.001, decay equal to 0.0001, number of epochs equal to 50, and batch size equal to 64. For fine-tuning of the deep models, the parameters are: learning rate = 0.0001, decay = 0.00001, number of epochs = 50 and batch size = 32. After the featurization using each of the CNN architectures in Section 3.1, we assessed the AutoML performance in terms of six broadly-accepted metrics available in literature for our multi-class problem: accuracy (Acc), recall (Rec), precision (Prec), F-Score (FS) [69], Kappa coefficient (κ) [70], and the Matthews correlation coefficient (MCC) [71], the later three recommended as more appropriate metrics for data exhibiting an imbalanced representation of classes [72–74].

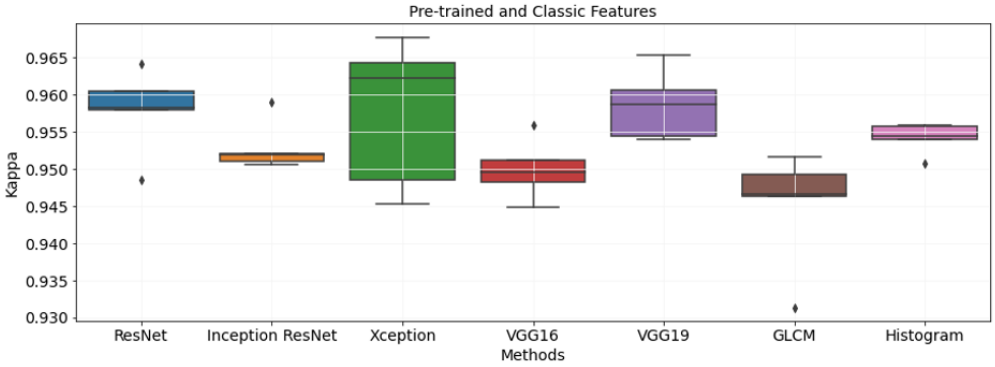


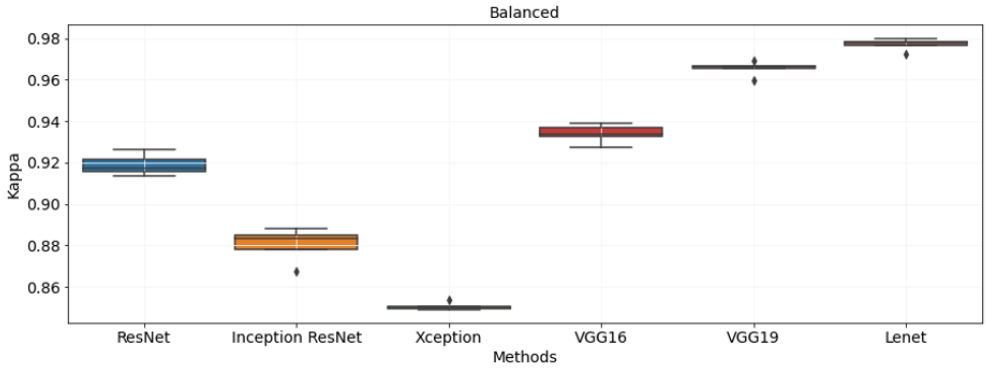
Figure 4. Comparison among different CNN architectures in terms of Kappa. All CNNs were pre-trained using ImageNet, but without fine-tuning. GLCM, and Histogram-based features bypass training.

Figure 4 presents the boxplots displaying the mean and variance of κ for the five deep architectures pre-trained with Imagenet, the histogram features, and GLCM. We focused on the κ coefficient, which is considered a vital metric to evaluate multiclass problems [70], and it allows easy interpretation, for example, a value less than 0 indicates non-agreement, between 0 and 0.20 as light agreement, between 0.21 and 0.40 as reasonable, between 0.41 and 0.60 as moderate, between 0.61 and 0.80 as substantial and between 0.81 and 1 as almost perfect agreement.

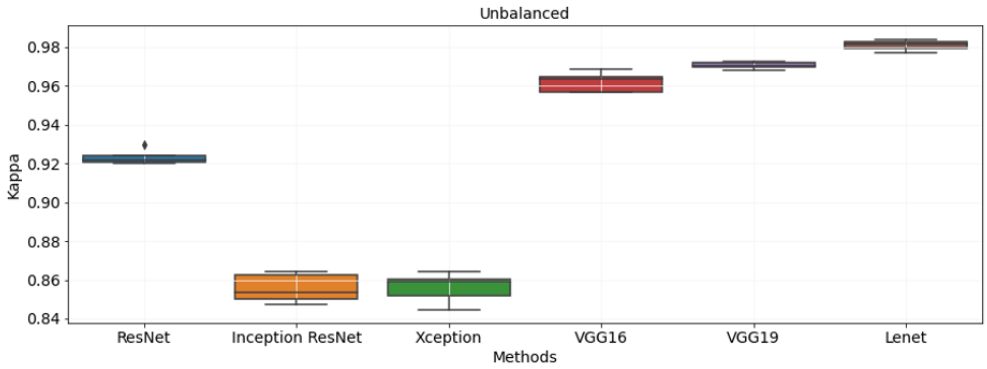
The IQR from each boxplots in Figure 4- 5 emphasizes the mean average κ over the 5 different AutoML classifiers. As an example of how to calculate each IQR, notice the blue IQR in Figure 4, which is the result of a 3-step computation: (a) calculate accuracy for the multiclass problem using ResNet associated to one AutoML classifier using k-fold cross-validation, therefore obtaining the average κ; (b) repeat the previous step for each one of the AutoML classifiers; (c) calculate the mean and variance over the average κ for each classifier.

Next, Figure 5 presents the boxplots summarizing the results when using fine-tuning for the deep architectures, and LeNet trained from scratch; notice that the κ dispersion indicates that they outperformed the pre-trained architectures in Figure 4. This was somewhat expected since the pre-trained architectures without fine-tuning extract features that are less customized to the problem at target. One of the hypotheses for such a divergence is that the ImageNet database, used in the pre-trained case, lacks images that follow patterns similar enough to those present in our experiments.

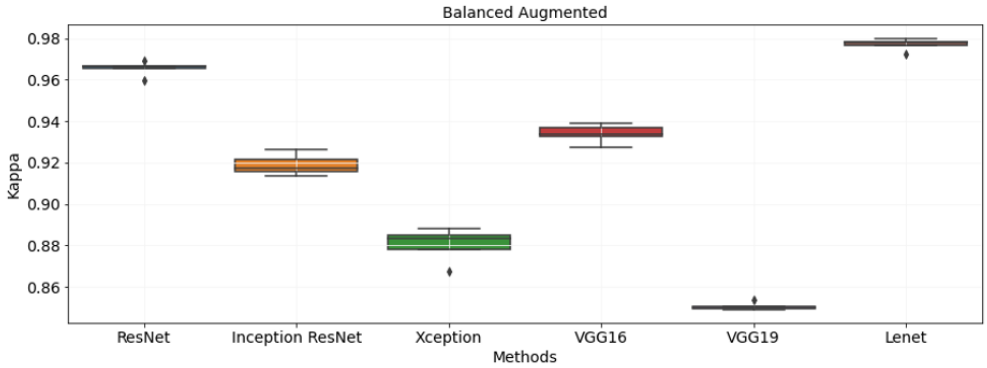
When evaluating the datasets used to train and fine-tune the CNNs, we observed similar κ coefficient across different architectures. Nevertheless, Figure 5 (a) and (c) indicate a lower dispersion in the boxplot when training with balanced sets. In addition, we show Table 3 that reports on the classification results following several metrics, including κ. Here, instead of calculating the mean across the 5 AutoML classifiers, we report metrics only for those with best performance given each featurization, in this case random forest (RF) and extra trees (ET). Overall, these results



(a) Balanced training set.



(b) Unbalanced training set.



(c) Balanced training set due to augmentation.

Figure 5. Comparison among CNN architectures in terms of κ: all CNNs used fine-tuning but LeNet.

Table 3. Result for each featurization, given the best classifiers (Clf) using different training sets: unbalanced (UB), balanced (BL), and augmented (BLA) datasets, and the pre-trained (PT) models after fine-tuning hyperparameters (*); best classifiers are random forest (RF) and extra trees (ET)

CNN	Train	Clf	Acc	Rec	Prec	FS	κ	MCC
HF	UB	RF	97.51±0.30	95.71±0.51	97.53±0.30	97.51±0.30	95.60±0.52	95.60±0.52
LeNet	UB	RF	99.10±0.13	98.80±0.13	99.11±0.12	99.10±0.13	98.40±0.23	98.41±0.23
VGG16*	BLA	ET	99.00±0.06	98.48±0.14	99.01±0.06	99.00±0.06	98.24±0.10	98.24±0.10
VGG19*	BLA	ET	99.01±0.11	98.45±0.25	99.02±0.11	99.01±0.11	98.25±0.20	98.25±0.19
Xception	PT	ET	98.17±0.10	97.10±0.29	98.21±0.11	98.17±0.10	96.78±0.18	96.78±0.18
ResNet	PT	ET	97.97±0.41	97.05±0.73	98.00±0.39	97.97±0.41	96.42±0.71	96.43±0.71
Inception	PT	ET	97.68±0.42	96.36±0.70	97.72±0.42	97.68±0.42	95.90±0.75	95.91±0.75

5. Conclusion and Future Work

This research work uses deep learning and AutoML algorithms encapsulated in camSortXS to explore thousands of patterns from real experiments emcompassing 4 X-ray scattering techniques: SAXS, WAXS, GISAXS, and GIWAXS. We were able to sort this new large labeled image collection of energy critical materials, and showed that different featurization options led to high mean average κ (over 0.9), which indicates almost perfect agreement across the 4 different classes and the different classifiers. By checking on 25 possible representations, classified using 5 different methods and evaluated using 7 metrics of performance, we achieve three main goals: (a) Characterize: Learn effective feature representations by transforming raw experimental data into compact signatures with the most relevant components; (b) Link: Compare, merge and integrate records from different experiments; (c) Screen: Autonomously sort experimental data which will inform upcoming inferential control of processes. Together, these efforts will exploit information from previously performed experiments to better guide next investigations.

We expect that camSortXS will allow for expansion of algorithmic and software design to build recommendation systems toward the construction of operators that can efficiently steer experiments. In this way, many calibration processes for obtaining successful experiments at the beamlines could be automated. Such intelligent guidance of complex experiments will accelerate tasks such as the acquisition of experimental data, as well as part of manufacturing processes, such as film production.

Author Contributions:

Conceptualization: D.U., F.A., R.S. and A.H.; methodology, D.U., F.A. and R.S.; software, D.U., F.A., R.S., H.K.; validation, F.A. and R.S.; formal analysis, D.U., F.A. and R.S.; investigation, D.U., F.A. and R.S.; resources, H.K. and A.H.; data curation, A.H. and E.R.; writing–original draft preparation, D.U., F.A. and R.S.; writing–review and editing, D.U., F.A., R.S., H.K., E.R. and A.H.; visualization, F.A. and R.S.; supervision, D.U.; project administration, D.U.; funding acquisition, D.U. and A.H.".

Funding: This research was partially funded by Center of Advanced Mathematics for Energy Research Applications (CAMERA), under the Contract No. DE-AC02-05CH11231 with the Advanced and Scientific Computing Research (ASCR) and Basic Energy Sciences (BES), both in the Office of Science of the U.S. Department of Energy. Further financial support from Brazilian entities: UFPI/PROPESQI 10/2018.

Acknowledgments: The X-ray scattering data was provided by Prof. Enrique Gomez (The Pennsylvania State University, Prof. Nitash P. Balsara (UC Berkeley), Dr. Ahmet Kusoglu (LBNL), Dr. Adam Weber (LBNL) under the Hydrogen and Fuel Cell Technologies Program of DOE EERE, Prof. Thomas P. Russell (University of Massachusetts Amherst) and Prof. Ting Xu (UC Berkeley) under the support of U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division, under Contract DE-AC02-05-CH11231 (Organic-Inorganic Nanocomposites KC3104).

Conflicts of Interest: The authors declare no conflict of interest. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Department of Energy or the University of California.

1. Freychet, G.; Kumar, D.; Pandolfi, R.J.; Naulleau, P.; Cordova, I.; Ercius, P.; Song, C.; Strzalka, J.; Hexemer, A. Estimation of Line Cross Sections Using Critical-Dimension Grazing-Incidence Small-Angle X-Ray Scattering. *Phys. Rev. Applied* **2019**, *12*, 044026. doi:10.1103/PhysRevApplied.12.044026.
2. Shyam, B.; Stone, K.H.; Bassiri, R.; Fejer, M.M.; Toney, M.F.; Mehta, A. Measurement and Modeling of Short and Medium Range Order in Amorphous Ta2O5 Thin Films. *Scientific Reports* **2016**, *6*, 2045–2322.
3. Schwarz, N.; Veseli, S.; Jarosz, D. Data Management at the Advanced Photon Source. *Synchrotron Radiation News* **2019**, *32*. doi:10.1080/08940886.2019.1608120.
4. Ushizima, D.; Xu, K.; Monteiro, P. Materials Data Science for Microstructural Characterization of Archaeological Concrete. *MRS Advancements - special issue: Materials Data Science* **2020**, pp. 1–14.
5. Xu, K.; Tremsin, A.S.; Li, J.; Ushizima, D.M.; Davy, C.A.; Bouterf, A.; Su, Y.T.; Marroccoli, M.; Mauro, A.M.; Osanna, M.; Telesca, A.; Monteiro, P.J. Microstructure and water absorption of ancient concrete from Pompeii: An integrated synchrotron microtomography and neutron radiography characterization. *Cement and Concrete Research* **2020**, pp. 106–282.
6. Araujo.; Silva.; Medeiros.; Parkinson.; Hexemer.; Carneiro.; Ushizima. Reverse image search for scientific data within and beyond the visible spectrum. *Expert Systems with Applications* **2018**, *109*, 35–48.
7. Guinier, A. *X-ray diffraction in crystals, imperfect crystals, and amorphous bodies*; Courier Corporation, 1994.
8. Yager, K.G.; Zhang, Y.; Lu, F.; Gang, O. Periodic lattices of arbitrary nano-objects: modeling and applications for self-assembled systems. *Journal of Applied Crystallography* **2014**, *47*, 118–129.
9. Lazzari, R.; Leroy, F.; Renaud, G. Grazing-incidence small-angle x-ray scattering from dense packing of islands on surfaces: Development of distorted wave Born approximation and correlation between particle sizes and spacing. *Physical Review B* **2007**, *76*, 125411.
10. Müller, B.; Deyhle, H.; Bradley, D.A.; Farquharson, M.; Schulz, G.; Müller-Gerbl, M.; Bunk, O. Nanomethods: scanning X-ray scattering: evaluating the nanostructure of human tissues. *European journal of nanomedicine* **2010**, *3*, 30–33.
11. Sinha, S.; Sirota, E.; Garoff, S.; Stanley, H. X-ray and neutron scattering from rough surfaces. *Physical Review B* **1988**, *38*, 2297.
12. Ni, B.; Shi, Y.; Wang, X. The Sub-Nanometer Scale as a New Focus in Nanoscience. *Advanced Materials* **2018**, *30*.
13. Liu, Q.; Wang, X. Polyoxometalate clusters: Sub-nanometer building blocks for construction of advanced materials. *Matter* **2020**, *2*, 816–841.
14. Betzig, E.; Patterson, G.H.; Sougrat, R.; Lindwasser, O.W.; Olenych, S.; Bonifacino, J.S.; Davidson, M.W.; Lippincott-Schwartz, J.; Hess, H.F. Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **2006**, *313*, 1642–1645.
15. Heberle, F.A.; Pabst, G. Complex biomembrane mimetics on the sub-nanometer scale. *Biophysical reviews* **2017**, *9*, 353–373.
16. Pabst, G.; Rappolt, M.; Amenitsch, H.; Laggner, P. Structural information from multilamellar liposomes at full hydration: full q-range fitting with high quality x-ray data. *Physical Review E* **2000**, *62*, 4000.
17. Liu, S.; Melton, C.N.; Venkatakrishnan, S.; Pandolfi, R.J.; Freychet, G.; Kumar, D.; Tang, H.; Hexemer, A.; Ushizima, D.M. Convolutional neural networks for grazing incidence x-ray scattering patterns: thin film structure identification. *MRS Communications* **2019**, pp. 1–7.
18. Ali, M. *PyCaret: An open source, low-code machine learning library in Python*, 2020. PyCaret version 2.1.
19. Dask developer team. Dask. https://dask.org/, 2020.
20. Ushizima, D.; Noack, M.; Hexemer, A. Data Science and Machine Learning for polymer films and beyond. Bulletin of the American Physical Society, 2020, Vol. 65, pp. 23–28.
21. Kozub, D.R.; Vakhshouri, K.; Orme, L.M.; Wang, C.; Hexemer, A.; Gomez, E.D. Polymer Crystallization of Partially Miscible Polythiophene/Fullerene Mixtures Controls Morphology. *Macromolecules* **2011**, *44*, 5722–5726. doi:10.1021/ma200855r.
22. Hexemer, A.; Müller-Buschbaum, P. Advanced grazing-incidence techniques for modern soft-matter materials analysis. *IUCrJ* **2015**, *2*, 106–125. doi:10.1107/S2052252514024178.
23. Santoro, G.; Yu, S. Grazing Incidence Small Angle X-Ray Scattering as a Tool for In- Situ Time-Resolved Studies. X-ray Scattering. IntechOpen, 2017.

24. Kiapour, M.H.; Yager, K.; Berg, A.C.; Berg, T.L. Materials discovery: Fine-grained classification of X-ray scattering images. IEEE Winter Conference on Applications of Computer Vision. IEEE, 2014, pp. 933–940.
25. Ushizima, D.; McCormick, M.; Parkinson, D. Accelerating Microstructural Analytics with Dask for Volumetric X-ray Images. 2020 IEEE/ACM 9th Workshop on Python for High-Performance and Scientific Computing (PyHPC) at Super Computing, 2020, pp. 41–48.
26. Araujo, F.H.; Silva, R.R.; Ushizima, D.M.; Rezende, M.T.; Carneiro, C.M.; Bianchi, A.G.C.; Medeiros, F.N. Deep learning for cell image segmentation and ranking. *Computerized Medical Imaging and Graphics* **2019**, *72*, 13 – 21.
27. Ushizima, D.; Bale, H.A.; Bethel, W.; Ercius, P.; Helms, B.; Krishnam, H.; Grinberg, L.; Haranczyk, M.; Macdowell, M.A.A.; Odziomek, K.; Parkinson, D.Y.; Perciano, T.; Ritchie, R.; Yang, C. IDEAL: Images across Domains, Experiments, Algorithms and Learning. *Journal of Minerals, Metals and Materials* **2016**.
28. Wang, B.; Yager, K.; Yu, D.; Hoai, M. X-ray scattering image classification using deep learning. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2017, pp. 697–704.
29. Pospelov, G.; Van Herck, W.; Burle, J.; Carmona Loaiza, J.M.; Durniak, C.; Fisher, J.M.; Ganeva, M.; Yurov, D.; Wuttke, J. *BornAgain*: software for simulating and fitting grazing-incidence small-angle scattering. *Journal of Applied Crystallography* **2020**, *53*, 262–276. doi:10.1107/S1600576719016789.
30. Herck, W.V.; Fisher, J.; Ganeva, M. Deep learning for X-ray or neutron scattering under grazing-incidence: extraction of distributions. *Materials Research Express* **2020**.
31. Chourou, S.T.; Sarje, A.; Li, X.S.; Chan, E.R.; Hexemer, A. *HipGISAXS*: a high-performance computing code for simulating grazing-incidence X-ray scattering data. *Journal of Applied Crystallography* **2013**, *46*, 1781–1795. doi:10.1107/S0021889813025843.
32. Pondenkandath, V.; Alberti, M.; Eichenberger, N.; Ingold, R.; Liwicki, M. Cross-Depicted Historical Motif Categorization and Retrieval with Deep Learning. *Journal of Imaging* **2020**, *6*. doi:10.3390/jimaging6070071.
33. Zhong, Y.; Xu, K. Contraction Integral Equation for Three-Dimensional Electromagnetic Inverse Scattering Problems. *Journal of Imaging* **2019**, *5*, 27.
34. Zhu, Y.; Li, G.; Wang, R.; Tang, S.; Su, H.; Cao, K. Intelligent Fault Diagnosis of Hydraulic Piston Pump Based on Wavelet Analysis and Improved AlexNet. *Sensors* **2021**, *21*. doi:10.3390/s21020549.
35. Fountsop, A.N.; Ebongue Kedieng Fendji, J.L.; Atemkeng, M. Deep Learning Models Compression for Agricultural Plants. *Applied Sciences* **2020**, *10*. doi:10.3390/app10196866.
36. Haralick, R.; Shanmugam, K.; Dinstein, I. Textural features for image classification. *IEEE Trans. Syst., Man, Cybern.* **1973**, *SMC-3*, 610–621.
37. Chen, Y.F.; Huang, P.C.; Lin, K.C.; Lin, H.H.; Wang, L.E.; Cheng, C.C.; Chen, T.P.; Chan, Y.K.; Chiang, J. Semi-Automatic Segmentation and Classification of Pap Smear Cells. *IEEE J. Biomed. Health Inform.* **2014**, *18*, 94–108.
38. Wang, H.; Feng, Y.; Sa, Y.; Lu, J.Q.; Ding, J.; Zhang, J.; Hu, X.H. Pattern recognition and classification of two cancer cell lines by diffraction imaging at multiple pixel distances. *Pattern Recognition* **2016**.
39. Alegro, M.; Theofilas, P.; Nguy, A.; Castruita, P.A.; Seeley, W.; Heinsen, H.; Ushizima, D.; Grinberg, L.T. Automating Cell Detection and Classification in Human Brain Fluorescent Microscopy Images Using Dictionary Learning and Sparse Coding. *Journal of Neuroscience Methods* **2017**, *282*, 20–33.
40. Kanadam, K.P.; Chereddy, S.R. Mammogram classification using sparse-ROI: A novel representation to arbitrary shaped masses. *Expert Systems with Applications* **2016**, *57*, 204 – 213.
41. Patil, S.; Udupi, V. Geometrical and texture features estimation of lung cancer and TB images using chest X-ray database. *International Journal of Biomedical Engineering and Technology* **2011**, *6*, 58–75.
42. Nabizadeh, N.; Kubat, M. Brain tumors detection and segmentation in MR images: Gabor wavelet vs. statistical features. *Computers & Electrical Engineering* **2015**, *45*, 286–301.
43. Saraswathi, D.; Srinivasan, E. An ensemble approach to diagnose breast cancer using fully complex-valued relaxation neural network classifier. *International Journal of Biomedical Engineering and Technology* **2014**, *15*, 243–260.
44. Jegadeeshwaran, R.; Sugumaran, V. Fault diagnosis of automobile hydraulic brake system using statistical features and support vector machines. *Mechanical Systems and Signal Processing* **2015**, *52–53*, 436 – 446.
45. Wu, Z.; Chen, H.; Lei, Y. Recognizing Non-Collaborative Radio Station Communication Behaviors Using an Ameliorated LeNet. *Sensors* **2020**, *20*. doi:10.3390/s20154320.

46. Lima, T.J.B.; Ushizima, D.; de Carvalho Filho, A.O.; de Araujo, F.H.D. Lung CT Screening With 3D Convolutional Neural Network Architectures. 2020 IEEE 17th International Symposium on Biomedical Imaging Workshops, 2020, pp. 1–4.
47. Tajbakhsh, N.; Shin, J.Y.; Gurudu, S.R.; Hurst, R.T.; Kendall, C.B.; Gotway, M.B.; Liang, J. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging* **2016**, *35*, 1299–1312.
48. Carvalho, E.D.; Filho, A.O.; Silva, R.R.; Araujo, F.H.; Diniz, J.O.; Silva, A.C.; Paiva, A.C.; Gattass, M. Breast cancer diagnosis from histopathological images using textural features and CBIR. *Artificial Intelligence in Medicine* **2020**, *105*, 101845.
49. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. CVPR09, 2009.
50. Tang, Y. Deep learning using linear support vector machines. Proc. of International Conference on Machine Learning, 2013.
51. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, pp. 2278–2324.
52. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
53. Chen, Z.; Xie, Z.; Zhang, W.; Xu, X. ResNet and Model Fusion for Automatic Spoofing Detection. INTERSPEECH, 2017, pp. 102–106.
54. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. *CoRR* **2016**, *abs/1610.02357*, [1610.02357].
55. Szegedy, C.; Ioffe, S.; Vanhoucke, V. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Computing Research Repository* **2016**, *abs/1602.07261*.
56. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **2014**, *15*, 1929–1958.
57. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
58. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* **1994**, *5*, 157–166.
59. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.
60. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
61. Feurer, M.; Klein, A.; Eggensperger, K.; Springenberg, J.; Blum, M.; Hutter, F. Efficient and Robust Automated Machine Learning. In *Advances in Neural Information Processing Systems 28*; Cortes, C.; Lawrence, N.D.; Lee, D.D.; Sugiyama, M.; Garnett, R., Eds.; Curran Associates, Inc., 2015; pp. 2962–2970.
62. Jin, H.; Song, Q.; Hu, X. Auto-Keras: An Efficient Neural Architecture Search System. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2019, pp. 1946–1956.
63. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 2017, pp. 3149–3157.
64. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* **2000**, *29*, 1189–1232.
65. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. doi:10.1023/A:1010933404324.
66. Maree, R.; Geurts, P.; Piater, J.; Wehenkel, L. Random subwindows for robust image classification. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, Vol. 1, pp. 34–40 vol. 1. doi:10.1109/CVPR.2005.287.
67. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. Proceedings of the 32nd International Conference on Neural Information Processing Systems; Curran Associates Inc.: Red Hook, NY, USA, 2018; NIPS'18, p. 6639–6649.

68. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* **2019**, *128*, 336–359.
69. Chinchor, N.; Sundheim, B.M. MUC-5 evaluation metrics. Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993, 1993.
70. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *biometrics* **1977**, *33*, 159–174.
71. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* **1975**, *405*, 442–451.
72. Sokolova, M.; Japkowicz, N.; Szpakowicz, S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. Australasian joint conference on artificial intelligence. Springer, 2006, pp. 1015–1021.
73. Gu, Q.; Zhu, L.; Cai, Z. Evaluation measures of the classification performance of imbalanced data sets. International symposium on intelligence computation and applications. Springer, 2009, pp. 461–471.
74. Bekkar, M.; Djemaa, H.K.; Alitouche, T.A. Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl* **2013**, *3*.