Materials Characterization 209 (2024) 113701

Contents lists available at ScienceDirect

Materials Characterization

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

Quantitative analysis of MoS_2 thin film micrographs with machine learning

Isaiah A. Moses [a], Wesley F. Reinhart [b,c,*]

[a] Materials Research Institute, The Pennsylvania State University, University Park, PA 16802, USA
[b] Department of Materials Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA
[c] Institute for Computational and Data Sciences, The Pennsylvania State University, University Park, PA 16802, USA

ARTICLE INFO

Keywords:
MoS_2 thin film
Morphological features
Machine learning
Transfer learning
Explainable AI

ABSTRACT

Isolating the features associated with different materials growth conditions is important to facilitate the tuning of these conditions for effective materials growth and characterization. This study presents machine learning models for classifying atomic force microscopy (AFM) images of thin film MoS_2 based on their growth temperatures. By employing nine different algorithms and leveraging transfer learning through a pretrained ResNet model, we identify an effective approach for accurately discerning the characteristics related to growth temperature within the AFM micrographs. Robust models with test accuracies of up to 70% were obtained, with the best performing algorithm being an end-to-end ResNet fine-tuned on our image domain. Class activation maps and occlusion attribution reveal that crystal quality and domain boundaries play crucial roles in classification, with models exhibiting the ability to identify latent features that humans could potentially miss. Overall, the models demonstrated high accuracy in identifying thin films grown at different temperatures despite limited and imbalanced training data as well as variation in growth parameters besides temperature, showing that our models and training protocols are suitable for this and similar predictive tasks for accelerated 2D materials characterization.

1. Introduction

Material properties are significantly influenced by conditions experienced during synthesis [1–5]. A systematic way of isolating the properties associated with different conditions is essential to enable the growth of materials with predefined properties on demand. We particularly seek approaches that eliminate intuition-based experimentation with different process variables, replacing them with data-driven approaches that are more efficient with time, effort, and other resources.

Several studies on thin film MoS_2 have revealed a number of growth parameters that determine the morphological features and properties of the grown materials. Instances include the evolution of the morphology of monolayer MoS_2 crystals grown by chemical vapor deposition (CVD) [6]. Domain shape variation from the triangular to hexagonal geometries has been shown to depend on the Mo:S ratio of the precursors [6]. Similarly, a MoS_2 domain shapes of mainly round, nearly round and hexagonal, truncated triangles, and triangles are observed at the temperatures of the MoO_3 precursor of 760 °C, 750 °C, 730 °C, and 710 °C, respectively [7].

The density and size of the domain have also been shown to decrease with temperature [7,8], with a random orientation of the MoS_2 domain associated with the growth temperature below 850°C [9] or at a much higher temperature [10]. In the former, the authors linked the phenomenon to the inability to achieve a thermodynamically stable state at the lower temperature, and in the latter, the inferred culprit is the step edges and step edge meanderings of sapphire substrate surface.

The grain size and crystal coverage of the the MoS_2 have also been shown to be tunable with the growth time [7]. The authors showed that the grain size increased when the growth time was increased from 20 min to 30 min. With the materials grown for 45 min, the grains merged to form a continuous MoS_2 [7]. Similarly, an increase in growth temperature [8] and O_2 flow rate [11] were shown to result in larger thin film crystal coverage.

In designing high throughput on-demand materials, deployment of data-based screening approaches have become more critical [12–17]. Data-driven approaches are being explored for materials characterization [18–22] and serve to provide greater clarity when searching the synthesis condition space compared to intuition-based experimentation [23–27]. With the use of the existing data consisting of the conditions and the corresponding materials properties, models that predict what

* Corresponding author at: Department of Materials Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA.
E-mail address: reinhart@psu.edu (W.F. Reinhart).

https://doi.org/10.1016/j.matchar.2024.113701
Received 11 October 2023; Received in revised form 19 January 2024; Accepted 24 January 2024
Available online 29 January 2024

conditions are necessary for a given properties can be developed. As observed, a number of these conditions play similar and intertwined roles in the materials properties. For instance, the time, temperature, and O_2 flow rate determine the MoS_2 thin film crystal coverage [7,8,11]. It will be interesting to use machine learning to isolate the distinct latent features associated with the different growth parameters. Additionally, identifying distinct latent features for these different growth parameters would result in the capability to classify material samples based on their growth conditions.

The Lifetime Sample Tracking (LiST) is a database hosted by the Penn State 2D Crystal Consortium (2DCC) facility, consisting of experimentally grown thin film transition metal chalcogenides materials, among others. Among the characterization methods used in the 2DCC and stored in LiST is Atomic Force Microscopy (AFM). AFM micrographs of MoS_2 thin films and their corresponding synthesis conditions are a set of data among other categories in LiST [10,28,29]. To accelerate the synthesis of MoS_2 with the desired properties, we deploy different machine learning (ML) models to classify AFM images of the material based on their growth temperature. The ultimate goal of the machine learning models is for the inverse design of materials, where the materials properties are tuned using the growth parameters. In essence, being able to predict the growth conditions from the morphology will enable the ability to determine the best growth conditions to achieve a hypothetical film morphology. This should accelerate the design and tuning of materials synthesis in the future. Despite the limited data available for the training, up to 71% test accuracy was obtained on the image classification. Most importantly, this study presents a simple approach that could help isolate underlying morphological features associated with different growth conditions for a broad range of materials, paving the way for rapid and cost-effective materials development.

2. Methods

2.1. Data preparation

Raw `spm` files of MoS_2 were retrieved from LiST [30]. These 262 AFM height maps were processed into greyscale images and either resized or randomly cropped to the common size of 224 × 224, depending on the augmentation method adopted, as discussed below. Training computer vision models on such a small dataset requires transfer learning, a common approach that utilizes CNN models pretrained on one image domain to extract features from a new image domain [31–33]. Many popular pretrained CNNs, such as the VGG [34], ResNet [35], and Inception model [36,37] architectures were trained on the ImageNet dataset [38]. ImageNet contains millions of color images of natural objects from thousands of categories. Using the size of the model architecture as the main basis for our choice, because of the small data volume in our characterization problem, the ResNet18 architecture pre-trained on ImageNet is used for transfer learning.

However, our data distribution is very different than the ImageNet data. To evaluate the effect of the pretraining domain, we consider pretraining on micrographs contained in the MicroNet dataset [39], which should be more similar to our image domain. The MicroNet dataset has been shown to give better performance on micrographs, indicating that the proximity of the two image domains should enhance the model performance [39]. We have therefore additionally used ResNet18 pretrained on the MicroNet dataset. This will enable us to compare how the same model architecture pretrained on different datasets perform on our characterization task. Features were extracted from the pretrained models for our shallow ML models. The pretrained convolutional models were also fine-tuned for the CNN model in our study (Fig. 1).

2.2. Data augmentation

The dataset consists of 262 instances of AFM height maps across 3 growth temperatures (Fig. 2). In addition to the limited data, there is a significant imbalance among the different classes with the 900 °C, 950 °C, and 1000 °C making up 11%, 50%, and 39% respectively (Table 1).

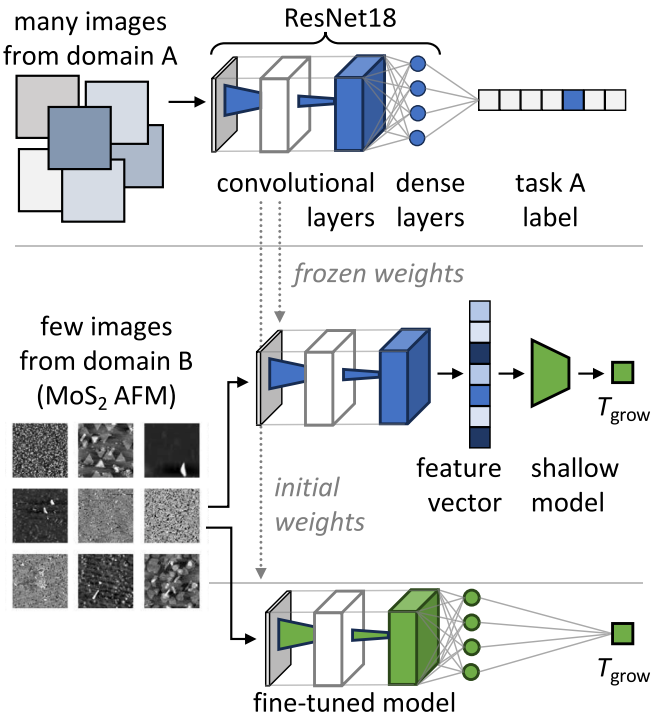

Fig. 1. An overview of the transfer learning approach. (top) A ResNet CNN model is trained on a different image domain with a large number of images. The task may be unrelated to the present task – all that matters is that convolutional filters are learned that can extract information (e.g., texture, color, shapes) from the images. (middle) The filters from the pretrained model can be used directly to extract relevant image features, which are interpreted in a supervised manner by a shallow model to predict a new label, such as the growth temperature. (bottom) Alternatively, the filters from the pretrained model can be fine-tuned on the new image domain to better capture relevant information for the task at hand.

The effect of limited and imbalanced data on the model performance can be partially mitigated with data augmentation approaches. Different data augmentation policies were therefore deployed to determine which method works best for our small, imbalanced dataset. The first was to randomly crop a common size of 224 × 224 from each of the original images. Multiple croppings were carried out, depending on the class of the image, in order to obtain a balanced representation of the different classes. This augmentation policy is termed *Aug1* (Table 1). Another augmentation policy examined is that developed by Cubuk, et al. [40], which we referred to as *Aug2* hereafter. The authors used a search algorithm to find the best policy, which is a combination of many sub-policies consisting of functions such as the translation, rotation, or shearing, and the probabilities and magnitudes with which the functions are applied, that give the best validation accuracy on a target dataset. Interestingly, they observed that the learned policy in a given dataset is transferable to another. We therefore examined how transferable the policy learned on ImageNet is to our present data domain. The third augmentation method used is a weighted random sampler or oversampling to correct the imbalance in the training set (*Aug3*). For *Aug4*, there is no biased augmentation applied to the data and only in CNN models do we have random rotations between 0 and 180°, horizontal and vertical flipping at 50% probability applied to the train and validation set on the fly.

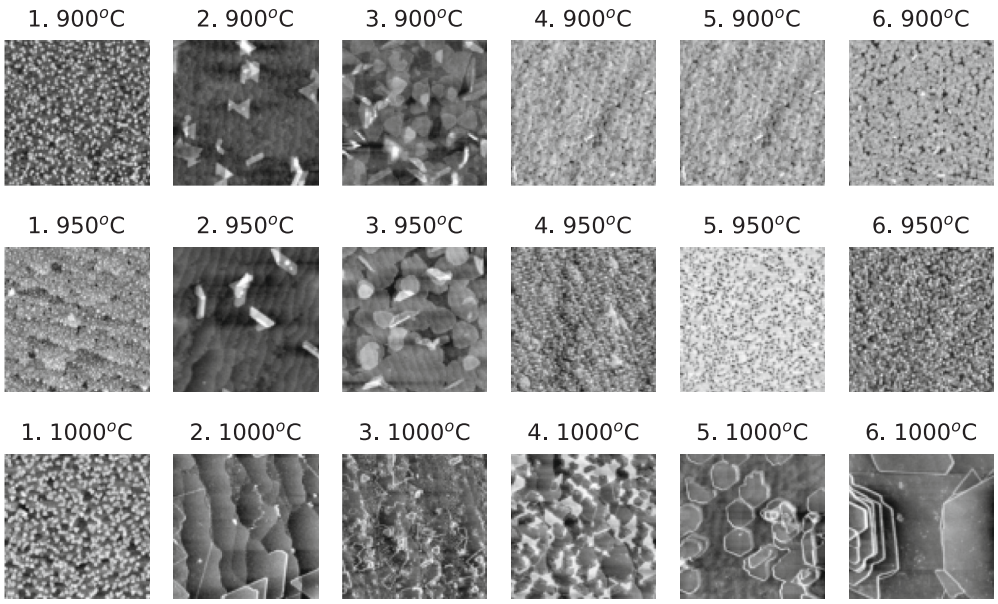

Fig. 2. Sample images from MoS_2 grown at 900, 950, and 1000 °C.

Table 1
Data augmentation policies and the corresponding data sets for the different classes, 900 °C, 950 °C, and 1000 °C. In *Aug1*, multiple random cropping of image size 224×224 is used to obtain balanced instances among the different classes, *Aug2* is augmentation policy learned on ImageNet [40], and in *Aug3* weighted random sampler and oversampling are used to correct the imbalance in train set for CNN and other models, respectively. *Aug4* is without biased augmentation. In CNN models, random rotations between 0 and 180°, horizontal and vertical flipping at 50% probability were additionally used on the train and validation set on the fly.

	900 °C			950 °C			1000 °C			Total
	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test	
Aug1	207	27	3	212	22	13	215	24	11	726
Aug2	207	27	3	208	24	13	210	23	11	734
Aug3	105	3	3	105	12	13	105	9	11	342
Aug4	23	3	3	105	12	13	83	9	11	262

2.3. Machine learning

A 10-fold cross-validation training scheme was used to train and evaluate the models, with 10 different models trained, one for each train-validation data splitting. 10% of the data was held out for testing while 90% was randomly split into 10 equal folds. A unique fold was used for the validation (to determine the performance for hyperparameter tuning using grid search) in each of the 10 models while the remaining 9 folds were used for training model parameters. The hyperparameters of the model with the best performance from the cross-validation procedure were selected for the production model. The 10 different training sets were then fitted independently into the production model and a held-out test set (not involved in the cross-validation procedure) was then used to evaluate the model performance in general.

Nine different ML models were considered: support vector classifier (SVC) [41,42], kernel ridge classifier (KRC) [43], radius neighbors classifier (RNN) [44], Gaussian process classifier (GPC) [45], k-nearest-neighbors classifier (KNN) [44], decision tree classifier (DTC) [46], gradient boost classifier (GBC) [47], multilayer perceptron (MLP) [48], and convolutional neural network (CNN) [49,50]. The shallow models were developed using the `scikit-learn` library version 1.2.2 [51] and the MLP and CNN were implemented in `pytorch`[52]. The optimized hyperparameters for the models are shown in the Supporting Information. The MLP model consists of 2 hidden layers, with each followed by a ReLU activation function. Additionally, we placed a drop out layer just before the output layer. For the CNN (fine-tuned pretraind ResNet model), the classifier outputs 3 classes for classification, but is replaced with a 100 nodes fully connected layer and an output layer for the regression models.

Using AFM images of 2D MoS_2 grown with MOCVD, we developed models to predict the growth temperature (one of 900 °C, 950 °C, or 1000 °C). We considered framing the task in several different ways to evaluate the efficacy of each: nominal classification, ordinal classification, and regression. Here nominal classification means the three growth temperatures were considered as distinct classes with no ordering. Unless otherwise specified, results are for nominal classifiers.

For ordinal classification, we implement NNRank [53] to account for ordering within the classes; the targets 900, 950, and 1000 °C are transformed into the vectors $[1,0,0]$, $[1,1,0]$, and $[1,1,1]$, respectively. At inference time, a threshold of > 0.5 is applied to the prediction and the values are counted from left to right, which provides the class label. Note that this scheme is only applied to the NN models (MLP and CNN). Finally, we perform regression by simply using the growth temperatures as continuous labels and evaluating the MSE. The class labels are obtained by binning the predicted growth temperature (e.g., $925 - 975$°C belongs to the 950°C class).

3. Results and discussion

3.1. Depth of image features

Given the poor performance observed from the randomly initialized weights of the CNN models (Supporting Information), we deployed transfer learning for the task. We first determined the best location in the pretrained model from which to extract image features for our models. Different portions ("blocks") of the ResNet were considered, providing

filters with different levels of abstraction. Due to the large number of channels in the pretrained model (see Table 2), Principal Component Analysis (PCA) was applied to reduce the dimension of input features to the shallow models, ideally reducing overfitting and thus improving predictive performance [14,54,55]. Cumulative explained variance thresholds of 85% and 99% were used to determine the number of features to keep for inference. We found that within a block, using fewer features gave better performance in 9 of 12 cases despite lower explained variance, likely because we had few training data compared to the size of the feature vectors. Depending on the model architecture and number of features used, minimal or significant deviations in model performance could be obtained from any of the ResNet blocks (e.g., 66%, 64%, 77%, and 78% accuracy from subsequent blocks, with typical standard deviation ±6%).

Separately, the dense layers of the pretrained model was replaced with new ones with fewer neurons and then fine-tuned on our training data. The model parameters are the same as the CNN classifier described in the preceeding section. The model fine-tuned on the ImageNet and the MicroNet gave a train accuracy of 88 and 76%, respectively, and a validation accuracy of and 73 and 70%, respectively. Finally, 100 features were extracted from the first dense layer. Note that we have compared the performance of this fine-tuned dense layer against those extracted from the pretrained blocks. This was an intentional choice to evaluate the degree to which fine-tuning was needed to achieve good performance in this task.

The performance of the selected classifiers on the different features shows that the features extracted from the fine-tuned dense layer gives the best performance overall, with 80%, 71%, and 70% accuracy using SVC, KRC, and RNN, respectively. Training the dense layer on a pretrained convolutional backbone might therefore be a better approach for extracting a low-dimensional image feature vector compared to PCA. These tuned features are therefore used in all of the following analysis.

3.2. Data augmentation

We then evaluated the effect of different data augmentation policies using the SVC, KNN, and CNN models (Table 1 and Fig. 3). In addition to the accuracies of the models, F1 score was used to evaluate the different augmentation policies. This is to ensure that the data imbalance is accounted for in comparing their performances. It is observed that both the accuracy and F1 score gave similar performance trend (Fig. 3 and Fig. S2). Significantly worse performances are obtained with *Aug1* and *Aug2*, especially in the shallow models, compared to *Aug3* and *Aug4*. Meanwhile, the performance observed between *Aug3* and *Aug4* is statistically indistinguishable.

The poor performance observed in the *Aug1* and *Aug2* might be related to the properties of the images learned by the models. While in the case of the natural images, activation of different classes are typically associated with unique features of the classes [56–58], the class activation in the models for the different synthesis conditions will be more likely due to differences in magnitude of the same feature, such as the domain size and thickness [3,4]. These relevant features of the AFM images may be disrupted by shearing, zooming, and resizing associated with *Aug1*, and the features location in the image might be omitted due to the cropping in *Aug2*.

Although *Aug3* and *Aug4* present about the same accuracy, *Aug3* has the desirable property of oversampling less represented classes. This should help mitigate systematic error related to class imbalance, a feature which is typical of distributions in materials synthesis, especially when exploring different growth conditions (e.g., poorly performing conditions will probably be undersampled). Therefore, the *Aug3* augmentation policy is selected for the rest of this study.

3.3. Pretraining domain

The previous two sections on the feature extraction and the data augmentation are initial verifications. Therefore, only 3 machine learning models were explored. We next seek to quantify how transfer learning from the ResNet18 model pretrained on the ImageNet data domain compares with the same model architecture pretrained on the seemingly more relevant MicroNet data domain. We therefore compared the performance of each pretrained model on the same nominal classification task across a wide range of predictive model types. In these experiments, we used the fine-tuned features from Table 2 in all cases except CNN, which was simply fine-tuned in an end-to-end manner using the original ResNet18 architecture (i.e., with a three-way classification layer attached to the end in place of the original classification layer). Based on the results shown in Table 3, the ImageNet model gives conclusively better performance than MicroNet, with at least 9% improvement and up to 32% improvement in the case of MLP (compared to a typical uncertainty of about 6%).

While standard deviations for individual observations are high, the fact that none of the nine model types shows a negative difference is compelling, especially because MicroNet was trained on greyscale micrographs of materials while ImageNet was trained on color images of macroscale objects. Previous work has suggested that ImageNet relies more heavily on texture rather than shape [59], while MicroNet has been primarily tested for segmentation tasks. We speculate that this focus on texture gives ImageNet filters that can be used for identifying distinguishing textures in the AFM height maps. The results presented here suggest that ImageNet may be surprisingly well suited for out-of-domain materials characterization data whose information content is primarily texture. All following results are based on transfer learning from the ImageNet pretraining since its features are strictly superior to MicroNet.

3.4. Model performance

We next investigate the performance of different algorithms in greater detail. As before, we rely on the features extracted from the fine-tuning procedure above, with additional shallow models trained on these static feature vectors of each image. The CNN model is the one exception to this, as it uses the original ResNet18 architecture and is fine-tuned on this task without modification to feature size. The classification accuracy across 10 different model instances of each type is shown in Fig. 4. Overfitting is observed across all model types, with

Table 2
Validation accuracy (in %) based on the features extracted from the different layers of the pretrained model (ResNet18 pretrained on ImageNet). Channels is the total size of raw feature vectors extracted from each block of the ResNet. PCA was applied to these channels, and then cumulative explained variance (CEV) of the components from PCA was used to determine the size of the input features for the listed shallow models. Separately, the dense layers of the pretrained model were replaced with fewer neurons and fine-tuned (last column).

	Block 2		Block 3		Block 4		Pooling		Fine-tuned
Channels	100,352		50,176		25,088		512		100
CEV	85%	99%	85%	99%	85%	99%	85%	99%	–
Features	190	235	156	235	94	219	28	142	100
SVC	66±6	59±7	64±5	62±8	77±6	58±5	78±5	71±6	80±7
KRC	45±5	57±4	48±11	55±5	58±11	52±5	57±7	58±12	71±7
RNN	21±4	15±2	35±11	15±3	42±9	20±5	57±7	39±7	70±9

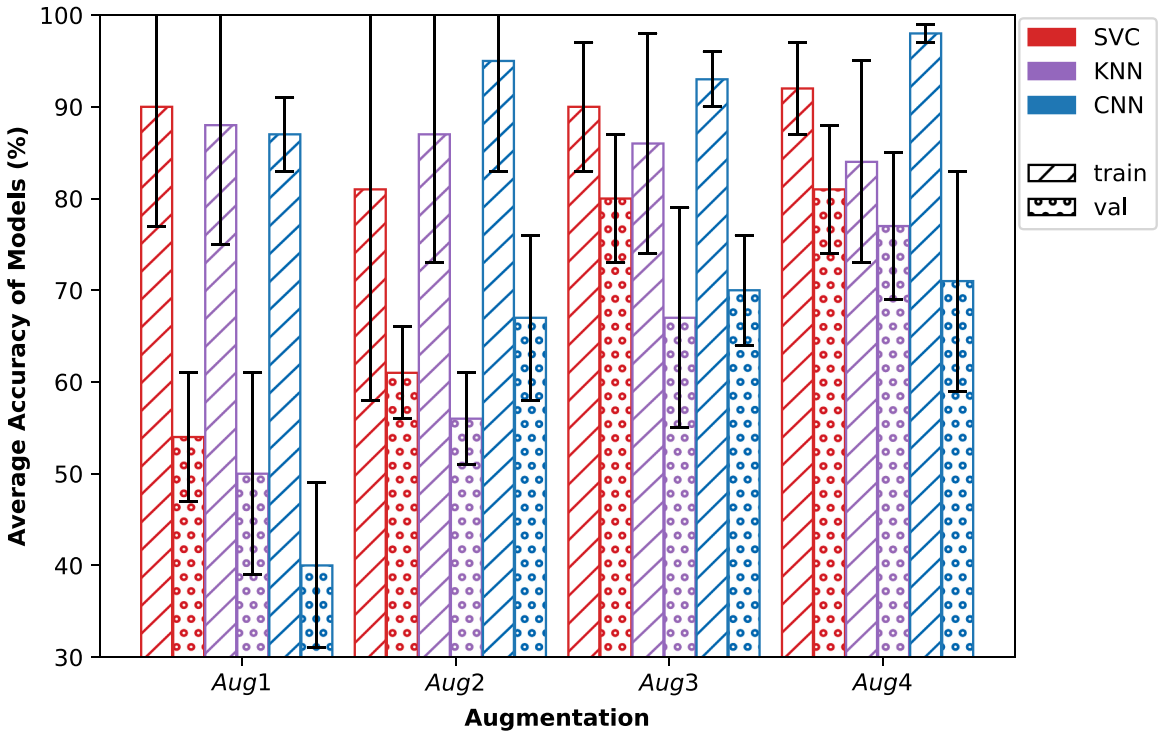

Fig. 3. Accuracy obtained from different augmentation policies across three different model types. Bars report averages over 10 folds, while error bars indicate standard deviation. Some models were trained with increased data size to have a balanced classes using different augmentation approaches, as indicated in Table 1.

Table 3
Validation accuracy (in %) over 10 folds obtained for the feature extraction (shallow and MLP models) or end-to-end learning (CNN) with ResNet18 pretrained on ImageNet and MicroNet. Values are reported as mean ± standard deviation. Difference is the fractional change in the average score between MicroNet and ImageNet. Best model performance in each row is shown in bold.

Models	SVC	KRC	RNN	GPC	KNN	DTC	GBC	MLP	CNN
MicroNet	**73±6**	65±10	63±9	52±12	59±10	71±9	71±9	65±8	63±8
ImageNet	80±7	71±7	70±9	59±10	67±12	78±4	78±11	**86±6**	70±6
Difference	+10%	+9%	+11%	+13%	+14%	+10%	+10%	+32%	+11%

training performance over 90% being typical, while validation typically only reaches around 60–85%. The greatest overfitting, in terms of the gap between train and validation performance, is seen in KRC and GPC, while SVC, DTC, and MLP exhibit the least. The best performing models in terms of validation performance is the MLP, with SVC coming in second but exhibiting training and validation scores one standard deviation below the MLP.

To understand how well the models can generalize to classifying images outside of the training data, we additionally examine their performance on a held-out test set (i.e., not used for training or hyperparameter selection). In this regard, MLP again showed the highest accuracy, with GBC and GPC appearing within one standard deviation. It is reassuring to see that MLP gave the highest scores in both validaiton and testing, inspiring confidence in its performance overall.

To understand the model performance on the different growth temperatures in greater detail, and particularly to check if the underrepresented classes have comparable accuracy, average confusion matrices of the held-out test set on 10 models are reported in Fig. 5. To focus the discussion, only the highly performant GBC and MLP models and the end-to-end CNN are examined in this regard. It is notable that the performance within each class does not vary substantially between different model types, as the overall accuracy are similar. For instance, the GBC, MLP, and CNN predict about the same number of samples grown at 950 °C and 1000 °C correctly (about 70% and 75% respectively). The samples grown at 900 °C are found to have the lowest in-class accuracy. This seems to be partially an artifact of underrepresentation in the test set; as shown in Table 1, classes are significantly imbalanced in the data, with the 900 °C classes having the least number of samples.

There is also some consistency among the models in misclassifying the 900 °C as 950 °C and not as 1000 °C. Similarly, 1000 °C is rarely misclassified as 900 °C. On the contrary, 950 °C is about equally likely to be misclassified as 900 °C as it is as 1000 °C by the MLP and CNN. This seems to suggest that the proximity of the growth temperature, which is expected to be reflected in the image features, makes it more likely for the model to group them together. Recall that this is for nominal classification, so this proximity is not reflected in the loss function. This could imply a fundamental bias in the data where the image feature learned by the models for a given temperature are more similar to that for the adjacent temperatures.

To further understand the classification fidelity of our models, we examine images that are correctly and incorrectly classified by the CNN in Fig. 6. Visual inspection suggests significantly different image features among the same growth temperature, demonstrating how difficult this classification task is. Some images grown at 950 °C show larger crystal domains typically associated with 1000 °C. Conversely, some images grown at 1000 °C show poor crystal formation and very small domain sizes exhibited mostly by the 900 °C growth temperature. Therefore, these wrongly classified images may be exceptional among the target class and would likely confuse even a human expert. However,

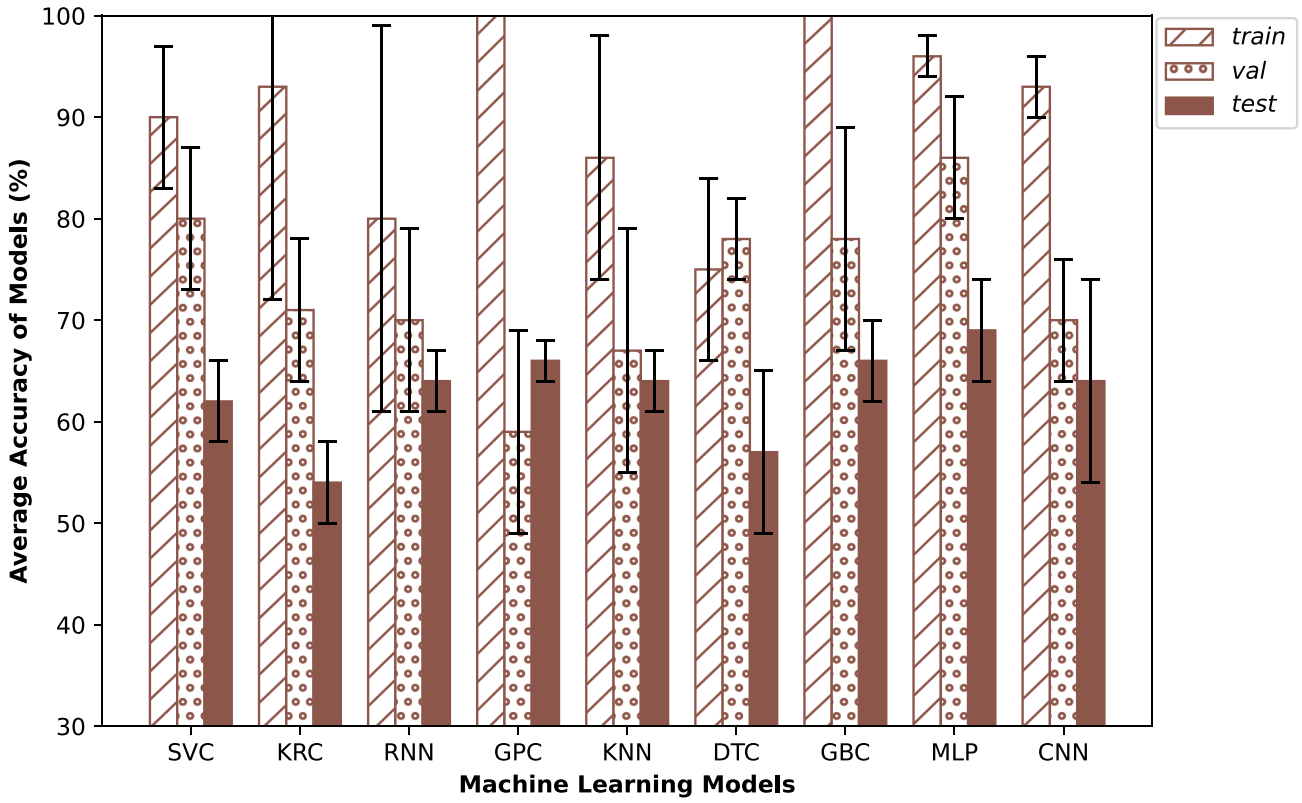

Fig. 4. The average train, validation (*val*), and test accuracy over 10 models for the different algorithms. The train-validation data was randomly split into 10 equal folds. A unique fold was used for the validation in each of the 10 models while the remaining 9 folds were used for training model parameters. Hyperparameters were tuned to obtain a trained model for each of the 10 splits. The trained models were tested with the test set.

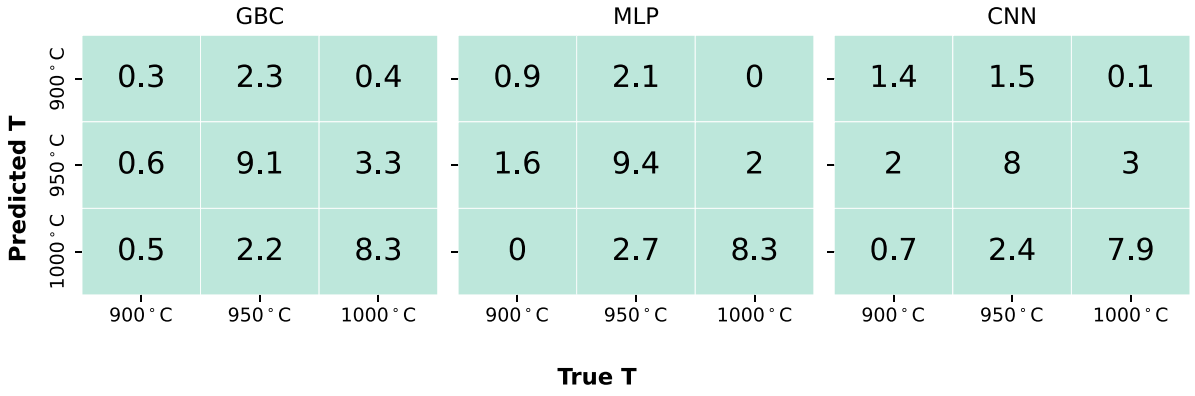

Fig. 5. The average confusion matrix for the test set predictions of production models trained on the 10 folds train data. Values indicate the number of samples in each bin. This is based on nominal classification.

they offer some preliminary insight into which features the classifier attributes to each growth temperature.

More fundamentally, other growth variables are not entirely fixed across the samples. For instance, the growth time varies significantly among the different samples (Fig. 6(b)). While the least growth time in the test set is as low as about 100 s, some samples are grown at much longer time, with up to 1650 s. Also, while most of the samples are grown on c-plane sapphire substrate, we also have some that are grown on A- and M-plane sapphire (Fig. 6(c)). These inconsistent growth parameters might have accounted for the significant differences observed among the samples grown at the same temperature and might have also resulted in some classification errors (e.g images 2, 5, 15, and 18). However, we do not observe any obvious trend in these growth parameters that leads to consistent misclassification, once again demonstrating how challenging this classification task is.

3.5. Ordinality

The preceding results were all based on nominal classification, without any notion of ordering. However, the classes consisting of the growth temperatures would appear to be ordered due to their continuous nature (i.e., ranging from 900 to 1000 °C). We therefore further quantify the effect of ordinal treatment of the class labels on model accuracy. In accounting for ordinality in shallow (i.e., non-NN-based) models, we adopted a simple approach based on training a regressor and then binning the results into classes. For the NN-based models, we further implemented the NNrank ordinal classification scheme. The results of this study are given in Table 4.

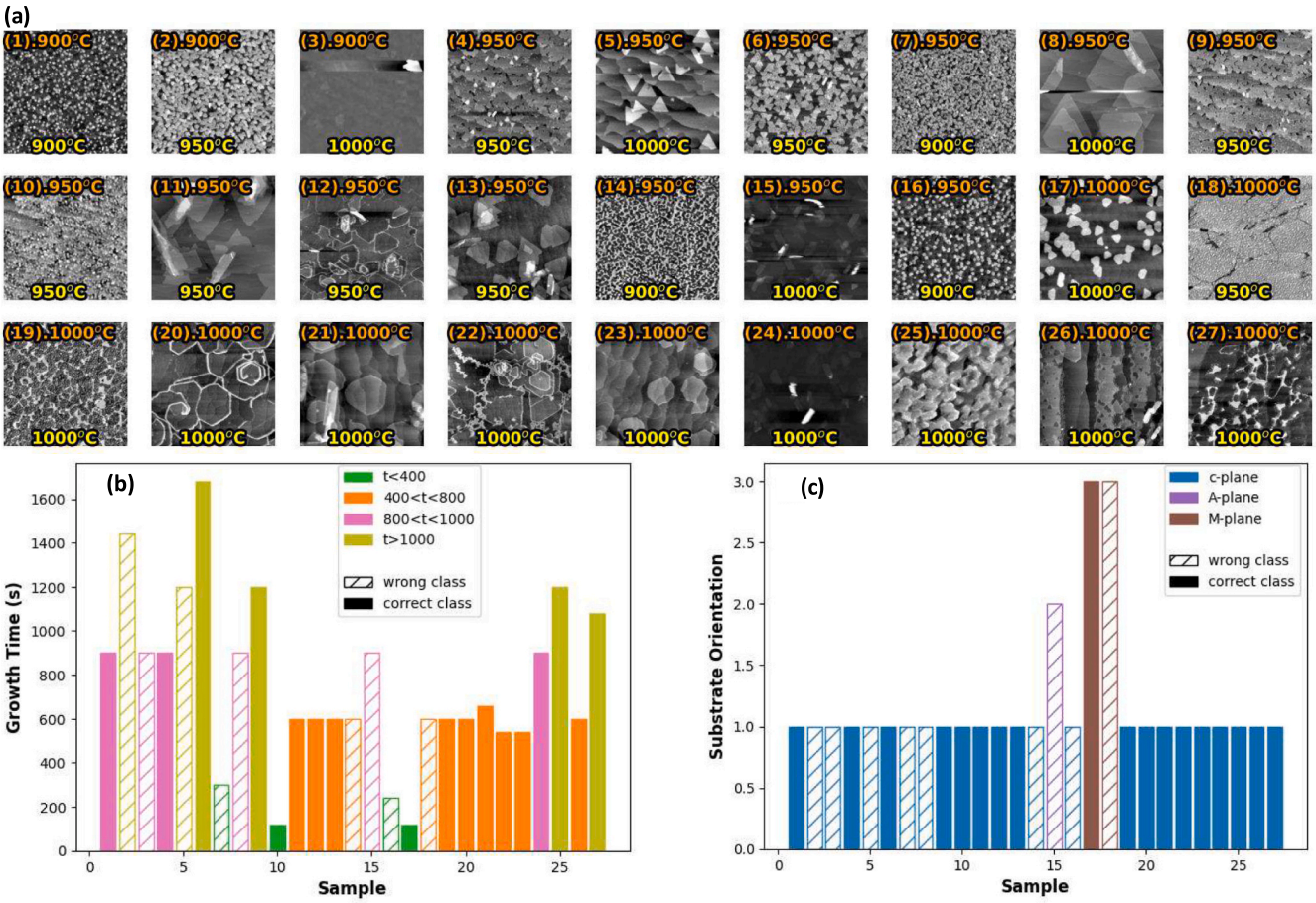

Fig. 6. Samples grown at 900 °C (1–3), 950 °C (4–16), and 1000 °C (17–27) in the test set. The predicted class by end-to-end CNN is shown at the bottom (yellow) for each image. (b) and (c) are the samples with their growth time and substrate orientation, respectively. The AFM # in (a) corresponds to the sample # in (b) and (c). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Performance of nominal and ordinal treatment of class labels, expressed as accuracy on held-out test data in % for classification and °C for regression. Best model performance in each row is shown in bold.

Models	SVM	KR	RNN	GP	KNN	DT	GB	MLP	CNN
Classification (%)	62±4	54±4	64±3	66±2	64±3	57±8	66±4	**69±5**	64±10
NNRank (%)	–	–	–	–	–	–	–	**71±4**	68±3
Regression (%)	50±8	60±5	64±4	48±0	58±6	54±6	**64±7**	42±8	61±7
RMSE (°C)	31±2	**26±1**	–	36±9	32±3	38±5	28±3	62±8	34±4

While results vary for each model type, some general trends emerge. Accounting for ordinality in model training leads to improvement in the test accuracy in only one of the shallow models (KR), but matches or degrades the performance for all others. Most of these are statistically indistinguishable, with only SVM, GP, and MLP exhibiting significant decreases. Overall, nominal classification gave superior performance over regression, with the top performing shallow models GP and GB giving 66% accuracy.

For the NN models, MLP outperformed CNN overall, with statistically indistinguishable accuracy using nominal classification and ordinal classification. While the end-to-end CNN performed significantly better than the MLP on the regression task, the performance on regression was the worst of the three schemes for each model, making it somewhat irrelevant. Somewhat counterintuitively, slightly higher accuracy could be obtained by binning the output of the GB regressor (64%) which had a higher RMSE compared to the KR regressor (28 ± 3°C versus 26 ± 1°C). This suggests that least-squares regression may be placing too much weight on outliers, which are less influential in the case of ordinal classification. It is even possible that the growth temperatures are not really ordinal after all, perhaps with 950 °C representing a value close to optimal while 900 °C and 1000 °C could be a similar distance away from optimal.

The best-performing model across any type or scheme was the MLP NNrank ordinal classifier with an accuracy of 71%. For the NNRank applied to the MLP and CNN, the average test accuracy of the CNN and MLP improved minimally with +2% and + 4%, repectively, over the nominal classification. This improvement is accounted for mainly in reduced classification errors of the 1000 °C images from 75% to 82% accuracy (Fig. 7).

In an effort to explain the surprising trend observed in the ordinal treatment of the data, we obtained the first 2 principal components of the data using principal component analysis (PCA) [60]. The image classes are embedded in the 2 components shown in Fig. 8. The figure shows overlap of all three classes and more significantly between neighboring classes, with very poor separation visible in the first two components. We visualize the micrographs in the PCA space in Fig. 9,

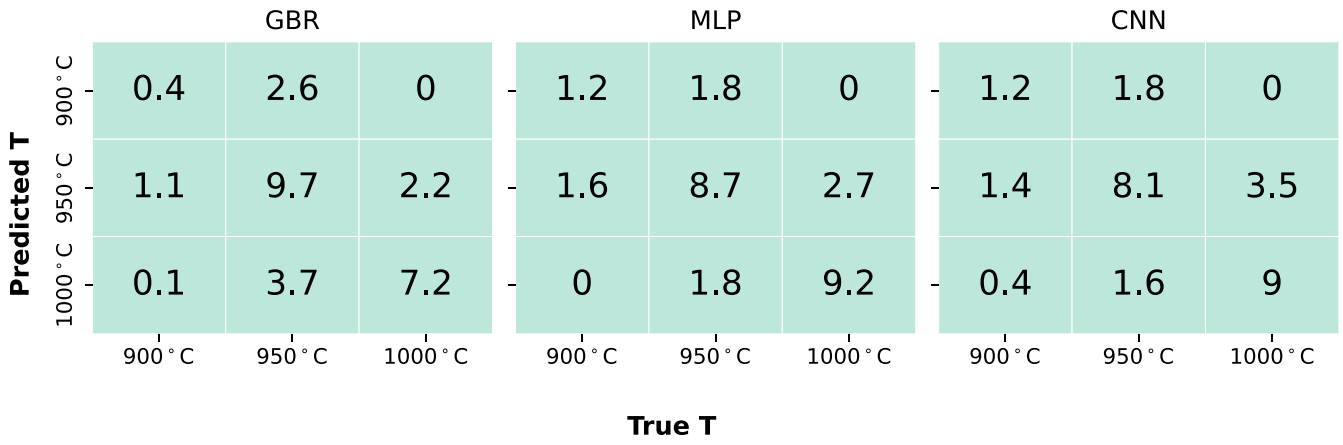

Fig. 7. The average confusion matrix of the 3 classes of temperature (900, 950, and 1000 °C) on the test set for the 9 different model architectures. This is based on ordinal classification with regression used for the GBR and NNrank for the MLP and CNN.

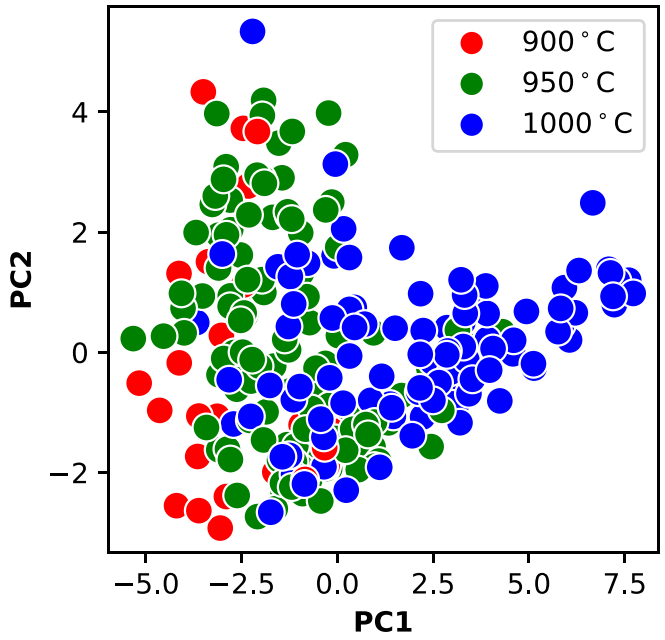

Fig. 8. The first two principal components of the image features showing the temperature class distribution in the reduced dimensional representation from the principal component analysis. Significant overlap is observed among the different classes in the embedding space.

indicating variations in the domain size (PC1) and density (PC2). Because these features vary significantly even within the same temperature class (e.g., see Fig. 8), the image feature vectors likely do not show consistent trends from 900 °C to 950 °C to 1000 °C, leading to no advantage in the ordinal treatment of growth temperature.

3.6. Model explanations

Beyond the capacity of the ML models to isolate the morphological features associated with the different growth temperature of the thin film MoS_2 based on their AFM images, we want to understand what features of the images the models used in the classification. Class activation maps (CAM) of the different classes are therefore obtained following the implementation by Zhou, et al. [61] The feature maps of the last convolutional layer are summed and then normalized by dividing by the maximum value to obtain a heatmap with the same dimensions as the layer. The bright yellow spot on the class activation maps represent the region with the highest activation which the model used for the classification.

Additionally, we obtained the occlusion attribution; the probability of a class of image as a function of an occluder object [62], using the implementation in Captum library [63]. To achieve this, we iteratively set a patch of the image to be zero-pixel values and then obtain the probability of the class. Stride size of 5×5 and the patch size of 15×15 were used. The probability is visualized as a 2D heat map. Both positive and negative attributions, indicating that the presence and absence of the area, respectively, increases the prediction scores are shown on the heat map. The occlusion attribution is applied to four sample images, for each class, correctly predicted by the CNN model. Green regions on the image have positive attributions while red regions have negative attribution.

The CAM and occlusion attribution in Fig. 10 show substantial agreement in identifying the activation region, with the latter giving more specific spatial attribution. The activation features are easier to perceive in images with bigger domain sizes, especially those grown at higher temperature. For some of the images from samples grown at higher temperature and which show clearly defined domains, some domain boundaries are highlighted, indicating the model's reliance on the boundaries in identifying such images. Also, regions with clean multi-steps crystals are shown to be important for the model in the classification (Fig. 10c), while the messy crystals post adverse effect to class attribution, as shown in the occlusion attribution.

From the experimental observations, the samples grown at higher temperatures are expected to exhibit greater domain sizes [3,4]. However, in the data used in training our models, there is significant variation in the quality of the samples, such that most images grown at higher temperature do not necessarily have greater domain sizes (Figs. 2 and 6). Additionally, if the model depends on domain size in identifying the images, it will be difficult to visually identify such features in images with less defined domains, and the only difference among the classes would only be the magnitude of the same feature. This is unlike the natural images where activation of different classes are typically associated with unique features of the classes that can be visually identified [56–58]. The models have therefore shown to be capable of identifying image features that humans could potentially miss.

4. Conclusion

This study focuses on the development of ML models for the classification of AFM images of thin film MoS_2 based on the growth temperatures of their samples. Many different strategies were explored for generating feature vectors, including using different pretraining image domains, extracting features from different depths in a pretrained

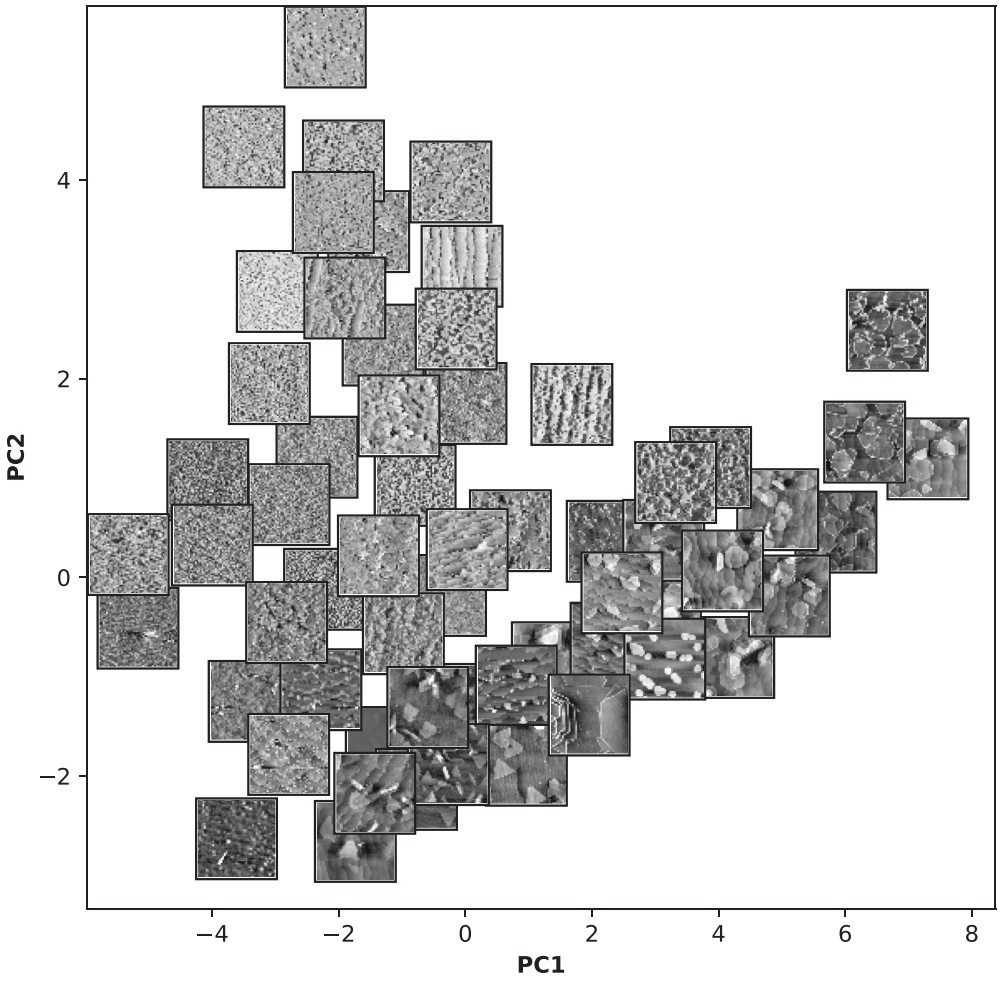

Fig. 9. The first two principal components of the image features showing the sample images, in the reduced dimensional representation from the principal component analysis. The embedding shows that the first dimension (PC1) is associated with the domain size, while the second dimension (PC2) seems to indicate the domain density.

ResNet, and end-to-end fine-tuning. A novel approach to transfer learning where the convolutional filters of the pretrained model were first fine-tuned before using them to extract features was also introduced. Our scheme yielded better results than the traditional approaches. Different augmentation strategies from the literature were evaluated to determine their effect on overall model performance. Beyond these pretraining schemes, nine different ML algorithms were evaluated to determine the most suitable approach for identifying morphological features associated with different growth temperatures.

The study also examined the impact of considering the ordinality of the classes on the accuracy of the models in identifying AFM images grown at different temperatures. We found that accounting for ordinality (i.e., by switching from classification to regression loss functions) improved the accuracy of some algorithms while decreasing performance for others. For instance, the best model overall was obtained using an NNrank ordinal classifier, but some nominal classifier were nearly as accurate. Furthermore, some algorithms had equivalent accuracy regardless of whether the data was treated as nominal classes or ordinal. Thus, there seems to be no clear advantage to using least-squares regression here, despite the data appearing in the form of continuous, ordered growth temperatures, which is a counterintuitive result.

To address class imbalance, weighted random sampling and oversampling techniques were employed, and robust ML models that generalize well to out-of-sample data were developed using model ensembles. The best-performing algorithms, MLP and end-to-end CNN, achieved classification accuracy of about 70% on held-out test data. The high accuracy obtained demonstrates the effectiveness of ML in accurately identifying thin films grown at different temperatures, despite the limitations of other inconsistent growth parameters and imbalances in the training data.

This study also sought to understand the features utilized by the ML models for classification by obtaining class activation maps and occlusion attribution. These strategies revealed that images from samples grown at higher temperatures, exhibiting well-defined domains, had the highest activation at the domain boundaries, aligning with experimental observations. Moreover, the models demonstrated the capability to identify latent features that humans could potentially miss, accurately classifying images with varying domain sizes that would be challenging for human experts. Future work may explore the relationship between these image features and additional attributes of the samples; the robustness of these features across growth chambers, characterization instruments, and even repeatability over time may be interesting ways to utilize the quantitative capability of deep learning to unlock new insights into challenging materials synthesis problems.

CRediT authorship contribution statement

Isaiah A. Moses: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Wesley F. Reinhart:** Conceptualization, Formal analysis, Funding acquisition,

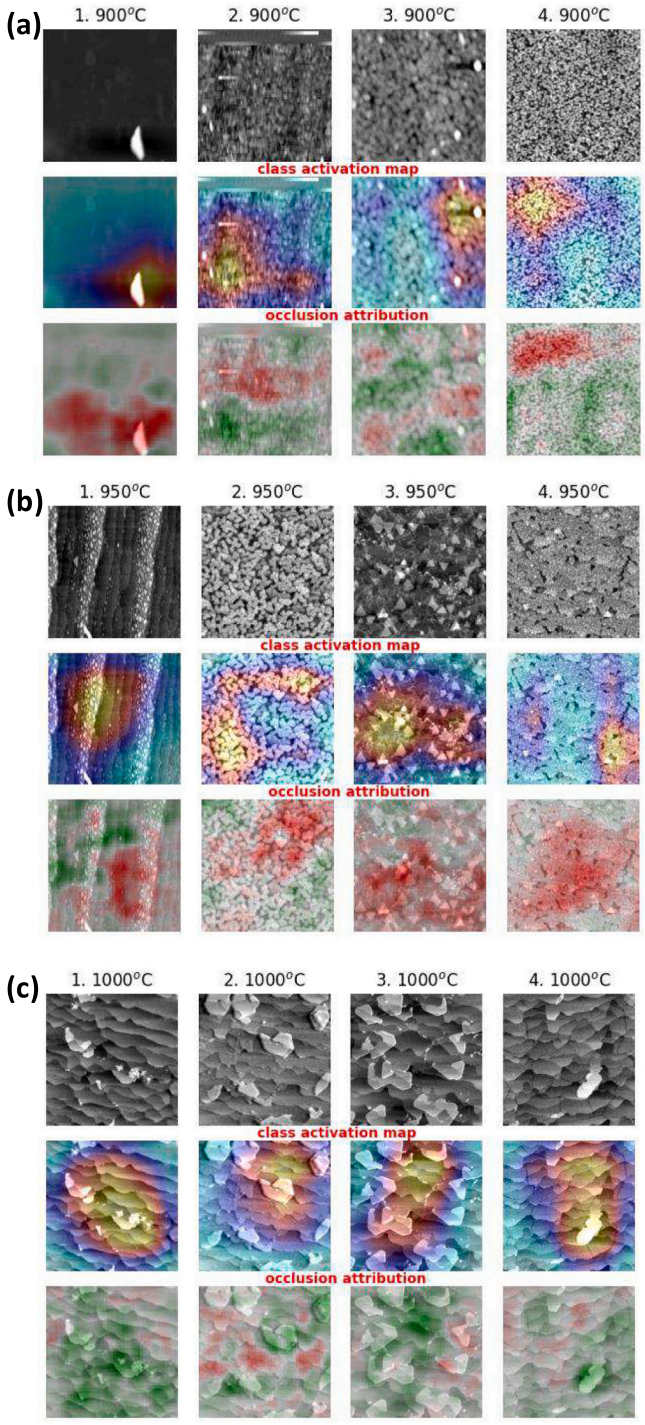

Fig. 10. Class activation maps (CAM) and occlusion attribution showing different regions of the images the model used for the classification. (a), (b), and (c) are for samples of images grown at 900 °C, 950 °C and 1000 °C, respectively. (i), (ii), and (iii) are the original AFM images, CAM, and AFM images overlaid with CAM, respectively.

Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Data availability

The raw data required to reproduce these findings are available to download from Ref.30 The processed data required to reproduce these findings are available to download from Ref.64 The codes used for this work can be accessed at https://zenodo.org/records/10534837

Acknowledgments

This study is based upon research conducted at The Pennsylvania State University Two-Dimensional Crystal Consortium – Materials Innovation Platform (2DCC-MIP) which is supported by NSF cooperative agreement DMR-2039351.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.matchar.2024.113701.

References

[1] A. Zavabeti, A. Jannat, L. Zhong, A.A. Haidry, Z. Yao, J.Z. Ou, Two-dimensional materials in large-areas: synthesis, properties and applications, Nano-Micro Lett. 12 (2020) 1–34.
[2] Y. Lei, T. Zhang, Y.-C. Lin, T. Granzier-Nakajima, G. Bepete, D.A. Kowalczyk, Z. Lin, D. Zhou, T.F. Schranghamer, A. Dodda, et al., Graphene and beyond: recent advances in two-dimensional materials synthesis, properties, and devices, ACS Nanosci.Au 2 (6) (2022) 450–485.
[3] S.M. Eichfeld, L. Hossain, Y.-C. Lin, A.F. Piasecki, B. Kupp, A.G. Birdwell, R. A. Burke, N. Lu, X. Peng, J. Li, et al., Highly scalable, atomically thin WSe2 grown via metal–organic chemical vapor deposition, ACS Nano 9 (2015) 2080–2087.
[4] X. Zhang, Z.Y. Al Balushi, F. Zhang, T.H. Choudhury, S.M. Eichfeld, N. Alem, T. N. Jackson, J.A. Robinson, J.M. Redwing, Influence of carbon in metalorganic chemical vapor deposition of few-layer WSe 2 thin films, J. Electron. Mater. 45 (2016) 6273–6279.
[5] X. Zhang, T.H. Choudhury, M. Chubarov, Y. Xiang, B. Jariwala, F. Zhang, N. Alem, G.-C. Wang, J.A. Robinson, J.M. Redwing, Diffusion-controlled epitaxy of large area coalesced WSe2 monolayers on sapphire, Nano Lett. 18 (2018) 1049–1056.
[6] S. Wang, Y. Rong, Y. Fan, M. Pacios, H. Bhaskaran, K. He, J.H. Warner, Shape evolution of monolayer MoS2 crystals grown by chemical vapor deposition, Chem. Mater. 26 (2014) 6371–6379.
[7] S. Xie, M. Xu, T. Liang, G. Huang, S. Wang, G. Xue, N. Meng, Y. Xu, H. Chen, X. Ma, et al., A high-quality round-shaped monolayer MoS 2 domain and its transformation, Nanoscale 8 (2016) 219–225.
[8] M. Suleman, S. Lee, M. Kim, V.H. Nguyen, M. Riaz, N. Nasir, S. Kumar, H.M. Park, J. Jung, Y. Seo, NaCl-assisted temperature-dependent controllable growth of large-area MoS2 crystals using confined-space CVD, ACS Omega 7 (2022) 30074–30086.
[9] T. Li, W. Guo, L. Ma, W. Li, Z. Yu, Z. Han, S. Gao, L. Liu, D. Fan, Z. Wang, et al., Epitaxial growth of wafer-scale molybdenum disulfide semiconductor single crystals on sapphire, Nat. Nanotechnol. 16 (2021) 1201–1207.
[10] Y. Xiang, X. Sun, L. Valdman, F. Zhang, T.H. Choudhury, M. Chubarov, J. A. Robinson, J.M. Redwing, M. Terrones, Y. Ma, et al., Monolayer MoS2 on sapphire: an azimuthal reflection high-energy electron diffraction perspective, 2D Mater. 8 (2020) 025003.
[11] X. Yang, S. Li, N. Ikeda, Y. Sakuma, Oxide scale sublimation chemical vapor deposition for controllable growth of monolayer MoS2 crystals, Small Meth. 6 (2022) 2101107.
[12] Y. Han, B. Tang, L. Wang, H. Bao, Y. Lu, C. Guan, L. Zhang, M. Le, Z. Liu, M. Wu, Machine-learning-driven synthesis of carbon dots with enhanced quantum yields, ACS Nano 14 (2020) 14761–14768.
[13] G.H. Gu, J. Jang, J. Noh, A. Walsh, Y. Jung, Perovskite synthesizability using graph neural networks, npj Computat. Mater. 8 (2022) 71.
[14] I.A. Moses, R.P. Joshi, B. Ozdemir, N. Kumar, J. Eickholt, V. Barone, Machine learning screening of metal-ion battery electrode materials, ACS Appl. Mater. Interfaces 13 (2021) 53355–53362.
[15] I.A. Moses, V. Barone, J.E. Peralta, Accelerating the discovery of battery electrode materials through data mining and deep learning models, J. Power Sources 546 (2022) 231977.
[16] Y. Yan, D. Lu, K. Wang, Accelerated discovery of single-phase refractory high entropy alloys assisted by machine learning, Comput. Mater. Sci. 199 (2021) 110723.
[17] S. Lu, Q. Zhou, Y. Guo, J. Wang, On-the-fly interpretable machine learning for rapid discovery of two-dimensional ferromagnets with high curie temperature, Chem 8 (2022) 769–783.
[18] J. Yang, H. Yao, Automated identification and characterization of two-dimensional materials via machine learning-based processing of optical microscope images, Extreme Mech. Lett. 39 (2020) 100771.
[19] B. Han, Y. Lin, Y. Yang, N. Mao, W. Li, H. Wang, K. Yasuda, X. Wang, V. Fatemi, L. Zhou, et al., Deep-learning-enabled fast optical identification and characterization of 2D materials, Adv. Mater. 32 (2020) 2000953.

[20] G. Jung, S.G. Jung, J.M. Cole, Automatic materials characterization from infrared spectra using convolutional neural networks, Chem. Sci. 14 (2023) 3600–3609.
[21] Z. Si, D. Zhou, J. Yang, X. Lin, 2D material property characterizations by machine-learning-assisted microscopies, Appl. Phys. A 129 (2023) 248.
[22] Y. Saito, K. Shin, K. Terayama, S. Desai, M. Onga, Y. Nakagawa, Y.M. Itahashi, Y. Iwasa, M. Yamada, K. Tsuda, Deep-learning-based quality filtering of mechanically exfoliated 2D crystals, npj Computat. Mater. 5 (2019) 124.
[23] B. Tang, Y. Lu, J. Zhou, T. Chouhan, H. Wang, P. Golani, M. Xu, Q. Xu, C. Guan, Z. Liu, Machine learning-guided synthesis of advanced inorganic materials, Mater. Today 41 (2020) 72–80.
[24] J.L. Beckham, K.M. Wyss, Y. Xie, E.A. McHugh, J.T. Li, P.A. Advincula, W. Chen, J. Lin, J.M. Tour, Machine learning guided synthesis of flash graphene, Adv. Mater. 34 (2022) 2106506.
[25] M. Lu, H. Ji, Y. Zhao, Y. Chen, J. Tao, Y. Ou, Y. Wang, Y. Huang, J. Wang, G. Hao, Machine learning-assisted synthesis of two-dimensional materials, ACS Appl. Mater. Interfaces 15 (2022) 1871–1878.
[26] N.C. Frey, J. Wang, G.I. Vega Bellido, B. Anasori, Y. Gogotsi, V.B. Shenoy, Prediction of synthesis of 2D metal carbides and nitrides (MXenes) and their precursors with positive and unlabeled machine learning, ACS Nano 13 (2019) 3031–3041.
[27] B. Ryu, L. Wang, H. Pu, M.K. Chan, J. Chen, Understanding, discovery, and synthesis of 2D materials enabled by machine learning, Chem. Soc. Rev. 51 (2022) 1899–1925.
[28] T.F. Schranghamer, N.U. Sakib, M.U.K. Sadaf, S. Subbulakshmi Radhakrishnan, R. Pendurthi, A.D. Agyapong, S.P. Stepanoff, R. Torsi, C. Chen, J.M. Redwing, et al., Ultrascaled contacts to monolayer MoS2 field effect transistors, Nano Lett. 23 (2023) 3426–3434.
[29] N. Trainor, C. Chen, H. Zhu, T.V. Mc Knight, T.H. Choudhury, J.M. Redwing, Epitaxial growth of wafer-scale transition metal dichalcogenide monolayers by metalorganic chemical vapor deposition, in: 2022 6th IEEE Electron Devices Technology & Manufacturing Conference (EDTM), 2022, pp. 160–162.
[30] I.A. Moses, W.F. Reinhart, Data: Quantitative Analysis of Chalcogenide Thin Film Micrographs with Machine Learning, 2023. https://m4-2dcc.vmhost.psu.edu/list/data/vdn4yTd70vdo.
[31] A.R. Kitahara, E.A. Holm, Microstructure cluster analysis with transfer learning and unsupervised learning, Integr. Mater. Manuf. Innov. 7 (2018) 148–156.
[32] Y. Gong, H. Shao, J. Luo, Z. Li, A deep transfer learning model for inclusion defect detection of aeronautics composite materials, Compos. Struct. 252 (2020) 112681.
[33] R. Cohn, E. Holm, Unsupervised machine learning via transfer learning and k-means clustering to classify materials image data, Integr. Mater. Manuf. Innov. 10 (2021) 231–244.
[34] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint (2014) arXiv:1409.1556.
[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[39] J. Stuckner, B. Harder, T.M. Smith, Microstructure segmentation with deep learning encoders pre-trained on a large microscopy dataset, npj Comput. Mater. 8 (2022) 200.
[40] E.D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, Q.V. Le, Autoaugment: Learning augmentation strategies from data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 113–123.
[41] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, in: ACM Transactions on Intelligent Systems and Technology (TIST) 2, 2011, pp. 1–27.
[42] J. Platt, et al., Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Adv. Large Margin Class. 10 (1999) 61–74.
[43] K.P. Murphy, Machine Learning: AProbabilistic Perspective, MIT Press, 2012.
[44] J. Goldberger, G.E. Hinton, S. Roweis, R.R. Salakhutdinov, Neighbourhood components analysis, Adv. Neural Inf. Proces. Syst. 17 (2004).
[45] C.E. Rasmussen, C.K. Williams, et al., Gaussian Processes for Machine Learning vol. 1, Springer, 2006.
[46] A. Cutler, D.R. Cutler, J.R. Stevens, Random forests, Methods and applications, Ensemble machine learning, 2012, pp. 157–175.
[47] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232.
[48] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Netw. 2 (1989) 359–366.
[49] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, L. Jackel, Handwritten digit recognition with a back-propagation network, Adv. Neural Inf. Proces. Syst. 2 (1989).
[50] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (1998) 2278–2324.
[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: an imperative style, high-performance deep learning library, Adv. Neural Inf. Proces. Syst. 32 (2019).
[53] J. Cheng, Z. Wang, G. Pollastri, A neural network approach to ordinal regression, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, pp. 1279–1284.
[54] F.R.S., K. P. LIII, On lines and planes of closest fit to systems of points in space, London, Edinburgh, and Dublin Philosoph. Magaz. J. Sci. 2 (1901) 559–572.
[55] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 374 (2016) 20150202.
[56] A. KwaÅ)niewska, J. RumiÅski, P. Rad, Deep features class activation map for thermal face detection and tracking, in: 2017 10th International Conference on Human System Interactions (HSI), 2017, pp. 41–47.
[57] M.B. Muhammad, M. Yeasin, Eigen-CAM: class activation map using principal components, in: 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–7.
[58] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, Y. Wei, LayerCAM: exploring hierarchical class activation maps for localization, IEEE Trans. Image Process. 30 (2021) 5875–5888.
[59] 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, arXiv preprint (2018). arXiv:1811.12231.
[60] K.L.I.I.I. Pearson, On lines and planes of closest fit to systems of points in space, London, Edinburgh, and Dublin Philosoph. Magaz. J. Sci. 2 (1901) 559–572.
[61] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning Deep Features for Discriminative Localization, Computer Vision and Pattern Recognition, 2016.
[62] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13 (2014) 818-833.
[63] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, O. Reblitz-Richardson, Captum: a unified and generic model interpretability library for PyTorch, arXiv preprint (2020) arXiv:2009.07896.

Further reading

[64] I.A. Moses, W.F. Reinhart, Processed Data for Quantitative Analysis of MoS_2 Thin Film Micrographs with Machine Learning, 2023; doi:10.5281/zenodo.8432222.